POSSE Macros Users' Guide for the Exploration of Observational Health Care Data

(Version 1.0, 2019)

William E. Miller, Kathleen B. Fedan and John M. Wood

Correspondence to: W.E. Miller, Respiratory Health Division, National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention (CDC), 1095 Willowdale Road, Morgantown, WV 26505-2888

Telephone: 304-285-5772
FAX: 304-285-5820
E-mail: wem0@cdc.gov

DISCLAIMER

The findings and conclusions in this users' guide are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health. Mention of product names does not imply endorsement by NIOSH/CDC. In addition, CDC and NIOSH do not warrant the reliability or accuracy of the software, graphics or text.

Acknowledgements

The writings of John Tukey provided inspiration for the POSSE methods. We are also indebted to the numerous works of Michael Greenacre, who transmitted much of the early French work on correspondence analysis to the English-speaking world, and added new techniques to the field. Special thanks are due to Steve Simon of the University of Missouri – Kansas City for his suggestions for clarifying portions of the users' guide. We are also grateful for the reviews provided by Gerry Hobbs, Anna Mnatsakanova, Steve Bertke and I-Chen Chen of the National Institute for Occupational Safety and Health (NIOSH). We also thank James T. Wassell of NIOSH who reviewed some early proposals for the methods and Dan Heath of SAS for his assistance in converting the macros for use in the newer SAS statistical graphical system. Many thanks are also due to Brian Tift, Nicole Edwards and Michelle Martin of NIOSH for their help with this project.

Preface

We are drowning in information and starving for knowledge.
-- Rutherford D. Rogers

The numbers of variables which are typically seen in data sets today, as compared to previous decades, suggest that we now, more than ever, live in a multivariate world. Therefore, applications of multivariate and exploratory methods might be more important than ever. This users' guide and the associated software represent an effort to make a few multivariate statistical tools more accessible to the scientific community, and in particular to those who use the SAS® programming language. The methods have been designed for those who have experience with regression modeling (including logistic regression), and who also wish to learn about the use of exploratory methods in a global search for patterns, with an emphasis on observational health-care data.

The methods presented in this guide are especially suited for the examination of observational studies[1,2]. In contrast to the classic situation of a controlled experiment with an assigned treatment and random sampling, observational studies involve non-random samples where the sources of variation are not well-known and where the hypotheses might be unclear. An exploratory analysis is often focused on examining the systematic variation, rather than on the error structure which is a critical component for statistical modeling. Also, unlike a confirmatory analysis where the goal might be to validate a hypothesis or replicate a result, an exploratory analysis might be used to generate new hypotheses and also be focused on identifying patterns which are unique to a sample, without regard to whether or not such patterns will be confirmed or replicated by additional data.

The methods presented here are collectively referred to as the 'POSSE' methods. This name suggests that, just as a posse is organized for the purpose of making a thorough search over a wide area (for an outlaw or a lost child), these methods are intended to provide a systematic and extensive search for patterns among variables and observations. It should be emphasized that the POSSE methods are usually not an aim unto themselves, and none of its results may actually be included among published results. The general goal of these methods is to screen variables and make a global search for interesting patterns, in order to produce results which are temporary. The outcome of applying the POSSE methods can also be quite varied, ranging from a simple reduction in the number of variables for a subsequent analysis to the identification of a subset of observations with an unusual pattern of results.

A multivariate statistics text will typically cover procedures for (a) finding clusters or groupings of similar observations, (b) determining groupings of correlated variables, (c) examining the interdependence of variables or groups of variables, and (d) finding transformations of the data that can simplify the overall structure. All these topics will be introduced here, to some extent, through the applications of two methods: (1) simple correspondence analysis and (2) homogeneity analysis, an approach to multiple correspondence analysis. These two methods will, however, involve a simplification where any continuous or quantitative variables are converted to categorical variables. That can seem like a limitation, along with some loss of information for the quantitative variables, but it will make it easier to explore the data. For example, some multivariate methods are designed to detect linear associations, but the POSSE methods will allow a researcher to also detect quadratic, cubic, and other complicated associations. Because the POSSE methods are applied near the beginning of an analysis, a researcher can then employ the POSSE results to inform the subsequent modeling that uses the original quantitative variables.

A quote attributed to Albert Einstein is that, "Everything should be made as simple as possible, but not

simpler." We have tried to balance our presentation by sparing the reader any unnecessary theory while also providing a number of references for those who want to obtain a deeper understanding of the methods. Just as you don't need to be a mechanic to drive a car, you don't need to have a complete understanding of the underlying theory to successfully apply these statistical techniques. The examples will provide some context for applying the methods and, in some cases, demonstrate how the various techniques complement each other. Proceeding carefully through the examples and having some practice with submitting the SAS® macros will be essential to learning how the POSSE methods can help to provide a more coherent data strategy. Because they use the newer statistical graphics procedures in SAS®, the POSSE macros will also work for the SAS® University Edition, which is freely available to teachers, students and academic researchers. For users of the R programming language, it should also be mentioned that many of the new features provided to SAS users by the POSSE macros are already available in the **ca** package[3].

Table of Contents

I. BACKGROUND INFORMATION 7
1. Introduction 7
2. Exploratory Data Analysis 7
3. Defining the POSSE Methods 8
3.1 Global Analysis (phase 3) 8
3.2 Naming the Toolbox 9
II. THE POSSE METHODS 10
4. Multivariate Categorical Data Analysis 10
4.1 Scaling Issues 10
4.2 Types of Categorical Scales 10
4.3 Detecting Patterns Using a Categorical Approach 11
5. Simple Correspondence Analysis 13
5.1 Correspondence Maps 13
5.2 Correspondence Analysis as Pattern Discovery 18
5.3 Correspondence Analysis as Scaling 18
5.4 Applications to Contingency Tables 20
6. Homogeneity Analysis 23
6.1 Basic Principles 24
6.2 Classifying Observations 26
6.3 Classifying Variables 28
6.4 Summary for Homogeneity Analysis 29
7. A Special Application for Deriving Categorical Variables 30
7.1 Interpreting the Cluster Tree and Cluster History 31
7.2 A Method for Choosing Categories 32
III. SOFTWARE FOR THE POSSE METHODS 34
8. The POSSE Macros 34
8.1 The *Data_Prep* Macro 38
8.2 The *Prelim_CA* Macro 40
8.3 The *Correspondence* Macro 41
8.4 The *Classification* Macro 44
8.5 The *Tabulation* Macro 45
9. Additional Recommendations 46
9.1 Row and Column Interactions 47

9.2 Stacked Tables and Screening Variables 48
9.3 Limitations 48
10. Key Terms 49
11. Concluding Remarks 50
IV. APPENDIX A: Accessing and Using the POSSE Macros 52
V. APPENDIX B: Example Data Sets 53
B.1 Intensive Care Unit (ICU) Data 53
B.2 Diabetes Data 54
B.3 Occupational Asthma Data 55
B.4 Coronary Risk Factor Data 56
B.5 Cancer Information Data 56
B.6 X-Ray Classification Data 57
VI. APPENDIX C: Examples 58
C.1: Converting a Table to a SAS Data Set 58
C.2: Determining the Categories for a Derived Categorical Variable 59
C.3: Creating a Categorical Data Set Using the *Data_Prep* Macro 63
C.4: Screening Variables 64
C.5: Examining the Scaling for a Set of Variables 68
C.6: Inspecting an Outlier 71
C.7: Deriving a Cluster Variable 75
C.8: Exploring Main Effects and Interactions 78
C.9: Assessing General Associations 81
C.10: Exploring Missing Values 85
C.11: Data Preparation for Multiple Quantitative Responses (repeated measures) 87
C.12: Exploring the Bias in Square Tables (repeated measures) 88
C.13: Examining the Changes in Cluster Profiles (repeated measures) 94
VII. APPENDIX D: Review of Chi-Square Statistic 103
References 105

…exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there.
-- John Tukey [4]

I. BACKGROUND INFORMATION

1. Introduction

This text will introduce the reader to the methods for the users' guide by describing exploratory approaches to data analysis and illustrating how complicated patterns can be detected using categorical methods. To provide some context, a few relevant features of multivariate analysis are described before introducing the principles underlying two types of *correspondence analysis*♦ and their applications. In addition, the parameters will be defined for the five SAS macros that implement the methods. This text acts as a prerequisite and reference for the examples given in Appendix C.

2. Exploratory Data Analysis

The area of exploratory data analysis is often identified with a collection of techniques developed by John Tukey[5] as a way to have a first look at data using various numerical summaries and graphical displays. Exploratory methods have developed, at least in part, in reaction to two challenges that grew out of applying statistical theory to data. The first challenge, sometimes referred to as 'multiplicity' in statistical literature, arose from the problems associated with correctly interpreting results after applying a large number of hypothesis tests to one data set. Classical methods which use hypothesis-testing and p-values are useful for confirmatory studies with clear and pre-defined objectives, but turn out to be poor tools for exploratory analyses. Over the years, a number of proposals have been made to remedy this situation. An early work is from Daniel[6], who argues that it is necessary to develop a single multivariate system in order to assess the effects associated with all possible outcomes from a host of factors.

The second challenge resulted from the need to address violations of assumptions that are implicit in statistical theory. This led to two developments, the first being the growth of robust statistics, which provides methods of estimation which are resistant to violations of assumptions. The second development was the area of diagnostics, which includes tools for assessing assumptions, finding outliers or identifying overly influential observations. By assessing the data through diagnostics, formal models and estimation could then be adjusted to account for any anomalies in the data.

Diagnostics are applied during the modelling that occurs later in an analysis, whereas the methods in this guide are used at an early stage of the data analysis. However, the aims and techniques of diagnostics are closely related to some of the applications that are presented in this guide. One way to think about this is to regard modeling and diagnostics as complementary processes. A fitted model tends to smooth over the bumps and warts in the data to find generalizable, broad patterns. Diagnostics tend to uncover the bumps and warts, helping to isolate those portions of the data which are not generalizable. Together, modeling and diagnostics can then provide a more complete view of the data.

To further describe exploratory methods, we refer to the distinction given by Wilk and Gnanadesikan[7]

♦ Italicized terms in the text are defined in Section 10.

between external versus internal comparisons. For example, when performing a standard regression analysis, we usually assess the estimated coefficients by individually comparing each of them to an external probability distribution, such as the *t*-distribution. However, when we assess the residuals from the model, we compare all of them simultaneously, such as in a quantile-quantile plot, where the residuals are plotted against the expected values for a normal distribution. Rather than compared piecewise in a multitude of local comparisons, as in a series of *t*-tests, the quantities are being assessed together in a single global comparison. This second comparison is an example of what Wilk and Gnanadesikan call an internal comparison. In this guide we denote this as a global comparison. Global comparisons, which involve simultaneous comparisons among like quantities, are characteristic of exploratory techniques (see, for example, Johnson and Tukey[8]).

3. Defining the POSSE Methods

3.1 Global Analysis (phase 3)

If we ignore, for the moment, the steps involved in the planning of a study, as well as the collection and cleaning of the data, then as shown below we might broadly depict the remaining tasks for the data analysis in terms of (1) the description of the data, (2) the modeling, and (3) the interpretation of the results. Mallows and Tukey[9] have further divided the data description into the three phases given in 1(a)-(c).

1.) Data Description (informal analyses)
   (a) univariate analyses (phase 1)
   (b) bivariate analyses (phase 2)
   (c) automated global examination of variables (phase 3)
2.) Modeling and Estimation (formal analyses)
3.) Interpretation of Results

The first phase involves the examination of single variables, the second phase the examination of pairs of variables, and the third phase a simultaneous global examination of all the variables. Mallows and Tukey state that a phase 3 procedure is "badly needed" and discuss several possible methods. One method they mention, which was less well-known at the time, is known in English as *correspondence analysis* or *simple correspondence analysis*, a largely graphical method developed by French statisticians such as Jean-Paul Benzécri. Other forms of correspondence analysis have also emerged periodically under the names of reciprocal averaging, optimal scaling, and dual scaling[10]. Other developments include canonical correspondence analysis[11], subset correspondence analysis[12], and *homogeneity analysis*[13]. Homogeneity analysis, which is an approach to multiple correspondence analysis, was developed by a group of statisticians who published under the pseudonym of Albert Gifi[14].

An article by Hill[15] refers to correspondence analysis as a neglected method, and over a decade later Greenacre and Hastie[16] state that they found it "surprising that the technique still remains relatively unknown outside the fields of psychology and ecology." This users guide attempts to widen its appeal by illustrating how some methods in correspondence analysis are especially suited to a phase 3 analysis. Both simple correspondence analysis and multiple correspondence analysis (i.e., homogeneity analysis) will be introduced in this users' guide.

3.2 Naming the Toolbox

Bergman and Magnusson[17] make a distinction between variable-oriented methods and person-oriented methods. For example, the multivariate method named principal component analysis is focused on finding associations among variables, and can be applied using only the correlation matrix for the variables, that is, without direct access to the observations themselves. Principal component analysis is then an example of a variable-oriented method. On the other hand, another multivariate method named cluster analysis is used to group similar observations and, therefore, information for the individual observations is required to perform a cluster analysis. Cluster analysis is then an example of a person-oriented method. Although both variable-oriented and person-oriented approaches are presented, the examples in this guide will contain many instances where the emphasis is on uncovering hidden structures with respect to groups of observations, or perhaps identifying a subset of observations that is primarily connected with a particular pattern of results.

The methods for this users' guide are designed to be a toolbox for a global analysis. We refer to the methods collectively as the POSSE methods for several reasons. First, because there is an emphasis on person-oriented methods, we can identify the POSSE name as an acronym for 'Person-Oriented Sample Screening and Exploration.' Secondly, because the Latin phrase 'in posse' means 'in possibility' or 'potentially,' we can view the methods as a means for exploring the potential for structure (i.e., systematic variation) in a data set.

The third and most accessible connotation for this name follows from the usual dictionary definition of posse as a group of people temporarily organized to make a search over a wide area, where the search, for instance, could be for a lost child. During the first day in its search for a lost child, the posse might not find the child, and yet the efforts expended can serve to substantially reduce the search area. In a similar way, researchers can apply the POSSE methods with the objective of reducing their search area by finding a subset of important or relevant variables during an early stage of an analysis.

In our attempt to design a toolbox for exploring data, we found it to be advantageous to have a single theoretical framework. Therefore, some form of correspondence analysis is included in each of the techniques used by the POSSE methods. These forms include using *simple correspondence analysis* to examine various types of tables, and using *homogeneity analysis* (i.e., multiple correspondence analysis) to analyze two-dimensional data frames or data matrices consisting of observations (rows) and variables (columns). The flexibility of correspondence analysis is illustrated in a variety of applications, including one application designed for data preparation.

In summary, the POSSE methods are intended to provide a global screening of variables, detect both linear and nonlinear patterns, furnish an initial examination of covariates, and identify special subgroups. In contrast to the classic situation of confirmatory statistical analysis, the POSSE methods are especially suited to studies where sources of variation are not well-known and the hypotheses are unclear.

II. THE POSSE METHODS

4. Multivariate Categorical Data Analysis

The tools described in this guide are designed to be accessible to those researchers who are comfortable with the use of regression models, their assumptions, and the use of basic diagnostic tools, such as the examination of residuals. Although this section will introduce some new concepts found in multivariate analysis, we will, in some cases, try to relate these concepts to methods that are learned in introductory statistics courses.

Someone who peruses textbooks on multivariate analysis[18-21] might come to regard it as a collection of loosely-related methods. Multivariate analysis includes methods which are used to examine the general interdependence of a single set of variables, and other methods which are designed to identify the relationship of a set of factors to either a single response variable or multiple response variables, in order to help determine the functional relationships for subsequent modeling. Readers will see both types of analyses in this guide.

4.1 Scaling Issues

One of the challenges of using multivariate analysis has to do with issues of scaling. To illustrate this, consider the situation where a researcher is analyzing a data set that contains fifty variables, where some of them are categorical variables and some of them are quantitative or continuous variables. In addition, the quantitative variables are not all normally distributed and have a variety of metrics, such as measures of heart-rate in beats per minute, weight in pounds, height in inches, and length in miles. How can we provide a single coherent analysis if the underlying measures are not comparable? A common approach for quantitative variables is to standardize them by transforming them to all have a mean of zero and a variance of one. However, that will not by itself address the fact that our variables are not all normally distributed, nor will it tell us how we might include the categorical variables in a single analysis.

In addition, other challenges emerge for the standardization of quantitative variables in some areas of multivariate analysis. Fleiss and Zubin[22] give an example of how systematic differences between two clusters in a cluster analysis can influence the standardization of the quantitative variables and, therefore, make it more difficult to identify the clusters. Chatfield and Collins[23] also illustrate how the results of a principal component analysis can depend on the scaling of the variables, and they show that the standardization of variables can make it more difficult to compare the results from different samples.

These challenges will be most critical at the beginning of an analysis when we have not yet reduced the number of variables. In Section 4.3, we will illustrate how starting with a categorical approach may facilitate this portion of the analysis, but we will first describe the types of categorical scales that will be encountered in this guide.

4.2 Types of Categorical Scales

Categorical variables can be measured on an unordered nominal scale, such as race or gender, or on an ordinal scale, such as categories based on ranges of age or weight. The ordinal age and weight variables are examples where each category represents some segment of a continuum. In this situation, it might

be important to determine how many categories are necessary for a variable. For example, as we move through the continuum of the weight variable for a human population, we might generally find those who are severely underweight with poor health, those who are average weight with better health, and finally those who are severely overweight with poor health. To reflect this situation in a health care study, we might need to specify at least three categories. (Section 7 will further discuss the derivation of categorical variables from quantitative ones.)

A second type of ordinal scaling is called *Guttman scaling*[10]. A Guttman scale implies a hierarchy of binary outcomes, so that a 'yes' or 'no' response for one variable depends on the outcome for another variable, such as in this sequence of questions:

> *Question 1a*: 'Do you smoke cigarettes?'
> *Question 1b*: 'If you do, do you smoke more than 20 cigarettes a day?'

Although there are theoretically 2^n total patterns of outcomes for n binary variables, this will reduce to only $(n + 1)$ possible patterns in practice for variables that follow a Guttman scale. For example, there are eight possible patterns of outcomes for the three unrelated binary variables, but only four possible patterns if they follow a Guttman scale. In some cases, it might be useful to combine Guttman-scaled variables into the categories of a single ordinal variable. For example, the information from the two dichotomous questions above might be combined into a new variable with the three ordinal categories and labels of 1='Non-Smoker', 2='Smoker', and 3='Heavy Smoker'.

4.3 Detecting Patterns Using a Categorical Approach

In addition to the problems connected with the scaling and standardization of continuous variables noted in Section 4.1, there are other reasons for considering a categorical approach near the beginning of an analysis. One is that some standard multivariate tools, such as principal component analysis, are typically searching for linear patterns using the standard Pearson correlation coefficient (r). They are, in this way, too specific for exploratory methods in their search for patterns[24].

For example, Figure 1 contains a scatterplot that appears to show a strong quadratic relationship between the variables X and Y, but the estimated Pearson correlation is r=0.11 with a p-value of 0.67. Suppose we derive two categorical variables with approximately equal numbers in each of three low-to-high categories from the original continuous variables, as illustrated by the grid-lines in Figure 1. The cross-tabulation for these new categorical variables is given in Table 1. The chi-square statistic, which tests for the independence between the row and column variables in the table, is equal to 10.6 with a p-value of 0.03. (See Appendix D for a review of the chi-square test of independence.) Even though there is a loss of information due to the conversion to categorical variables, as well as much smaller cell frequencies in Table 1 than is desirable for a chi-square test, we can now detect the quadratic relationship that exists for the original variables. In a global analysis with many variables, when we do not know the nature of the relations beforehand, both linear and nonlinear relationships can be detected using a categorical approach. For example, when there are sufficient data, the specification of four categories during the conversions of quantitative variables can allow us to detect linear, quadratic, cubic and other possible patterns, such as threshold or ceiling effects, which exist for the original quantitative variables. This information can then be used subsequently to construct better-fitting models for the original variables.

Another reason for considering a categorical approach near the beginning of an analysis has to do with the advances that have been made in categorical data analysis, some of which are illustrated in this guide.

In addition, the frequency tables produced by the POSSE methods will often present the results in ways that make the interpretation clearer. For instance, a table of frequencies can make it apparent what portion of the data is involved in a particular association, as well as how many observations are involved.

Figure 1. The scatterplot for computer-generated data points along with a quadratic trend line, where the correlation is equal to 0.11 with a p-value of 0.67. The grid represents some categorizations of the X and Y variables. The vertical lines of the grid represent the cut-points that partition the responses for X into three low-to-high categories with approximately equal numbers, and the horizontal lines carry out a similar partitioning of the Y responses. The values in the boxes represent the resulting numbers of points found in the sections of the grid. The three values found in a column of the grid (reading from bottom to top) correspond to the counts found in a row of the cross-tabulation of Table 1.

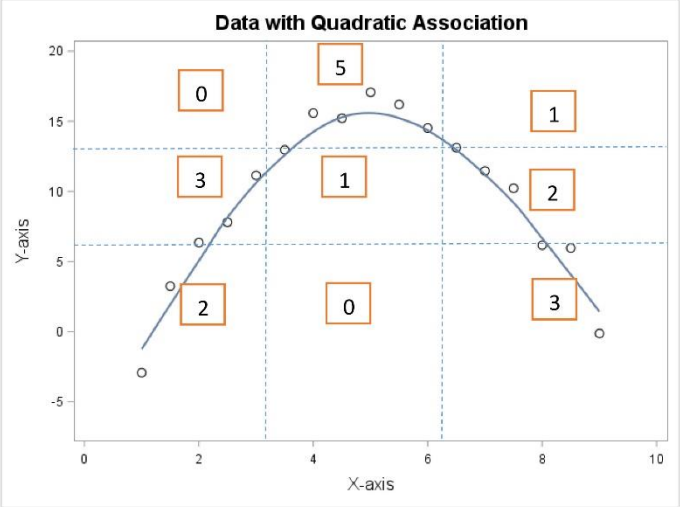


Table 1. The table which results from categorizing the two variables depicted in Figure 1. The associated chi-square (χ^2) value is equal to approximately 10.56 with a p-value of 0.03. The average row profile, which is calculated using the row totals, and the average column profile, which is calculated using the column totals, are shown in bold. (The row and column totals are equal in this example, but this will generally not be the case in correspondence analysis.)

Category Frequencies	YLow	YMed	YHigh	Total	Row Profiles Elements
XLow	2	3	0	5	0.400 0.600 0.000
XMed	0	1	5	6	0.000 0.167 0.833
XHigh	3	2	1	6	0.500 0.333 0.167
Total	5	6	6	17	**0.294 0.353 0.353**
Column Profile Elements	0.400 0.000 0.600	0.500 0.167 0.333	0.000 0.833 0.167	**0.294** **0.353** **0.353**	

5. Simple Correspondence Analysis

Correspondence analysis is a vast topic with a number of areas of current research. A presentation of the theory of correspondence analysis can be found in Greenacre[25] and a number of more or less intensive introductions are available[10, 16, 26-28]. In this section, we will review only those features of correspondence analysis that are applied in the POSSE methods, and we will limit ourselves to a more intuitive presentation of its basic principles. Correspondence analysis can be seen as much less rigorous than the formal methods of modeling and estimation. As Greenacre[25] states:

> By contrast, correspondence analysis has no aspiration of modelling the data statistically. The data serve as the population and every 'degree of freedom', as it were, is worth looking at… Also, being a more modest technique, correspondence analysis can be applied to data with less stringent properties, for example sparse matrices [tables] and matrices with some low cell frequencies.

Correspondence analysis includes both graphical and numerical summaries. The next section describes the graphical component, a plot referred to as a *correspondence map*. Later, we will introduce three numerical quantities: (1) the *inertia for a table*, (2) the *inertia for a dimension* and (3) the *contribution to an inertia*. For now, readers can regard inertias as measures of the variation or spread of the points in a correspondence map. In later sections, we will see how these inertias are related to each other when we discuss the splitting or decomposition of inertias, and readers will learn that relatively large values for the inertias or their components are associated with interesting features of a table. Readers will also learn how these quantities are connected with a correspondence map.

5.1 Correspondence Maps

Multivariate analysis often refers to multi-dimensional spaces. In practice, it can be difficult to clearly discern the positions of objects in even three dimensions. One of the aims of correspondence analysis is to find the best two-dimensional view of the results, and then display this in a map. We will begin our introduction to correspondence analysis by describing this *correspondence map*.

The chi-square statistic, sometimes written as χ^2, has an important connection with correspondence analysis. The points in the correspondence map result from a transformation of a space defined by a distance formula that is related to the chi-square statistic. This transformation makes it easier to calculate distances between points by applying the usual Euclidean distance formula.

$$Euclidean\ Distance = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

In the following example, we will demonstrate the distance calculations for the columns of a table, but these calculations will apply equally to the rows of a table. The ordered set of proportions for a particular row or column is defined as its *row profile* or *column profile*[28]. For example, in Table 1 the second column profile (from left to right), is [.500 .167 .333], where the three elements of this profile add to one.

One interpretation for the chi-square statistic in Table 1 is provided by looking at its column profiles. If the rows and columns are independent (i.e., <u>not</u> associated), then we would expect the three column profiles in Table 1 to approximate the average column profile shown in the last column, and therefore, for the three column profiles to be approximately equal to each other (i.e., by comparing two profiles

element by element). In like manner, we would expect the three row profiles to be approximately equal to the average row profile and, therefore, to each other.

The chi-square statistic, as applied to the counts or frequencies in a table, is a summation of terms (see Appendix D). Each term is calculated by squaring the difference between an observed count and its expected value, and then dividing by the expected value. In correspondence analysis, the form of the distance measure is similar. For example, under the assumption of independence, the expected value for the column profiles is the average column profile. Therefore, the 'chi-square distance' of the second column profile to the average column profile in Table 1 is:

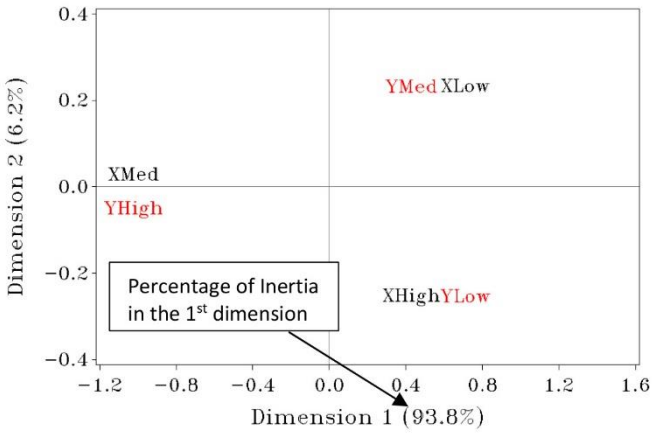
$$Distance = \sqrt{\frac{(0.5-0.294)^2}{0.294} + \frac{(0.167-0.353)^2}{0.353} + \frac{(0.333-0.353)^2}{0.353}} \approx 0.49$$

We can extend this definition to define the distance between two column profiles, or between two row profiles. For example, the chi-square distance between the first and third column profiles of Table 1 is:

$$Distance = \sqrt{\frac{(0.4-0.0)^2}{0.294} + \frac{(0.0-0.833)^2}{0.353} + \frac{(0.6-0.167)^2}{0.353}} \approx 1.74$$

The squared differences for the components of these two column profiles are divided by the elements of the average column profile, which represents the expected value for the column profiles under the assumption of independence. The chi-square distance formula can also be applied to pairs of rows, where the expected values in the formula are then given by the elements of the average row profile. (Note that the chi-distance can be described as a weighted Euclidean distance, where the weights are the elements of either the average column profile or average row profile.)

Figure 2. The correspondence map for Table 1, which shows the labels for the three row points in black and the labels for the three column points in red. The quantities in parentheses along the axes show the percentage of the total inertia for each dimension.



The correspondence map in Figure 2 represents a geometric transformation of the data in Table 1 that centers the average column and row profiles at the origin with coordinates (0,0). Applying the usual Euclidean distance formula from the origin (0,0) to the 'YMed' point with coordinates (0.430,0.242) in Figure 2 also yields the first distance calculation above of 0.49. (Note that these coordinates are obtained from a SAS printout not shown here.)

$$Euclidean\ Distance = \sqrt{(0.430 - 0)^2 + (0.242 - 0)^2} \approx 0.49$$

Applying the Euclidean distance formula to the coordinates (also obtained from a SAS printout not shown here) for the 'YLow' and 'YHigh' points in Figure 2 yields the second distance calculation above of 1.74. Therefore, an analyst who views the correspondence map of Figure 2 can approximate the relative sizes of the chi-square distances of the profiles for Table 1.

The correspondence map produced by the POSSE macros is called a 'symmetric' map. In this map, the chi-square distances between any two column points (i.e., which represent two column profiles), as well as between any two row points, are meaningful indicators of differences between categories. Although the distances between the row points and column points in Figure 2 do not have the same fixed meaning, there is a relatively intuitive interpretation available for symmetric maps. When a row point and a column point both lie in the same direction and away from the origin, it is often the case that an association exists for the underlying categories. For instance, if a row point and column point lie in the same general direction and are far from the origin in a map, their associated categories will tend to co-occur more often than we would expect by chance alone. A row point and column point which are on opposite sides of a map will often imply that an outcome for the row category tends to be present more often when the outcome for a column category is absent, and vice versa. When variables are binary, these results will often imply a positive or negative correlation between the variables. When two variables are associated with different axes, with their respective labels lying close to only one of the axes, this will often imply that the variables are relatively uncorrelated or independent. For other situations, the overall association for two variables can depend on the configurations and the order of categories in the correspondence space. In Section 5.4, the status of associations suggested in a map will be quantified when we examine the numerical quantities that are produced by the POSSE macros.

The next two examples present results for modifications of Table 1. The first example of Table 2 demonstrates how proportional changes to Table 1 can leave the correspondence map essentially unchanged, even though the chi-square values might change. Table 2 is obtained by multiplying all the counts in Table 1 by three and then splitting the 'YMed' category proportionally, where two-thirds of the counts are now in one category and one-third in the other. This yields the new 'YMed1' and 'YMed2' column profiles in Table 2, which have identical column profiles of [.500 .167 .333] and are identical to the column profile for the 'YMed' column of Table 1. Figure 3(a) is the correspondence map for Table 2, which is essentially unchanged from Figure 2, because 'YMed1' and 'YMed2' are located at the same position of the original 'YMed' category. This result illustrates the principle of distributional equivalence, which is unique to correspondence analysis. If we think in reverse by going from Table 2 to Table 1, then this principle implies that, if two column profiles are identical, the columns can be combined without affecting the row points in the correspondence map. Likewise, two identical row profiles would allow us to combine the rows without affecting the column points. Readers will learn how this principle operates in a homogeneity analysis in Section 6.2.

The second example of Table 3 demonstrates what happens when we change a single cell of Table 1. Table 3 is obtained by multiplying only the middle cell of Table 1 by ten. In this case, both the middle row and middle column profiles have now changed from those in Table 1. The row profile of [0 .167 .833] for the 'XMed' row in Table 1 has changed to [0 .667 .333] in Table 3, and the column profile for the 'YMed' column has changed from [.500 .167 .333] to [.200 .667 .133]. When we examine the correspondence map for Table 3 in Figure 3(b), we see that the 'XMed' and 'YMed' labels are now closer to each other, and the 'YMed' label is closer to the first axis. The 'XMed' and 'YMed' labels are also closer to the origin, which suggests that our multiplication has caused these row or column profiles to

contribute more to the determinations of the average row or column profiles.

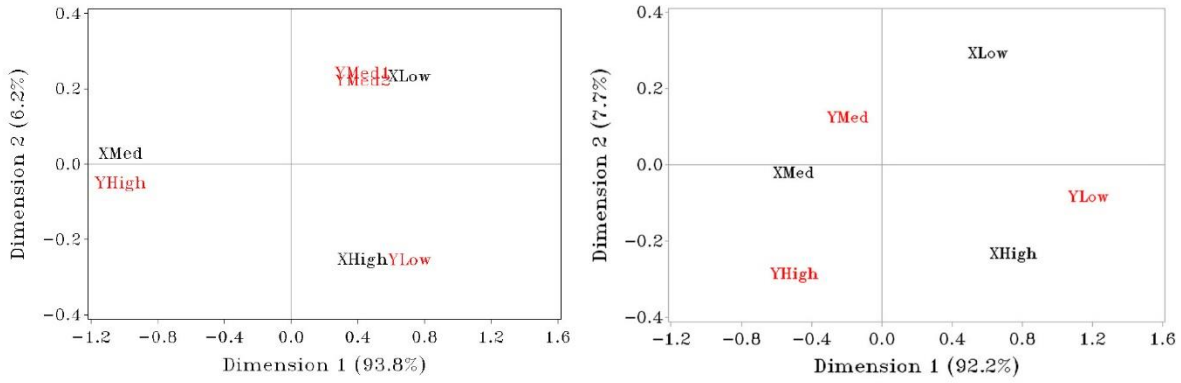
Table 2. The table which results from multiplying the counts in Table 1 by three and then splitting the middle Y category of Table 1 proportionally, where two-thirds of the counts are in the first new category and one-third in the second new category. The average row profile and average column profile are shown in bold. (Note that χ^2=31.7 for this table.)

Category Frequencies	YLow	YMed1	YMed2	YHigh	Total	Row Profiles Elements
XLow	6	6	3	0	15	0.400 0.400 0.200 0.000
XMed	0	2	1	15	18	0.000 0.111 0.056 0.833
XHigh	9	4	2	3	18	0.500 0.222 0.111 0.167
Total	15	12	6	18	51	**0.294 0.235 0.118 0.353**
Column Profile Elements	0.400 0.000 0.600	0.500 0.167 0.333	0.500 0.167 0.333	0.000 0.833 0.167	**0.294** **0.353** **0.353**	

Table 3. The table that results from multiplying the one cell frequency found in the middle row and middle column of Table 1 by ten. The average row profile and average column profile are shown in bold. (χ^2=9.7 for this table.)

Category Frequencies	YLow	YMed	YHigh	Total	Row Profiles Elements
XLow	2	3	0	5	0.400 0.600 0.000
XMed	0	10	5	15	0.000 0.667 0.333
XHigh	3	2	1	6	0.500 0.333 0.167
Total	5	15	6	26	**0.192 0.577 0.231**
Column Profile Elements	0.400 0.000 0.600	0.200 0.667 0.133	0.000 0.833 0.167	**0.192** **0.577** **0.231**	

Figure 3(a)-(b). The correspondence maps for Tables 2 and 3.



Because the average row and column profiles in Table 3 have changed from Table 1, all the distances are altered to some degree. However, the chi-square distance between the 'XLow' and 'XHigh' labels changed very little from Figure 2 to Figure 3(b) (from about 0.56 to about 0.54), because the associated

row profiles are unchanged. Likewise, the chi-square distance between the 'YLow' and 'YHigh' labels changed very little (from about 1.74 to 1.69).

These examples support the following general perceptions concerning the correspondence maps obtained with the POSSE macros:

(a) The row profiles largely influence the relative positions in the map of the row points, while the column profiles influence the relative positions of the column points.
(b) The sizes of the counts (sometimes referred to as the 'masses') influence the coordinate system of the map and determine the relative positions of the points to the axes and to the origin.
(c) Larger counts that are found in one row or column of a table can have the effect of a kind of 'rotational inertia' that draws axes closer to the row or column points that are identified with those larger counts. It can also change the relative positions of the points with respect to the origin.

As suggested earlier, users will also need to refer to the numerical quantities that are generated by the POSSE macros, which will be presented in subsequent sections, in order to clarify some of the results that are suggested by a map.

One other concept concerning correspondence maps has to do with the *dimension of a solution*. Although a chi-square test for a two-way table with r rows and c columns is associated with $[(r-1) \times (c-1)]$ degrees of freedom, the solution for a correspondence analysis has *dimension* equal to the minimum of $(r-1)$ and $(c-1)$. Since the dimension for the solution for Tables 1 thru 3 are all equal to two, the total or 100% of the inertia or variation for these tables is captured in a two-dimensional correspondence map. In this case, the distances depicted in the maps of Figures 2 and Figures 3(a)-(b) represent the exact chi-square distances, so the maps can be regarded as exact two-dimensional views of the results. In other words, there is no projection involved. When the dimension of a solution is greater than two, the results are a projection onto a plane and the distances in those maps represent approximations of chi-square distances.

The percentages in parentheses along the axes in each of the maps in Figure 2 and Figures 3(a)-(b) give the percentages of the inertias for the dimensions. These percentages will be described in the next section, but the fact that these sum to 100% also indicates that all of the inertia is captured in the first two dimensions in these maps. The fact that over 90% of the inertia in Figure 2 is associated with the first dimension suggests that the association between the levels of X and Y is essentially one-dimensional, which is reinforced by the approximate fit of the points to a one-dimensional quadratic line in Figure 1.

The POSSE methods examine results up to the first three dimensions. In those cases where the second or third dimensions represent less than 10% of the inertia, the POSSE results for those dimensions can be difficult to interpret unless the sample size is very large (e.g., $n > 1000$). In practice, users can usually choose to ignore the results for a dimension that represents less than 10% of the inertia.

5.2 Correspondence Analysis as Pattern Discovery

The search for patterns in a correspondence map is aided by some numerical quantities obtained with correspondence analysis. The previous section has illustrated the connection between the chi-square statistic and distances in the correspondence map. The connection with the numerical quantities is somewhat more straightforward. For instance, the total *inertia for a table* is defined simply as the chi-square statistic divided by the sample size. The χ^2 value for Table 1 is 10.56, so the total inertia for Table 1 is $10.56 / 17 \approx 0.62$. The total inertia also has an interpretation in terms of the correspondence map. This quantity of 0.62 can be calculated as the weighted squared distances of the column points to the origin, where the weights are the elements of the average row profile. The same quantity of 0.62 can also be calculated as the weighted squared distances of the row points to the origin, where the weights are the elements of the average column profile. (Note that each weight in these calculations simply represents the proportion of the 17 total samples which is due to a particular column or row profile.)

The total *inertia for a table* can then be split into the smaller *inertias for the dimensions*. The presentation of the inertias in correspondence analysis is somewhat analogous to the role of a source table in an analysis of variance, where the sums-of-squares are used to indicate the proportion or percentage of the variation that is captured or explained by the various factors. In an analysis of variance, the treatment and other factors sums-of-squares represent the systematic variation, and the error or residual sum-of-squares represents an estimate of the unexplained or random variation. In simple correspondence analysis, the first few dimensions with relatively large inertias are generally associated with the systematic variation, while any remaining dimensions with much smaller inertias are connected with the unexplained or random variation.

The focus in the POSSE methods is often on comparing the relative proportions of the inertias for the variables and their categories, and on detecting associations among the variables for those first few dimensions with large inertias, rather than focusing on the actual sizes of the inertias in each dimension. During an analysis, the inertia for each dimension is then split into the absolute contributions, each of which represents the proportion of the inertia for that dimension which is due to a particular category. This proportion is calculated for each dimension and each category of the row and column variables of the table. This absolute contribution, which will be illustrated in Section 5.4, will henceforth be referred to as the *contribution to the inertia*.

5.3 Correspondence Analysis as Scaling

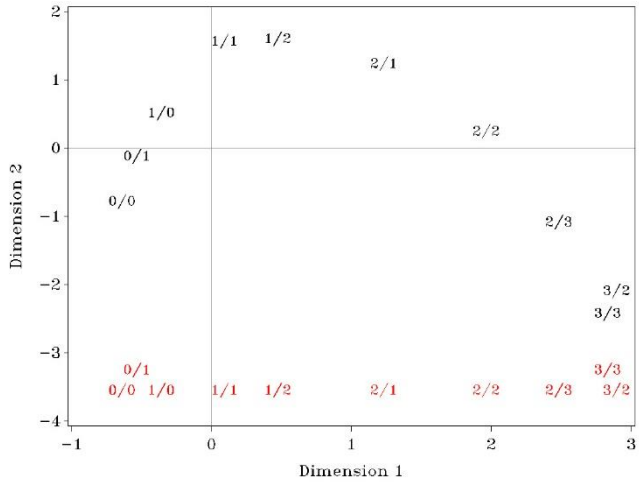
Let's suppose that we are specifically interested in how we might apply a scale to categorical data. For instance, if a variable has three categories, do we believe that the difference between the first and second categories has the same meaning as the difference between the second and third categories? Or is there some way that we can quantify our belief that the difference between the first two categories is not the same as the difference between the second and third categories? These questions lead us to a property of correspondence analysis that, as Friendly[29] has stated, can "provide an optimal way of transforming categorical water into quantitative wine."

Our example involves an unpublished data set which contains x-ray image ratings by doctors using a classification system for pneumoconioses[30] (i.e., silicosis and other dust-related lung diseases). This system was developed by the International Labour Organization (ILO) and includes a classification of 'profusion' which corresponds to the grading by a doctor of the frequency of small shadows in chest X-

rays. This profusion classification is composed of 12 ordered categories. Because of small frequencies, the bottom two categories are often collapsed into one category, and the top two categories are combined as well, resulting in 10 ordered categories, which range from the lowest profusion category of paired zeros '0/0' to the highest profusion category of paired threes '3/3'.

Figure 4 is a correspondence map of the ratings for over 200 images, each of which was rated by nine doctors. The labels near the top are at the original first-dimension and second-dimension coordinates for the 10 categories, but the order of the labels along the first dimension is emphasized here by projecting those labels along the bottom of the graph. The surprising result here is that there is nothing in the analysis that would indicate beforehand the ordering of these categories, and yet the correct order of the 10 categories is, for the most part, captured by correspondence analysis in the order of the category labels displayed along the first-dimensional axis. (Note that, in some correspondence analyses, an ordinal variable might be ordered from the highest category to the lowest category along the first dimension, instead of from lowest to highest.)

Figure 4. The correspondence map for the ratings of over 200 x-rays by nine doctors (unpublished data): original coordinates (in black) and the labels projected along the first dimension (in red).



The one misplacement in this analysis occurs for the last two categories '3/3' and '3/2', which are in the wrong order. It might be that the doctors have a difficult time distinguishing the high-profusion '3/3' and '3/2' categories. Furthermore, because the spacing is greater among the categories of '1/2', '2/1' and '2/2' then among the categories of '0/0', '0/1' and '1/0', we might also conclude that these doctors have an easier time distinguishing among the former categories than the latter. In terms of the x-ray images, this could imply that distinguishing between the '0/1', and '1/0' classifications is more difficult than distinguishing between the '1/2' and '2/1' classifications.

Although it is easier to see how this works for ordinal variables, correspondence analysis also can provide a scaling for nominal variables, such as race or gender, thereby allowing us to talk about the ordering of categories and the distance between categories for a nominal variable with respect to its perceived associations with other variables.

The pattern of the original coordinates in Figure 4 clearly suggests an arch, also sometimes referred to as a *horseshoe effect* in correspondence analysis. This often occurs with ordinal data in correspondence

maps. Greenacre[25] provides a number of explanations for this effect. An approach that uses the horseshoe effect in an analysis of binary variables will be presented in Section 6.3.

5.4 Applications to Contingency Tables

There are other possible applications for simple correspondence analysis[25, 28], but the most common applications are to two-way tables, also called contingency tables. This application is the focus of the POSSE methods.

At the end of Section 5.2, we described the process of determining the *contribution to the inertia* as first finding the *inertia for a dimension* and then splitting it into the components due to each category of the variable. The contribution to the inertia indicates how much a category is contributing to the inertia for each dimension. It should also be emphasized that the contribution for a category can be larger when its count is either lower or higher than would be expected, given the row and column counts under the assumption of independence.

A second way of comparing the relative proportions of the inertias is by first finding the total inertia due to a certain category over all the dimensions, and then splitting this value among all the axes or dimensions. The POSSE methods use the results of this process to calculate a statistic called the *quality*, which indicates how much of the inertia for a category is near the plane defined by the correspondence map. This quality is not shown numerically, but is incorporated visually into the POSSE maps by linking its size with the size of the label for a category, where a larger size for a label indicates a better fit for that category. The example which follows will demonstrate this feature. Those categories with better fits (i.e., with larger labels), along with larger contributions to the inertias given by the printed output, will be of particular interest during an analysis.

In a regression analysis, we often want to examine the effects of covariates. Such effects might be examined in categorical analysis using stratification. For example, suppose our results are represented by the following contingency table for the two variables X and Y, where the rows represent the two categories for the X variable, the columns represent the three categories for the Y variable, and the letters *a* – *f* represent the counts.

$$X\ levels \begin{matrix} \overbrace{\begin{matrix} a & b & c \end{matrix}}^{Y\ levels} \\ d & e & f \end{matrix}$$

If this table were stratified by a third variable Z with three categories or levels P, Q, and S, the resulting contingency table might look like the following, where each block represents a stratum defined by categories P – S of the third variable.

$$stratified\ table \begin{cases} \boxed{\begin{matrix} a1 & b1 & c1 \\ d1 & e1 & f1 \end{matrix}} & P \\ \boxed{\begin{matrix} a2 & b2 & c2 \\ d2 & e2 & f2 \end{matrix}} & Q \\ \boxed{\begin{matrix} a3 & b3 & c3 \\ d3 & e3 & f3 \end{matrix}} & S \end{cases}$$

This is an example of a stratified or 'multiway' table, and the analysis of such a table allows us to analyze

associations with respect to a third variable. In the stratified table above, we have split the counts in the cells of the first table in order to produce the stratified table. Therefore, the sum of all the cell counts in the stratified table will be equal to the original total frequency N. Alternatively, suppose that P, Q and S instead represented three binary row variables. We could then view the three blocks as three separate cross-tabulations of Y with the three binary row variables P, Q, and S. This would be an example of a 'stacked' table where the three cross-tabulations are stacked on each other, and where the sum over all the cells would then be equal to 3 × N.

It is also possible to present a mixture of stacked and stratified results in a contingency table. As an illustration, we use the Titanic data set (N = 2,201) which is analyzed by Friendly[29], where the variables include survival (1=died, 2=survived), gender (1=female, 2=male), age (1=child, 2=adult), and class (i.e., travel accommodations of 1=1st class, 2=2nd class, 3=3rd class, 4=crew). Table 4 is a cross-tabulation with the two stacked tables of age and class in the rows, and where the column variable of survival is stratified by gender. Since there are two stacked tables, the sum over all the cells is equal to 2 × N = 4,402. In other words, every subject is counted twice: once as either being a child or adult, and again by class of service. (Further guidance about the use of stacked and stratified tables are provided in Section 9.2 and the examples of Appendix C.)

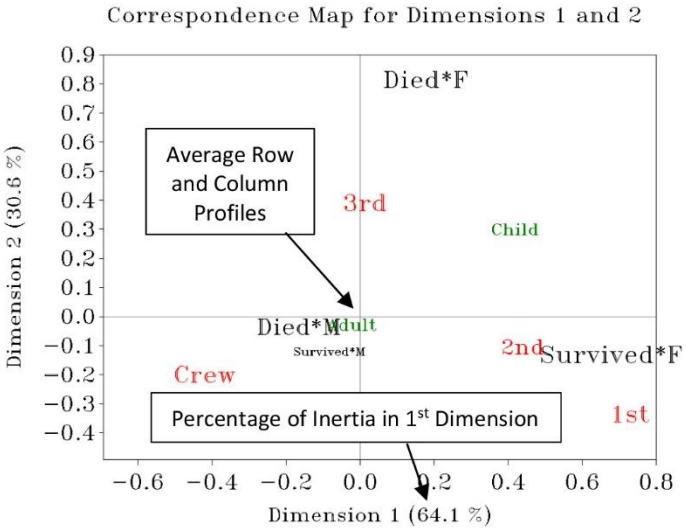
Table 4. The cross-tabulation of the Titanic data[29], where the categories of the *age* and *class* variables are stacked in the rows of the table, and the categories of the *survival* variable are stratified by those of the *gender* variable in the columns.

	Female		Male		
	Survived	Died	Survived	Died	Total
Child	28	17	29	35	109
Adult	316	109	338	1329	2092
1st Class	141	4	62	118	325
2nd Class	93	13	25	154	285
3rd Class	90	106	88	422	706
Crew	20	3	192	670	885

As mentioned in Section 5.2, the dimension of the correspondence solution for a given contingency table is equal to min($r - 1$, $c - 1$), where r and c are the numbers of rows and columns, respectively. Therefore, the dimension for the solution for Table 4 is min(6 – 1, 4 – 1) = 3. The correspondence map for the first two dimensions for Table 4 can be seen in Figure 5. The information within parentheses along the axes indicates that the first dimension is capturing about 64% of the inertia and the second dimension is capturing about 30% of the inertia for this table. Average row and column profiles are represented by the origin of the map. An examination of the table of contributions shows that over 78% of the contributions to the first-dimensional inertia for the stratified outcomes (i.e., the columns of the table) is due to the surviving females. Over 80% (0.428 + .386) of the contribution to the first-dimension inertia for the other variables (i.e., the rows) is due to the first-class and crew travel accommodations. Figure 5 suggests this result by plotting the labels for the surviving females and first-class accommodation at one end of the first-dimensional axis, in contrast to the label of the crew accommodation at the other end of the axis, where this label is closer to the label representing the males who died. The second dimension draws attention to the relatively high mortality among females with 3rd class accommodations. Although the

categories for the two row variables are not independent and so the degrees of freedom would be incorrectly specified for a chi-square test of independence, the results for this stacked and stratified exploratory analysis — augmented by the sizes of labels — provide a clear depiction of the relative strengths of the associations.

Figure 5. The correspondence map and contributions to the inertias for Table 4. Sets of contributions to the inertias are given for both the stratified outcomes (in the columns of Table 4) and the other variables (in the rows of Table 4). Each set of contributions sum to one for each dimension. The percentages along the axes suggest the percentages of explained variable (similar to the role of the R^2 statistic in regression analysis). The size of the labels is related to the fit for a category, where a bigger size indicates that more of its inertia is near the plane defined by the correspondence map. (Except for the text boxes, the figure and output were produced using the ***correspondence*** macro described in Section 8.3.)



Contributions of Outcomes to Inertia

Variable Levels	Contrib. to 1st Dim.	Contrib. to 2nd Dim.
Died*F	0.020	0.903
Died*M	0.184	0.005
Survived*F	0.784	0.045
Survived*M	0.013	0.047

Contributions of Other Variables to Inertia

Variable Levels	Contrib. to 1st Dim.	Contrib. to 2nd Dim.
1st	0.428	0.169
2nd	0.136	0.011
3rd	0.000	0.611
Crew	0.386	0.151
Child	0.047	0.055
Adult	0.002	0.003

Three basic numerical quantities have been presented for the simple correspondence analysis of a table. These can be described as follows.

(a) The total inertia for a table, which is defined as the chi-square value for the table divided by the sample size.
(b) After the total inertia for a table is split into the smaller inertias for the dimensions, these smaller inertias divided by the total inertia. The results, found along the axes of the correspondence maps, are expressed as the percentages of the total inertia for the various dimensions.

(c) After the inertia for each dimension is split into the still smaller inertias due to the various categories of the variables, they are then divided by the inertia for the dimension. The results, called the contributions to the inertias, are interpreted as the proportions of the inertia for a dimension which are due to the various categories. Tables of contributions to the inertia for each dimension are presented separately for the row and column variables.

6. Homogeneity Analysis

Greenacre[28] introduces some extensions of simple correspondence analysis to various approaches to 'multiple' correspondence analysis. One approach, which takes advantage of the *scaling* properties discussed in Section 5.3, is *homogeneity analysis*. Like simple correspondence analysis, homogeneity analysis can be used for pattern discovery. Additionally, we will see that it provides tools for the classification or clustering of observations and variables. One method of classifying variables will rely on the *Guttman scaling* and *horseshoe effect* that were introduced in Sections 4.2 and 5.3.

An important concept here is the discrimination provided by a variable when one wishes to find clusters or groupings of similar observations. To illustrate this, suppose that we would like to use the answers to the following four questions in order to predict the gender of individuals.

Question 1: 'Do you like to watch public television?'
Question 2: 'Are you a vegetarian?'
Question 3: 'Do you smoke cigarettes?'
Question 4: 'Are you a hockey fan?'

During this process we might ask which responses help us most to predict an individual's gender. In other words, which responses provide the best discrimination with respect to gender? Later, we might ask what combinations of these responses can best boost our ability to predict the gender, and whether there is some way to weight the responses in order to maximize our ability to predict the gender.

We can extend the scenario above to the situation where, given the results for a set of variables, we want to use a model to predict the gender of a subject. This leads to a multivariate method called discriminant analysis.[19] However, in contrast to this, the objectives for our applications of homogeneity analysis will be more open-ended and more exploratory. Although an analysis might begin by initially choosing a subset of variables using subject-matter information, homogeneity analysis will be employed in order to identify a subset of variables that are particularly good at separating the subjects into clusters without regard to their associations with any other variable. Therefore, in contrast to the process described above, the strategy using homogeneity analysis for the example above would be to find clusters based on the four questions and then, only afterwards, explore whether the resulting clusters were associated with gender or any other variables. (Those readers who are familiar with data-mining methods will recognize discriminant analysis and our use of homogeneity analysis as forms of 'supervised' and 'unsupervised' learning.[21])

6.1 Basic Principles

Although the computational procedure for homogeneity analysis is similar to that for simple correspondence analysis, it involves new derivations, interpretations and applications. For example, in homogeneity analysis the *inertia* will indicate the ability for a dimension to provide good discrimination for an exploratory clustering of the observations. While simple correspondence analysis is applied to tables, homogeneity analysis is applied to a data matrix, where the columns represent the categories of the variables and the rows represent the subjects or observations. In addition, the data matrix is comprised of zeros and ones (sometimes called an 'indicator matrix') rather than of frequencies. To illustrate how this indicator matrix might be constructed, suppose a subject is classified into the fourth of five categories for variable X, the first of three categories for variable Y, and the third of four categories for variable Z. The information could then be presented as follows.

X	Y	Z
4	1	3

The same information could also be presented as the following ordered pattern of ones and zeros, where the ones indicate the categories into which the subject's variables have been classified.

X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Z_1	Z_2	Z_3	Z_4
0	0	0	1	0	1	0	0	0	0	1	0

It can be shown that simple correspondence analysis can determine a linear combination of the column variables that maximizes the discrimination among the row categories, in the sense of providing a maximum dispersion of the row points in a correspondence map that distinguishes the categories in an optimal way[16, 27]. (As the X-ray example in Section 5.3 demonstrates, the resulting scaling can also provide an informative ordering and spacing of categories in a map.) We can extend this idea to homogeneity analysis, which can then determine a linear combination of the variables (i.e., in the columns) which maximizes the discrimination among the observations with different profiles (i.e., in the rows). This is sometimes referred to as an *optimal score* (i.e., an optimal linear combination of variable categories) which can be applied to the observations. In addition, each dimension has a separate optimal score. Therefore, because the POSSE methods calculate the results for the first three dimensions, each observation will have a score for each of the three dimensions.

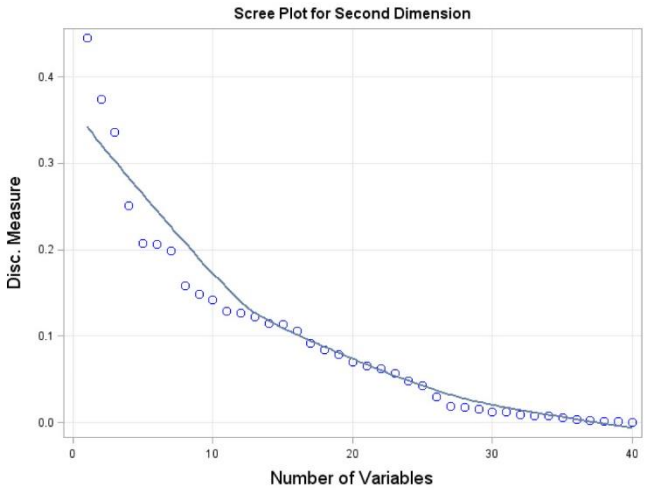
The optimal scores can be used to determine the coordinates for the observations in a *homogeneity map*. Some properties are very similar for simple correspondence maps and homogeneity maps. For example, if the categories for two variables (i.e., in the columns of the data matrix) have labels which are in close proximity in a homogeneity map, they will often be correlated. Although the observations can be seen in a separate plot during the cluster analysis to be described in the next section, only the variable categories can be seen in the homogeneity maps produced using the POSSE software. However, if the observations were also plotted in a homogeneity map, we would find that the distances between observations are related to the similarities of their profiles of zeros and ones.

Other properties or features for the maps produced using simple correspondence analysis and homogeneity analysis are quite different. As was mentioned in Section 5.1, the POSSE methods produce a symmetric map for simple correspondence analysis where the distances between row and column variables are not clearly defined. However, the map that the POSSE methods produce for homogeneity analysis is an asymmetric one where there is an unambiguous connection between the variables and observations. For instance, during a cluster analysis each observation will be located at the average position of all the categories to which it belongs.

In simple correspondence analysis, the maps depict both row and column variables and their interactions. In homogeneity analysis, there is only one set of variables, where the spread of its categories in each dimension is related to the size of the inertia for that dimension. A single variable generally provides good discrimination when its categories are spread far apart and far away from the origin in a homogeneity map. The degree of discrimination possessed by a variable (over all its categories) to a given dimension is called its *discrimination measure* and the mean of the discrimination measures for all the variables is equal to the inertia for that dimension. Therefore, the size of the inertia for a given dimension will provide an overall indication of the degree of discrimination for the variables that are contributing to that dimension. Because a POSSE analysis examines the first three dimensions, each variable will have three discrimination measures, one for each dimension.

Discrimination measures can be interpreted in terms of squared correlations, but in the next section we will focus on the ability of homogeneity analysis to identify variables with larger inertias which have their categories further apart and, therefore, provide better discrimination for the observations[13]. This will sometimes allow us to find meaningful groupings of the observations using a clustering method.

Figure 6. A scree plot from a homogeneity analysis using part of the data which are analyzed in Section C.13 of Appendix C. The plot shows the discrimination measures, plotted from highest to lowest, for 40 variables. Note that, unlike the contributions to the inertia, a discrimination measure is not a proportion and is calculated over all the categories of a variable. (This figure was produced using the ***classification*** macro described in Section 8.4.)



One tool that is used to evaluate the discrimination measures is the *scree plot*[21], which plots the estimated discrimination measures (on the y-axis) against the order—largest to smallest—of the discrimination measure for each variable. An example can be seen in Figure 6. The location of the 'elbow' in scree plots is sometimes used to indicate the number of variables to be retained for further analysis. In Figure 6, the location of the elbow suggests that the seven variables with the largest discrimination measures, which are all approximately ≥ 0.20, should be retained for further analysis. In some cases, a subset of variables might be chosen when their discrimination measures are clearly much larger than the remaining variables, even when this doesn't involve the elbow.

(Technical Note: More information about homogeneity analysis and its maps can be found in Michailidis and de Leeuw[13] and Greenacre[28].)

6.2 Classifying Observations

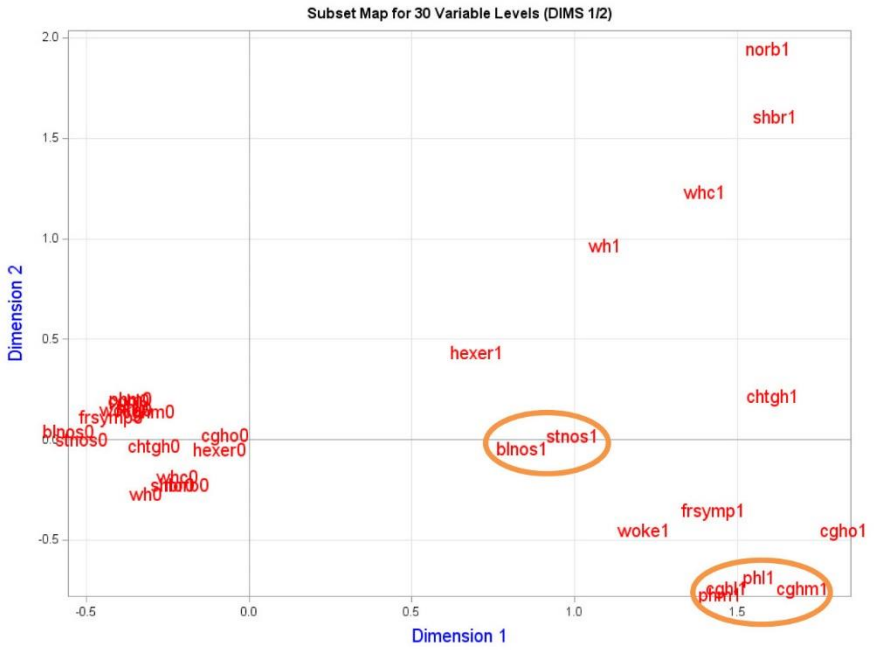
A cluster analysis can sometimes be applied to the observations during homogeneity analysis (examples can be found in Sections C.7 and C.13). A cluster is a group of similar subjects or observations. In the context of homogeneity analysis, two observations are similar if their profiles of ones and zeros are similar. As mentioned in the previous section, the homogeneity maps plot only the variable categories. However, the assignment of observations to particular clusters relies on the distances between the unseen points that represent these observations in a homogeneity map. These distances, in turn, are influenced by how far apart the categories are in the map, which is related to the discrimination measures. Therefore, the variables with large discrimination measures will generally have more influence on the formation of clusters.

Here we present a general outline for applying a cluster analysis using the POSSE methods. Oftentimes, the objective of the cluster analysis will be to develop a cluster variable to be used as the response variable in a simple correspondence analysis. This cluster variable will be interpreted in terms of the *profile for a cluster*. For example, suppose that the clusters have been formed using three binary symptom variables. In this case, a cluster profile of [.05 .85 .50] for a particular cluster would indicate that 5% of the cluster members are positive for the first symptom, 85% positive for the second symptom, and 50% positive for the third symptom. (Note that, unlike a row or column profile, a cluster profile does not necessarily sum to one.)

There will generally be three steps in the development of a cluster variable, including (1) the removal of irrelevant variables, (2) the removal of redundant variables, and (3) the formation of the clusters using the remaining variables. The objective of the first step is to retain those variables that are contributing the most to the discrimination of the observations by examining the scree plots and some additional results. The second step performs *subset homogeneity analysis*, a type of subset correspondence analysis[12]. This is a technique was developed to provide more consistency between various possible subanalyses involving parts of a table by maintaining the marginal totals of the original table.

The homogeneity map for the asthma study example in Section C.13 can be seen in Figure 7. We have already discussed the principle of distributional equivalence in Section 5.1 with respect to identical profiles. In homogeneity analysis, this principle implies that, if the profiles of two variables are identical, they can be combined without affecting the geometry of the observations. It generally follows from this principle that two correlated variables can sometimes be replaced by one of the variables without large distortions in the results. The practical consequence of this is the ability to choose a single variable from among a cluster of correlated variables, thereby removing redundant variables from the cluster analysis. The ellipses in Figure 7 identify two collections of correlated categories. The ellipse with the two categories labeled 'stnos1' and 'blnos1' are, respectively, associated with positive responses for the questions related to whether the subject has (a) 'a stuffy nose or drainage at the back of the nose' or (b) 'a blocked, itchy and runny nose.' By comparing their discrimination measures, an analyst can choose and retain one variable from among a collection of redundant variables. This reduction in the number of variables before the formation of clusters will make it easier to interpret the clustering results.

Figure 7. A map from a subset homogeneity analysis for a collection of symptom variables[48]. (Except for the ellipses, this figure was produced using the ***classification*** macro of Section 8.4.)



The last step determines the number of clusters by first applying homogeneity analysis to the remaining variables to generate the row coordinates (i.e., the coordinates for the observations) for the first three dimensions, and then using these coordinates in a K-Means cluster analysis[21]. K-Means clustering is a type of 'seed-point' clustering. It starts with a random selection of observations and, by using an iterative reassignment process, results in clusters where the within-cluster variation has been minimized.

The number of clusters can be determined by examining the resulting values for the *cubic clustering criterion* (CCC) developed by Sarle[31], who also provides the following criteria for using the CCC statistic:

(a) The average number of observations per cluster should be at least 10.
(b) A maximum CCC statistic which is greater than 2 indicates good clustering.
(c) A maximum CCC statistic between 0 and 2 indicates possible clusters which should be interpreted cautiously.
(d) Very negative CCC values (e.g., -30) might be due to outliers.
(e) When the CCC values continue to increase with increasing number of clusters or when the CCC values are all negative, the clusters are not well-defined.

Once the number of clusters is chosen, the POSSE methods will produce a plot of the observations that identifies their cluster membership. A figure is also produced which can identify outliers among the clusters. Note that the removal of all outliers could be accomplished by simply continuing to increase the number of clusters, but this does not necessarily produce useful clustering results or maximize the CCC statistic. The purpose of the outlier information is to identify extreme outliers, which an analyst might decide to remove from an analysis.

(Technical note: In contrast to our application of *subset homogeneity analysis*, subset correspondence analysis is usually applied in a more general way in order to look at a subset of categories instead of a subset of variables. Also, for those readers who are familiar with multivariate methods, the process of

reducing the variables before clustering is somewhat analogous to a preconditioning procedure that is applied sometimes during spectral clustering[32].)

6.3 Classifying Variables

The idea of classifying or clustering variables is not a new one, but it is less commonly used in multivariate analyses. The POSSE methods also apply this idea in a broader and more informal way than is usual. The POSSE methods provide two ways to conceptualize the classification of variables. The first way is found in the second step in the cluster analysis of observations presented in the previous section, and illustrated by Figure 7. In that figure, clusters of correlated categories suggested that the binary variables were correlated, and could be grouped together.

The second way of classifying variables is less straightforward and has to do with how they depend on one another. This second way is connected with the concept of a Guttman scale that was described in Section 4.2. It extends the scaling ideas in Section 5.3 where we discussed some results using an ordinal variable with the resulting *horseshoe effect* of Figure 4. When a horseshoe effect is evident in a map resulting from a homogeneity analysis, it might imply an ordinal structure for a group of variables.

An example comes from Weller and Romney[10], who use data from Kay[33] to show the ownership of seven consumer goods for 40 French Polynesians: (a) a small stove, (b) a large stove, (c) a bicycle, (d) a motorcycle, (e) a refrigerator, (f) a radio, and (g) a car or automobile. When a homogeneity analysis is performed for these seven binary or dichotomous variables, a POSSE macro submission produces the map in Figure 8 for the first two dimensions. In the figure, those variable labels ending with a '1' represent the categories of ownership for the consumer goods and those ending with a '0' indicate the categories for non-ownership. Note that a 'horseshoe effect' is evident in the arched display of the labels. A POSSE macro submission is also used to create Table 5, a tabulation of the 14 unique profiles of ones and zeros which were observed for the 40 participants, where the blank cells in the table represent the zeros. The first line of Table 5 illustrates that there are two participants who possessed all seven consumer goods, whereas the last line indicates that one participant had none of them.

The fact that the ones are mostly found in the upper triangular portion of the matrix is an indication that many of these variables follow either a Guttman or quasi-Guttman scale. In the context of this study, the results could be interpreted in terms of the order of acquisition. For example, it would be unlikely for a person to purchase an automobile if that person didn't already own a refrigerator. On the other hand, there are several persons who own a large stove, but not a motorcycle, so that dependency is less clear. In Appendix C, we will see an example of a subset of variables which generally follow Guttman or quasi-Guttman scales, where a pattern of ones is evident (although not necessarily in upper or lower triangular portions of a matrix). Also, as suggested in Section 4.2, we will see that it can be useful to combine a set of such variables into a new variable.

(Technical note: For those interested readers, the scaling results illustrated in this section are also related to a topic in ecological research called ordination[34].)

Figure 8. The map from a homogeneity analysis of the consumer goods data[33]. This shows a 'horseshoe effect' or 'arch effect' for the binary variables which indicate the ownership for seven types of consumer goods. (This was produced using the ***classification*** macro described in Section 8.4.)

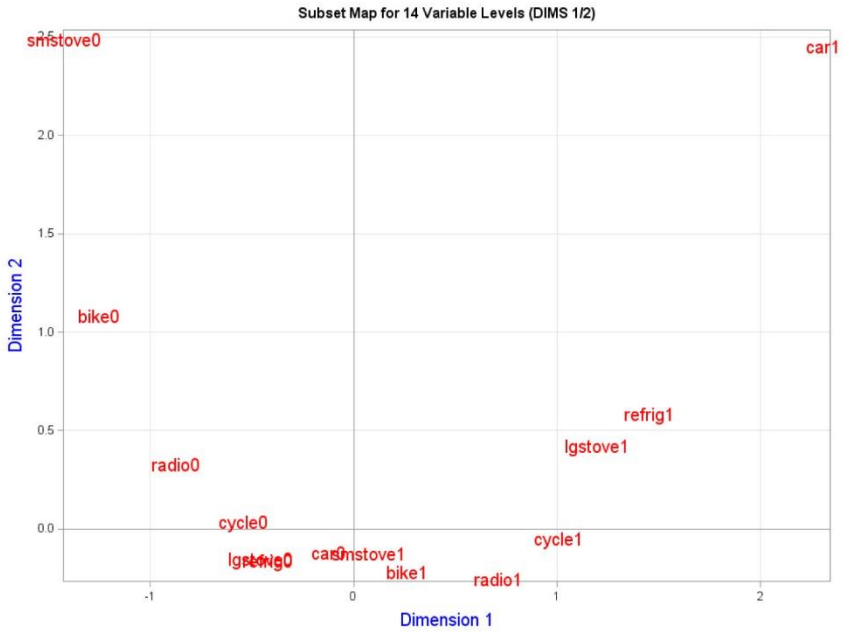


Table 5. The 14 unique profiles of ones and zeros for the 40 participants in the consumer goods study[33], where the blank cells represent the zeros.

Profiles	*small stove*	*bike*	*radio*	*cycle*	*large stove*	*refrig.*	*car*	Frequency
1	1	1	1	1	1	1	1	2
2	1	1	1	1	1	1		3
3	1	1	1		1	1		2
4	1	1	1	1		1		2
5	1	1	1	1	1			2
6	1	1	1		1			1
7	1	1	1	1				4
8	1	1			1			1
9	1	1	1					6
10	1	1		1				1
11	1	1						8
12		1						1
13	1							6
14								1

6.4 Summary for Homogeneity Analysis

Some of the key ideas for homogeneity analysis that were covered in this section can be summarized as follows:

(a) In homogeneity analysis, the relative sizes of the inertias for the dimensions indicate the degrees of discrimination provided by the variables that contribute to a dimension.

(b) The contribution of a variable (over all its levels) to a given dimension is called its discrimination measure.
(c) The mean of the discrimination measures for each dimension is equal to the inertia for that dimension. The POSSE methods produce results for the first three dimensions, where each dimension has its inertia and its corresponding set of discrimination measures.
(d) Homogeneity analysis can be combined with a clustering procedure.
(e) Scaling results from a homogeneity analysis can sometimes identify an ordinal structure in groups of binary variables which follow Guttman or quasi-Guttman scales.

7. A Special Application for Deriving Categorical Variables

To motivate the presentation in this section, we examine the data from an example from Weller and Romney[10] which come from a study conducted by Srole[35]. This study examines the relationship of the mental-health status of 1,660 offspring to the socioeconomic class of their parents. As can be seen in Table 6, the row percentage of subjects with mental impairment is about twice as large for the 'Very Low' socioeconomic class as it is for the 'High' or 'Very High' classes (i.e., 33% versus 16-18%). However, the profile of percentages for the two highest categories are very similar, so it might make sense to combine these categories. How do we decide whether or not to collapse or combine those two categories of economic status, or if other rows in the table can be combined as well? A cluster analysis method developed by Greenacre[28, 36] is applied in the POSSE methods to help make these determinations.

Table 6. The frequencies and approximate row percentages for a cross-tabulation of mental health categories with economic-status categories[35].

Frequencies (Approximate Row Percentages)	**Mental Health**				
	Well	Mild	Moder-ate	Im-paired	**Totals**
Economic Status					
Very High	64 (24%)	94 (36%)	58 (22%)	46 (18%)	262
High	57 (23%)	94 (38%)	54 (22%)	40 (16%)	245
Above Average	57 (20%)	105 (37%)	65 (23%)	60 (21%)	287
Below Average	72 (19%)	141 (37%)	77 (20%)	94 (24%)	384
Low	36 (14%)	97 (37%)	54 (20%)	78 (29%)	265
Very Low	21 (10%)	71 (33%)	54 (25%)	71 (33%)	217
Totals	307	602	362	389	1660

In this following, the terms of 'cluster' and 'category' and 'row of a table' will sometimes be used interchangeably. For example, a new category that results from combining two previous categories might be referred to as a cluster. Also, it is important to note that, unlike the usual clustering methods that assign individual subjects or observations to various clusters, this method assigns the subsamples represented by the rows of a contingency table, where the assignment is based on the 'chi-square distance' defined in Section 5.1. As recommended by Greenacre[36], Ward's minimum-variance method of clustering is applied to the estimated chi-square distances. Although the methods are otherwise identical, Greenacre

expresses his results in terms of the chi-square scale, whereas the POSSE results are expressed in terms of the between sum-of-squares which are calculated from Ward's method.

7.1 Interpreting the Cluster Tree and Cluster History

Ward's method provides an estimate of the R^2 statistic, which will be an important indicator of how much information is lost by reducing the number of categories for a variable. The R^2 statistic should be familiar to readers from the analysis-of-variance method and has a similar interpretation here. The two goals of a cluster analysis can very roughly be defined as seeking groups where (1) subjects from different groups are dissimilar and (2) subjects from the same groups are similar. Another way of expressing these two goals is to say that we would like (1) the between-group variation to be large and (2) the average variation within groups to be small. The POSSE approach quantifies the first goal in a between-cluster sum-of-squares statistic which is a measure of the distance between two clusters. The second goal is quantified in an estimate of a statistic called the 'Root-Mean-Square Total-Sample Standard Deviation' in the SAS output[37]. This statistic, which will hereafter be referred to it as the 'root-mean-square', estimates the average variation within each cluster. In practice, it might be necessary to strike a balance between minimizing the root-mean-square while minimizing the reduction in the R^2 statistic.

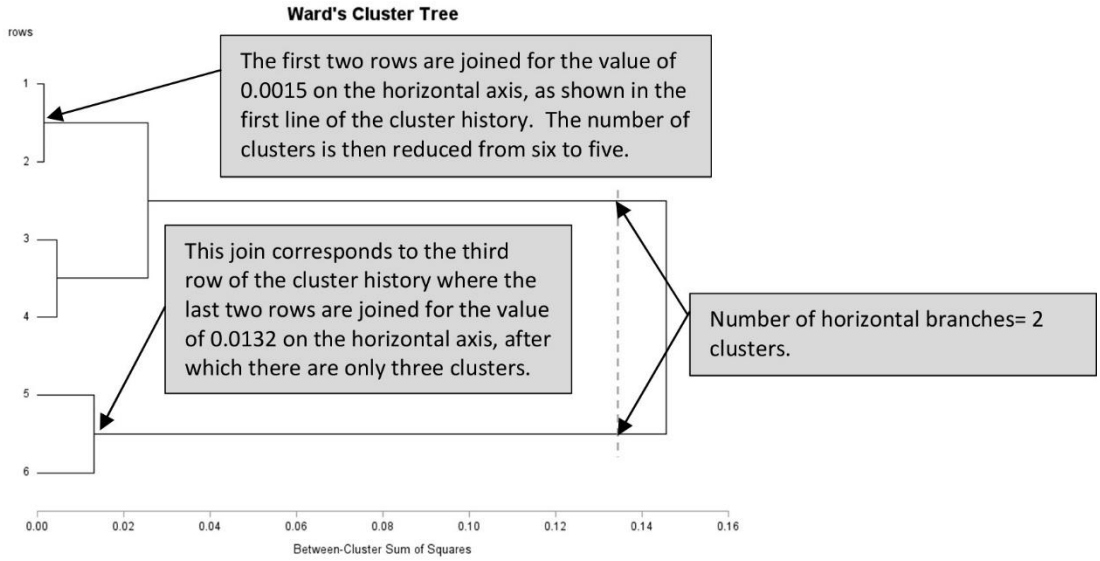
The clustering method used by the POSSE methods is an example of hierarchical clustering, where the process involves starting with the rows as single clusters and then merging clusters at successive stages until there is finally just a single cluster. Hierarchical clustering results can be displayed in a diagram called a cluster tree or dendogram. The cluster tree and cluster history seen in Figure 9 indicate the sizes of the between-cluster sum-of-squares at which rows or groups of rows of Table 6 are merged. Beginning at the extreme left of the cluster tree, where the between-cluster sum-of-squares is zero, we start with the original six clusters that are represented by the six rows (from top to bottom) in Table 6. Therefore, at the beginning of the process we have 'singleton' clusters, that is, each cluster is composed of one row.

It will be useful at this point to picture an imaginary vertical line which is moving from left to right along the cluster tree in Figure 9. When that vertical line reaches a between-cluster sum-of-squares value of 0.0015 on the horizontal axis, then the first two clusters are joined. These two clusters have the smallest between-cluster sum-of-squares, so they are merged first. This merge corresponds with the first row of the cluster history in Figure 9, where the R^2 value of 0.992 indicates that over 99% of the explained variability is retained when we reduce the number of clusters from six to five. (The example in Section C.2 will demonstrate how the root-mean square comes into play during this cluster analysis.)

After the first join, we have four singleton clusters and one cluster with two members. The two clusters to be joined next are the third and fourth clusters which, as the second line of the cluster history illustrates, have the next smallest between-cluster sum-of-squares equal to 0.0046. The R^2 statistic now indicates that about 97% of the explained variability is retained when we reduce the number of clusters from the original six clusters to four clusters.

This process continues until all of the clusters are joined into one cluster, signified by the farthest right-hand vertical branch in the cluster tree, which lines up with the value of 0.1458 on the horizontal axis. The number of clusters at any point in the process can be determined by how many horizontal branches are intersected by our imaginary vertical line. For example, a vertical line pictured at a between-cluster sum-of-squares value of 0.02 corresponds to the formation of three clusters at that point in the clustering process.

Figure 9. The cluster tree and cluster history for the six categories of socioeconomic class[35], which are represented by the rows of Table 6. The joining of clusters 1 and 2 occurs at a value of 0.0015 for the between-cluster sum-of-squares, and the corresponding R^2 of 0.992 indicates that over 99% of the explained variability is retained when going from six to five clusters. (Except for the text boxes, output was produced using the ***data_prep*** macro described in Section 8.1.)



Root-Mean-Square Total-Sample Standard Deviation	0.097635

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Between Cluster Sum of Squares
5	**1**	**2**	2	0.0081	**.992**	0.0015
4	**3**	**4**	2	0.0242	**.968**	0.0046
3	**5**	**6**	2	0.0691	**.899**	0.0132
2	**CL5**	**CL4**	4	0.1341	**.765**	0.0256
1	**CL2**	**CL3**	6	0.7645	**.000**	0.1458

7.2 A Method for Choosing Categories

Hierarchical clustering was used in the previous section to examine whether the number of categories for a known categorical variable can be reduced. However, the usual application of this feature using the POSSE methods will be in the derivation of a categorical variable from a quantitative variable during the data preparation. In cases without any prior information, we might simply choose to transform a continuous variable into a relatively large number of categories with approximately equal numbers of observations, and then see if we should reduce the number of categories when we examine its relationship to the response variable using the Greenacre[36] method. Generally speaking, although this is somewhat arbitrary, we will use a rule-of-thumb of reducing the number of categories for a variable when the R^2 statistic in the cluster history (Figure 9) still captures at least 90% of the explained variability for a variable with respect to the response variable. It is also possible that a researcher will want to determine the number of categories for a variable with respect to its relationship with more than one other variable. In cases where the resulting numbers of categories differ for the separate cluster analyses for two variables, the researcher could choose the larger number of categories, or try to find a compromise based on comparisons of the respective R^2 statistics and root-mean-squares. During the data preparation for a POSSE analysis, other factors can come into play when determining the number of categories for a variable. For example, subject-matter information might also be used to determine the

number of categories. Also, if the hierarchical clustering results suggest the combination of disjoint categories (such as first and last rows of Table 6), subject-matter information can be helpful in deciding whether such a combination is plausible or makes scientific sense.

The POSSE methods also provide an easy way to examine the sample distribution of a quantitative variable, and in cases where there are distinct modes in the distribution, we might decide to choose cut-points in the 'valleys' between the peaks which are indicated by the bar-chart or histogram for a variable. In still other cases, there might be ceiling effects or threshold effects in the distribution, that is, where observations bunch up at or near a maximum or minimum value. In these cases, we could decide to choose cut-points near those values.

III. SOFTWARE FOR THE POSSE METHODS

8. The POSSE Macros

The POSSE macros have been developed to provide the user with a variety of options and features. For example, the ***prelim_ca*** macro produces three correspondence analyses with some simplified output using one macro submission; similar output would require a user to make three separate submissions using SAS® procedures. Other features found in the POSSE macros are not as easily obtained in SAS. For example, the ***correspondence*** macro allows a user to show and highlight the observations in a simple correspondence map. This ability is not readily available in SAS. There are also some features in the POSSE macros which are absent in SAS, such as providing estimates of the *discrimination measure* that are given in the ***classification*** macro, or invoking the clustering procedure which is found in the ***data_prep*** macro.

Although the correspondence maps are identical for the ***prelim_ca*** and ***correspondence*** macros, the macros differ in their tabular summaries. The ***prelim_ca*** macro provides the total inertias of tables, whereas the ***correspondence*** macro provides the contributions to the inertias for the rows and columns of the tables. Tables 1 and 2 of Section 5.1 illustrate differences between the two macros. In Section 5.1, we saw that Tables 1 and Tables 2 have basically the same correspondence maps, and this is true whether we use the ***prelim_ca*** or ***correspondence*** macros. Because Table 2 has a sample size that is three times that of Table 1, their chi-square values are different, but applying the ***prelim_ca*** macro to each of them would indicate that the total inertias, as defined in Section 5.2, are identical for the two tables. However, the results using the ***correspondence*** macro would indicate that the contributions to the inertias for the various categories are not the same. The total inertias given by the ***prelim_ca*** macro tell us about the overall associations between the row and column variables, which is why the total inertias for Tables 1 and 2 are the same. The contributions to the inertias given by the ***correspondence*** macro tell us how the various categories of the row and column variables are influencing the associations. This point also has implications for the analysis of the stacked tables, which is further explained in Section 9.2.

Tables 7 through 13 summarize the main features of the POSSE macros and are meant to serve as a quick reference guide. Table 7 describes the basic features for each macro and Table 8 describes some common tasks accomplished by the macros and the location of corresponding examples in Appendix C. Tables 9 through 13 describe the parameters in each of the five macros. Appendix A provides additional information for storing and invoking the POSSE macros for those readers who are unfamiliar with SAS macros.

Note that the entries for the parameters are <u>not</u> case-sensitive (e.g., 'yes' or 'YES' or 'Yes' all work). All the variables specified in the POSSE macros must be numerical and categorical, with the exception of the ***data_prep*** macro where quantitative variables might be used to derive categorical ones. In order to successfully run the macros with SAS software, values and/or variable names initially must be inserted—after the equals sign—for each of the 'required' parameters (i.e., the ones preceded by a /**/ comment). The parameters that are not preceded by comment symbols are either 'optional' or might be required when a value for another optional variable is entered, or might be required the last time a macro is submitted in order to save a permanent SAS data set. Detailed descriptions of each parameter's function, as well as suggestions for when to include values for the optional parameters, are found in each subsection. Users should also check the SAS LOG for error messages if a POSSE macro generates unexpected or null results.

Table 7. The names and basic features for the five POSSE macros.

Macro Name	Features
data_prep	▪ Displays distribution plots for continuous variables ▪ Derives categorical variables from continuous ones using cut-point or ranking methods ▪ Performs a special type of cluster analysis which suggests an appropriate number of categories for a variable with respect to the response variable (where the output will include a cluster tree and a table with the cluster history) ▪ Outputs the final set of categorical variables to a data set for a POSSE analysis
classification	▪ Performs homogeneity analysis and subset homogeneity analysis (where the output will include homogeneity maps and tables of inertias) ▪ Produces scree plots during a homogeneity analysis which can indicate which variables are the best discriminators with respect to the observations ▪ Creates a cluster variable along with profiles which are used to characterize the clusters
prelim_ca	▪ Performs up to three simple correspondence analyses for (a) the response variable versus the explanatory variables, (b) the response variable versus the covariates, and (c) the explanatory variables versus the covariates (where the output will include up to three correspondence maps and tables of inertias) ▪ Clarifies potential confounding by identifying variables which are associated with the response and explanatory variables
correspondence	▪ Performs a correspondence analysis for the response variable versus the explanatory variables and covariates (where the output will include correspondence maps for up to a three-dimensional analysis and tables with the contributions to the inertia for the response variables and the other variables) ▪ Produces correspondence maps which can display both the variable levels and the observations ▪ Highlights observations in a map which are identified with a category of the response variable and/or with a category of a second variable ▪ Stratifies the response variable by another variable
tabulation	▪ Displays cross-tabulations and stratified cross-tabulations of variables ▪ Allows for the creation of complete tables for sparse data (i.e., by including the rows or columns for categories with no outcomes)

Table 8. Some common tasks addressed by the five POSSE macros and where they are illustrated in the examples of Appendix C.

Task	Macro Used	Example
Determining the Number of Categories	The number of categories can be determined for a categorical variable that is derived from a quantitative variable by using the clustering method found in the ***data_prep*** macro.	Section C.2
Identifying Dependencies	General dependencies/associations among ≥ 3 variables can be explored using the ***classification*** macro.	Section C.9
Stratifying the Analysis	The ***correspondence*** macro can be used to stratify the response variable by a covariate.	Section C.10
Creating a Cluster Variable	A cluster variable can be created from a set of variables by using the ***classification*** macro.	Section C.7
Identifying Unusual Observations	Unusual observations can be identified in a correspondence map by highlighting or circling observations using the ***correspondence*** macro.	Section C.6

Table 9. Summary for the ***data_prep*** macro parameters.

Parameters	
rawdata=	Name for the original data set (required for first submission of macro).
newdata=	Set to YES after creating a new categorical variable in order to begin saving any newly-created variables. (Required after creating a new variable).
contvar=	Name for continuous variable(s). Specify only one continuous variable here when deriving a categorical variable. (Required when deriving a categorical variable.)
plotdist=	YES yields the distribution plot(s) for the continuous variable(s) listed above (optional).
catname=	Name for new categorical variable created from continuous variable.
firstlevel=	First category equals '1' (default='0').
	To derive a new categorical variable, use one of the two following options.
ranklevs=	Number of categories for new categorical variable using ranking procedure. (This first option uses uniform ranking to derive a new variable.)
numcutpts=	Number of cutpoints for continuous variable (= number of categories − 1). (This second option uses up to five cutpoints to derive a new variable.)
cutpoint1=	1st (i.e., lowest) cutpoint.
…	…
cutpoint5=	5th (i.e., highest) cutpoint. (At least one cutpoint must be specified when using the 'numcutpts' option.)
cluster=	Set to YES to perform cluster analysis for next two insertions below.
response=	Name for the categorical cluster analysis response variable. (Required when using the 'cluster' option.)
predvar=	Name for the categorical cluster analysis predictor variable. (Required when using the 'cluster' option.)
outdata=	Name for the output data set (to save both existing and new variables).
id=	ID variable for observations (required when outputting the data).
savevars=	Variables to be saved to the output data set (ID is automatically saved).

Table 10. Summary for the ***prelim_ca*** macro parameters.

Parameters	
dataset=	Name of data set to be analyzed.
response=	Name of outcome or response variable.
fmtresp=	Format for response variable (optional).
explanvars=	Explanatory variables (> 2 total categories required).
covars=	Covariates (> 2 total categories required).
fmtothr=	Formats for explanatory variables and covariates (required).
id=	ID for observations (required).
onedim=	Set to YES when the response variable has only two levels.

Table 11. Summary for the ***correspondence*** macro parameters.

Parameters	
dataset=	Name of data set to be analyzed.
response=	Name of outcome or response variable.
fmtresp=	Format for response variable (optional).
explanvars=	Explanatory variables (required).
covars=	Covariates.
fmtothr=	Formats for other variables (required).
id=	ID for observations (required).
onedim=	Set to YES for 1-dimensional solution.
twodim=	Set to YES for 2-dimensional solution.

showobs=	Display observations in the correspondence maps.
stratavar=	Insert a stratification variable.
fmtstrata=	Format for strata (optional).
highlightobs=	Highlight observations for a category of the explanatory variables or covariates.
circlelevel=	Circle observations for a category of the response variable.
noplot=	Insert 'ONE' or 'TWO' to suppress either the first or second correspondence maps.

Table 12. Summary for the ***classification*** macro parameters.

Parameters	
data=	Name of data set to analyze.
var=	Variable names for the initial homogeneity analysis.
print=	Set to YES to print the subset of variables defined by the 'ndim1-3' parameters.
ndim1=	Number of variables from 1st dimension.
ndim2=	Number of variables from 2nd dimension.
ndim3=	Number of variables from 3rd dimension.
sub=	Variable names for subset homogeneity analysis. (Insert ALL to use all in 'var=').
noprint=	Set to YES to exclude the zero levels from the printed output.
haclust=	Names of variables for the cluster analysis.
fitclust=	Set to YES to calculate and plot the CCC statistic for a series of 2-7 clusters.
nclust=	Insert the number of clusters indicated by the CCC plot.
id=	ID variable for the observations (required for the cluster analysis).
allbin=	Set to YES when all variables are binary to simplify the output.
out=	Output the clustering results to the data set named here.
printfreqs=	Set to YES to print frequencies of profiles ordered by the 1st dimension. (Section 6.3)

Table 13. Summary for the ***tabulation*** macro parameters.

Parameters	
dataset=	Name of the data set to be analyzed.
response=	The column variable in the cross-tabulation (required).
fmtresp=	Formatting for the column variable (optional).
secondvar=	The row variable in the cross-tabulation (required).
thirdvar=	The row-stratification variable (optional).
byvar=	The by-variable which stratifies the results into multiple tables (optional).
id=	The ID variable for observations.
perc=	Set to NO to suppress the row percentages.
rowresp=	Set to YES to make response the row variable (and 'secondvar' the column variable)
range1=	The range (e.g., '1 to 5') for the column variable (optional).
range2=	The range (e.g., '0 to 1') for the row variable (optional).
range3=	The range (e.g., '0 to 2') for the row-stratification variable (optional)).

8.1 The *Data_Prep* Macro

(Note: If the data for a study are in tabular form, it must be converted as seen in Appendix Section C.1.)

The ***data_prep*** macro can be used to create a data set with categorical variables for a POSSE analysis. It includes the ability to examine the distributions of quantitative variables, to derive a categorical variable from a continuous one, and to perform a special type of cluster analysis (see Section 7.2 and appendix Sections C.2–C.3). In order to derive a categorical variable from a continuous one, the user must choose one of two of the following methods: (a) the 'ranklev=' parameter or (b) the 'numctpts=' parameter together with up to the five parameters of 'cutpoint1=' through 'cutpoint5='. This macro includes the following parameters, and Table 9 includes a quick reference guide for the parameters. Note that a categorical variable created for a POSSE analysis should contain no numbers in its name. For example, a variable should be named as 'varone' instead of 'var1'.

```
%data_prep(
/**/    rawdata=,
        newdata=,
        contvar=,
        plotdist=,
        catname=,
        firstlevel=,
        ranklevs=,
        numcutpts=,
        cutpoint1=,cutpoint2=,cutpoint3=,cutpoint4=,cutpoint5=,
        cluster=,
        response=,
        predvar=,
        outdata=,
        id=,
        savevars=);
```

'rawdata= <SAS data set>': The name of the original SAS data set. This is required for the first submission of the macro. Because only one categorical variable can be derived from a continuous variable during a single submission, the 'newdata=yes' entry must be turned on after the first such derivation to save the latest version of the data set, and allow the subsequent submissions to continue to collect the derived variables into the new data set.

'newdata=': This parameter can be set to 'yes' after the first submission of the macro in order to begin saving any newly-created variables to an output data set. (This should be left blank during the first submission of the macro during a SAS session.)

'contvar=': Name(s) for the continuous or quantitative variable(s). Only one continuous variable should be specified when the objective is to derive a categorical variable during a submission. However, it might be useful for the first submission of the macro to list all the quantitative variables here in order to examine their distributions together using the 'plotdist=' entry described below.

'plotdist=': This parameter can be set to 'yes' to obtain information about the distribution of the continuous variable(s) found in the 'contvar=' entry.

'catname=': The name for the new categorical variable to be derived from the continuous variable named in the 'contvar=' entry.

'firstlevel=': An entry of '1' specifies that the first category of the newly-created categorical variable will be assigned a value of '1'; otherwise, the macro defaults to assigning '0' to the first category.

<As mentioned above, in order to derive a categorical variable from a continuous variable, the user must choose one of two methods: (a) the 'ranklevs=' parameter or (b) the 'numctpts=' parameter together with up to the five parameters of 'cutpoint1=' through 'cutpoint5='.>

'ranklevs=': This parameter specifies the number of categories using a ranking procedure to derive a categorical variable from a quantitative variable. This will result in each category having the same or nearly the same number of observations, depending on the sample size and the number of tied values for the quantitative variable. Any observations with missing values are automatically placed in a separate category. For example, if 'ranklevs=3' and some observations have missing values, then there will be four categories for the new categorical variable, with the last category representing the missing outcomes.

'numcutpts=' and 'cutpoint1=' to 'cutpoint5=': These parameters are used to define the number of cut-points and the chosen values for up to five cut-points to create a categorical variable from a continuous variable. The cut-points must be listed from lowest to highest. The cut-points should also be defined within the range of the data, where a specified cut-point will represent the minimum of its category. For example, a value equal to the first cut-point defined inside the range of the data will belong to the second category. Note that the number of categories for the new variable will be one more than the number of cut-points listed. Any missing outcomes in the continuous variable will be placed in a separate category. For example, if there are three cut-points and some missing values, then there will be five categories in the new categorical variable, with the last category representing the missing outcomes.

'cluster=', 'response=' and 'predvar=': An entry of 'cluster'='yes' performs a cluster analysis on the rows of a two-way table using the method of Greenacre[36] described in Section 7. The parameters of 'response' and 'predvar' are required for the cluster analysis. The rows of the table represent the levels of the 'predvar=' variable and the columns represent the levels of the 'response=' variable. The categorical variable assigned in the 'response=' entry must already be available in the original data set or have been created in a previous submission of the macro. The categorical variable assigned in the 'predvar=' entry can either be available in the original data set or be created during the same macro submission by using the variable named in the 'catname=' entry which is derived from the variable specified in the 'contvar=' entry. The clustering results include a cluster tree which indicates the stages at which the rows are combined (see Section 7.1).

'outdata=', 'id=' and 'savevars=': Once all the derived categorical variables have been created, the latest version of the data set is saved to a dataset with the name provided in the 'outdata=' entry. The subset of variables to be saved into this new data set must be specified in the 'savevars=' entry. The 'id' identification variable for the observations is required when outputting to a data set. The 'id' variable, which must be numeric, can be an existing variable in the original data set or created in a data step prior to invoking the macro.

8.2 The *Prelim_CA* Macro

The ***prelim_ca*** macro performs preliminary correspondence analyses yielding maps and tables of inertias for the variables. At least two of the three variable types (response, explanatory variable, and covariate) are required for a macro submission. If a response variable and both explanatory variables and covariates are specified, it will perform three individual correspondence analyses and produce three separate tables of inertias and three maps for the correspondence analyses involving (a) the response variable versus the explanatory variable, (b) the response variable versus the covariates, and (c) the explanatory variables versus the covariates. If only explanatory variables are specified, then one table of inertias and one map will be produced for the correspondence analysis involving the response variable versus the explanatory variables. Likewise, if only two of the three types of variables are specified, then results are produced for only one correspondence analysis. Note that only one response variable can be specified. When multiple explanatory variables or multiple covariates are specified5.4, they are analyzed as stacked variables (see Section 5.4).

Although the basic features of the correspondence maps are the same for the ***prelim_ca*** and ***correspondence*** macros, their printouts are different. The printout for the ***correspondence*** macro (see Figure 5) includes the *contributions to the inertias*, which measures the contribution for each category to the *inertia for a dimension*. However, the printout for the ***prelim_ca*** macro (see Figures C.4 and C.5) will display – for each table for a collection of tables – the *inertia for a table*, where the inertia is calculated over all the categories of its rows and columns. For example, if there are eight variables listed as explanatory variables in the macro, then this part of the analysis will display the eight inertias resulting from the eight separate cross-tabulations of the response variable with the explanatory variables.

When the ***prelim_ca*** macro indicates that a covariate is associated with either the response variable or an explanatory variable, then it should be included in a subsequent analysis using the ***correspondence*** macro. In addition, a covariate which is strongly related to an explanatory variable is often a good candidate to be a stratification variable for a ***correspondence*** macro submission.

The ***prelim_ca*** macro includes the following parameters. Table 10 includes a quick reference guide for the parameters, and detailed descriptions are provided here below the macro code.

```
%prelim_ca(
/**/    dataset=,
        response=,
        fmtresp=,
        explanvars=,
        covars=,
/**/    fmtothr=,
/**/    id=,
        onedim=);
```

'dataset= <SAS data set>': The name of the data set to be analyzed (required).

'response=': The name of the outcome or response variable for the analysis. It is assumed that the response variable has at least three categories. When it has only two categories, then the 'onedim=yes' option must be invoked. (See Section 5.4 for details about the 'dimension' of a solution in correspondence analysis.)

'fmtresp=': The name of the format used for the response variable (optional).

'explanvars=': The name(s) for the explanatory variable(s), where variable names are separated by blank spaces. The total number of categories for the variables listed here must be at least three. For example, if there is one explanatory variable, it would need to have at least 3 categories. This condition would also be satisfied by specifying two binary explanatory variables with a total of four categories between them.

'covars=': The name(s) for the covariate(s). The total number of categories for the variables listed here must be at least three.

'fmtothr=': The name of the data set which contains the formatting for the explanatory variables and the covariates (required). Examples in Appendix C will demonstrate how this data set is created using the PROC FORMAT procedure.

'id=': The name of the ID variable which provides the observation identification numbers (required).

'onedim=': This parameter is set to 'yes' when the response variable has only two categories. In this case, the macro will perform a one-dimensional correspondence analysis. (See Figure C.4 in Section C.4 for an example of a one-dimensional correspondence map.)

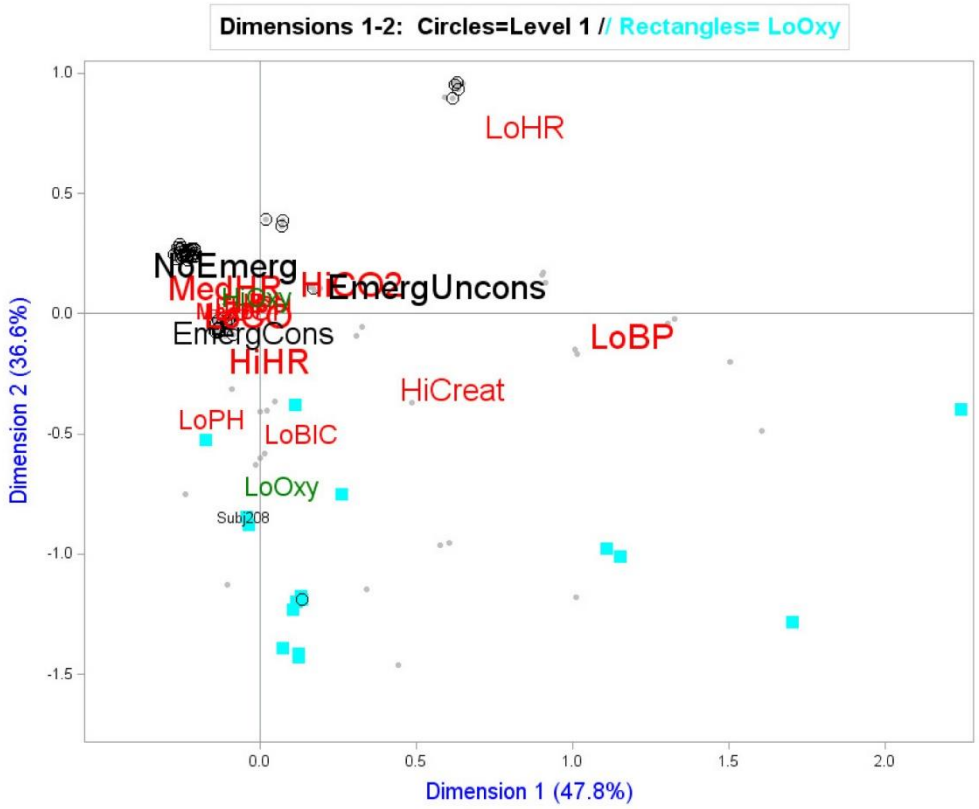
(Note that the ***prelim_ca*** macro is primarily used for the screening of a large number of variables. When there are only a few variables for a data set, a user might want to begin an analysis by directly applying the ***correspondence*** macro, which has more options and features.)

8.3 The *Correspondence* Macro

When a regression model is determined, the resulting equation can be used to predict the values of the dependent (i.e., predictor) variables. In an analogous fashion, the solution from a correspondence analysis can be used to project the observations as supplementary points onto the correspondence map. Software programs which perform standard correspondence analysis will generally only plot the results for the rows and columns of the table, but not the observations themselves. A unique feature of the POSSE methods is the ability to easily display the observations in a map using the ***correspondence*** macro (see Figure 10).

Since the variables are categorical, it often happens that some observations will have the same coordinates in a map. In order to prevent the over-plotting of multiple points, which could then prevent us from assessing where clusters of observations occur, the results are 'jittered' by adding a small random increment to the coordinates of each point[38]. The ability to highlight the observations in a POSSE map provides additional information about the patterns in the data and assists in detecting clusters among the displayed observations. This feature is most useful when there are many variables and can be applied to identify groups of subjects with similar profiles within the map. An example is provided by Figure 10, which comes from Example C.6 in Appendix C. In this figure, the circled points represent patients who have been classified as non-emergencies (i.e., the first category of the response variable), while at the same time observations illustrated as solid squares represent patients who have been classified as having low oxygen levels.

Figure 10. A correspondence map which plots the observations (from Section C.6 of Appendix C). The circled points represent patients that have been classified into the first category of the response variable, and the solid squares represent patients who have classified as having low oxygen levels.



The ***correspondence*** macro can perform a correspondence analysis on both stacked and stratified tables, and provides tables which contain the estimated *contributions to the inertias* (Section 5.4 and Figure 5). Note that the first eight parameters are identical to the parameters for the ***prelim_ca*** macro. Table 11 includes a quick reference guide for the following parameters, and detailed descriptions are given here below the macro code.

```
%correspondence(
/**/    dataset=,
/**/    response=,
        fmtresp=,
/**/    explanvars=,
        covars=,
/**/    fmtothr=,
/**/    id=,
        onedim=,
        twodim=,
        showobs=,
        stratavar=,
        fmtstrata=,
        highlightobs=,
        circlelevel=,
        noplot=);
```

'dataset= <SAS data set>': The name of the data set to be analyzed (required).

'response=': The name of the outcome or response variable for the analysis (required). By default, the macro assumes that the response variable has at least four categories. If the response variable has three categories, the 'twodim=yes' option must be used. If the response variable has two categories, the 'onedim=yes' option is used. (See Section 5.4 for details for the 'dimension' of a solution in correspondence analysis.)

'fmtresp=': The name of the format used for the response variable (optional).

'explanvars=': The name(s) for the explanatory variable(s) (required). The total number of categories for the variables listed here and in the 'covars=' entry must be at least four, unless you are also using the 'stratavar=' option.

'covars=': The name(s) for the covariate(s)

'fmtothr=': The name of the data set which contains the formatting for the explanatory variables and the covariates (required). Examples in Appendix C will demonstrate how this data set is created using the PROC FORMAT procedure.

'id=': The name of the ID variable for the observation numbers (required).

'onedim=': This parameter is set to 'yes' when the response has only two categories. In this case, the macro will perform a one-dimensional correspondence analysis.

'twodim=': This parameter is set to 'yes' when the response has only three categories. In this case, the macro will perform a two-dimensional correspondence analysis.

'showobs=': This parameter is set to 'yes' to display the observations in the correspondence maps.

'stratavar=': The name of the variable which is used in the stratification of the response variable (optional).

'fmtstrata=': The name of the format for the stratification variable (optional).

'highlightobs=' and 'circlelevel=': Inserting the label for a category—as specified in the data set given in the 'fmtother=' entry—into the 'highlightobs=' option will cause all observations in that category to be replaced by solid squares in the correspondence map. Inserting a category number of the response variable into the 'circlelevel='option will cause all observations in that response category to be circled in the correspondence map. In both cases, the subject IDs and coordinates are also given in the output.

'noplot=': An insertion of either 'one' or 'two' will suppress the display of either the first correspondence map (1^{st} versus 2^{nd} dimensions) or the second correspondence map (2^{nd} versus 3^{rd} dimensions).

8.4 The *Classification* Macro

The ***classification*** macro performs homogeneity analysis and cluster analysis (see Section 6 and appendix Sections C.7 and C.13). Unlike the ***prelim_ca*** and ***correspondence*** macros, format labels for the categories are not necessary or used in the ***classification*** macro, but the category number is automatically appended to the name of the variable. In order to be able to clearly distinguish the results in a homogeneity map produced by the 'sub=' entry, the variable names are required to be no more than seven letters in length. The variable names used in the macro should also contain no numbers. If the original variables do contain numbers or have long names, they can be converted to temporary variable names by using the SAS ARRAY statement as seen in the example of Section C.13. The ***classification*** macro includes the following parameters. Table 12 includes a quick reference guide for the parameters, and detailed descriptions are provided here below the macro code.

```
%classification(
/**/    data=,
        var=,
        print=,
        ndim1=, ndim2=,ndim3=,
        sub=,
        noprint=,
        haclust=,
        fitclust=,
        nclust=,
        id=,
        allbin=,
        out=,
        printfreqs=);
```

'data=<SAS data set>': The name of the data set to be analyzed (required).

'var=': The names of the variables to be used in the initial homogeneity analysis. The variable names must contain no numbers and be no more than seven letters in length.

'print=' and 'ndim1=' through 'ndim3=': Set 'print'= 'yes' after examining the results of the homogeneity analysis and the scree plots to print out the variables with the highest discrimination measures for the three dimensions (Section 6.1). The number of variables to print from each of the three dimensions is set by the 'ndim1=' through 'ndim3=' parameters.

'sub=': The names of the variables to use in a subset homogeneity analysis. The 'var=' line is required and the variables listed in the 'sub=' line should be a subset of those in the 'var=' line. An insertion here of 'sub=all' will cause a full homogeneity analysis to be performed on all the variables found in the 'var=' line. The homogeneity maps produced plot only the categories of the variables, not the observations.

'noprint=': When this is set to 'yes', some printed output for the subset homogeneity analysis is simplified by excluding the coordinates for the zeroth categories of the variables.

'haclust=': The names of the variables to be used in a cluster analysis. (The variables listed here need not be a subset of the variables in the 'var=' line.)

'fitclust=': When this is set to 'yes', the CCC statistic is calculated and plotted for a series of 2-7 clusters. This information can be used to choose the number of clusters, which can then be inserted into the 'nclust' parameter of the macro. A positive value greater than two for the CCC statistic will generally indicate that there are well-defined clusters. More details are given in Section 6.2 and in examples found in Appendix C.

'nclust=': The chosen number of clusters. (Leave this blank if you want to explore the results for the 'fitclust=' option.)

'id=': The name of the ID variable for the observation numbers. This is only required for the cluster analysis, but not for the homogeneity analyses using the 'var=' and 'sub=' parameters.

'allbin=': If all the variables are binary or dichotomous, then this can be set to 'yes' to simplify some of the output for the cluster analysis.

'out=': The name of the data set which will contain the clustering results for the number of clusters chosen in the 'nclust=' line. The output will include the variables used in the cluster analysis, the cluster variable and the coordinates for the three dimensions resulting from the homogeneity analysis. These coordinates can be used if the user decides to use a different clustering method (i.e., other than the default K-means method).

'printfreqs=': When this is set to 'yes', the frequencies are given for profiles of zeros and ones, as ordered by their first-dimensional coordinates, in order to determine if there are any binary variables which follow a Guttman or quasi-Guttman scale (Sections 3.2 and 6.3). When this option is used, the variables in the 'haclust=' entry must be listed in the order of the positive outcomes for the variables found (left to right) along the first dimension of the map produced by the subset homogeneity analysis. An example is given in Section C.5.

(Computational resources: Because it is being applied to a data matrix instead of a table, a homogeneity analysis can be relatively computer-intensive. In applications with a 64-bit operating system with 8.0 GB of RAM, datasets with up to 9000 observations and 250 total categories for the variables were successfully analyzed. Depending on the available computing resources, it might be necessary to analyze a random sample for larger data sets.)

8.5 The *Tabulation* Macro

The ***tabulation*** macro can be used in conjunction with the other POSSE macros to clarify or supplement the interpretation of results. The ***tabulation*** macro creates cross-tabulations and stratified cross-tabulations of variables in order to elucidate the nature of possible associations or patterns (see appendix Sections C.4, C.7, C10–C.12). The macro includes the following parameters. Table 13 includes a quick reference guide for the parameters, while detailed descriptions are given here below the macro code.

```
%tabulation(
/**/    dataset=,
/**/    response=,
        fmtresp=,
/**/    secondvar=,
        thirdvar=,
```

```
                byvar=,
        /**/    id=,
                perc=,
                rowresp=,
                range1=,range2=,range3=);
```

'data set=<SAS data set>': The required name of the data set to be used.

'response=': The column variable in the cross-tabulation (required).

'fmtresp=': The formatting for the column variable.

'secondvar=': The row variable in the cross-tabulation (required).

'thirdvar=': The row-stratification variable.

'byvar=': The by-variable which further stratifies the row and column results into multiple tables.

'id=': The name of the ID variable for the observation numbers (required).

'perc=': Set to 'no' when row percentages are not wanted.

'rowresp=': Setting this to 'yes' switches the roles of the 'response=' and 'secondvar=' variables by making the 'response=' entry the row variable for the table and the 'secondvar=' entry the column variable.

'range1=' through 'range3=': When the frequencies in a row or column of a table are all zero, SAS will automatically delete them from a cross-tabulation. However, this practice can make it more difficult to compare a set of complete and incomplete tables. To provide complete tables for all cross-tabulations, when some rows or columns in some cross-tabulations would otherwise be missing, you can specify the ranges for those variables listed in the 'response=', 'secondvar=' and 'thirdvar=' lines. For example, assume that the 'response=' and 'secondvar=' variables have three and four categories, respectively, with some categories for either variable having no outcomes. Then specifying 'range1=1 to 3' and 'range2=1 to 4' will result in a complete table. Note that, if the first category of the response variable had been '0', then we would have specified 'range1=0 to 2'.

9. Additional Recommendations

Those readers who have been analyzing data for many years will recognize the importance of context. For example, one statistical method that is successful in one study might not be appropriate for a similar study because of what might seem to be small differences in the sampling design. Therefore, we are hesitant to provide any stringent rules here, such as specifying that certain POSSE macros should be applied for particular data strategies. However, a general rule-of-thumb is that analysis follows design, so we can safely caution users to always allocate enough time to fully understand the study design and the data-collection methods before beginning any POSSE analysis.

We recommend that the user often check the output in the SAS LOG window for any error messages. One common error message, which appears in green in the log window, is "WARNING: Row partial

contributions to inertia are tied or nearly tied. There may be more than one possible set of BEST statistics." This warning results from the fact that tied values are fairly common for categorical variable and, therefore, this warning can usually be safely ignored. However, error messages that appear in red in the LOG window should be addressed. Two common mistakes are (1) failing to insert the ID variable when it is required and (2) submitting the ***correspondence*** or ***prelim_ca*** macros using a binary response variable but without also specifying the 'onedim=yes' option.

9.1 Row and Column Interactions

A general principle given by Greenacre[28] is that, when a simple correspondence analysis is performed, rows and columns interact, but rows do not interact with other rows of the same table and, similarly, columns do not interact with each other. When we analyze a stacked table, we can use our results to indicate how the response categories in the column are interacting with the stacked variables in the rows, but we cannot use our results to indicate how the variables in the rows might be interacting.

A more comprehensive view of this issue is possible by discussing the transitive properties of correlations and associations. To illustrate this for three quantitative variables, suppose that variable X is positively correlated with variable Y, and Y is also positively correlated with a third variable Z. Does this then imply that X and Z are positively correlated? The general answer is no. However, if the first two correlations are perfect (e.g., r = 1.0), then the answer would be yes and, according to Jacobsen[39], if the first two correlations are both greater than 0.71 — or if both are less than -0.71 — then the third correlation (X and Y) would necessarily be positive. If the first two correlations were both 0.5, then there is some latitude in the possibilities for the third correlation. It could be zero, which means X and Z are independent, or it could even be as low as -0.5, a moderately negative correlation.

To examine this for categorical outcomes, suppose that 10 of the 20 cases of disease in a health study are from Native-American subjects, and 10 of the 20 cases are also from subjects over 65 years old, and that our analysis indicates each of the factors of race and age are also associated with being a case. Are these conditions (i.e., being Native-American and being over 65) also related to each other? In other words, do they tend to co-occur? Because these reported frequencies represent summary statistics, they don't tell us how the race and age variables interact. One extreme possibility is for the ten Native-American cases and the ten over-65 cases to be one and the same, implying a positive association. The other extreme is for the outcomes to be mutually exclusive, with none of the Native-American subjects coinciding with the over-65 subjects, implying a negative association. An examination of the POSSE output, the maps and some stratified tables will usually provide some insight regarding the nature of any patterns, but we generally need to be careful when interpreting such results for simple correspondence analysis. For instance, we cannot assume a positive correlation exists for two row categories (or two column categories) when their labels are in close proximity in a map when using simple correspondence analysis.

Note, however, that this general principle for interactions does not apply to homogeneity analysis because the categories — represented by the columns of a data matrix — are interacting with each other through the observations themselves, which are represented by the rows of the data matrix. Therefore, if our purpose is to examine the overall associations among a single group of variables, we can always choose to use the ***classification*** macro to perform a homogeneity analysis.

9.2 Stacked Tables and Screening Variables

As was mentioned in Section 8, although the correspondence maps are identical for the ***prelim_ca*** and ***correspondence*** macros, the macros differ in the numerical summaries that they generate. The ***prelim_ca*** macro provides the total inertias of tables, whereas the ***correspondence*** macro provides the contributions to the inertias for the rows and columns of tables. This distinction is especially important for the stacked tables that are defined in Section 5.4. In Section 5.4, the ***correspondence*** macro was used to produce the results given in Figure 5 for the Titanic example. That submission provided the contributions to the inertia which indicated the nature of the associations by identifying the categories which were loading on the various dimensions of the correspondence analysis. This analysis is similar to the approach for stacked tables given in Section 4.3 of the Clausen[26] monograph on correspondence analysis.

On the other hand, the numerical summaries for stacked tables using the ***prelim_ca*** macro would agree with the approach given in Chapter 17 of the Greenacre[28] text. In this case, the total inertia for each of the individual stacked tables is given. In particular, the ***prelim_ca*** macro will provide the total inertias for a series of stacked tables which cross-tabulate the response variable versus the other variables. These total inertias then provide overall measures of the associations (i.e., over all the categories for each variable). This makes the ***prelim_ca*** macro a useful tool for screening out any variables with weaker associations with the response variable, as is demonstrated in appendix Section C.4. Although this is somewhat arbitrary, in many cases where there are ten or more explanatory variables, we have sometimes found it useful to retain only those variables with total inertias greater than the mean total inertia for all the variables. We have used a similar approach for each of three possible correspondence analyses (see Section 8.2) produced by the ***prelim_ca*** macro. A user might also decide to retain only a few variables when their inertias are much larger than those for the other variables. In addition, when there is an especially large inertia between the explanatory variables and one of the covariates, we recommend using that covariate as a stratification variable in the ***correspondence*** macro.

The analysis of stacked tables works because of the unique way that correspondence analysis handles the duplication or redundancy of information, which was demonstrated by the correspondence maps (Figures 2 and 3) presented in Section 5.1. Redundancies generally do not distort the results in correspondence analysis. This stands in contrast to what happens when there are redundancies for the predictors in regression analysis. In that case, it leads to the phenomenon of collinearity and parameter estimates with inflated variances.

9.3 Limitations

Although these methods will hopefully be useful for a wide range of applications, there are some applications which might be inappropriate or would be analyzed more effectively using other software. These would include special applications like time-series analysis, or analyses that require customized software such as geographical information studies. In addition, although the POSSE methods can provide some guidance for subsequent modeling, they should not be the sole tool for model-building.

One other limitation is that, because continuous variables are represented by categorical variables in a POSSE analysis, any univariate outliers would be masked by the conversion to categories. Therefore, the distributions of the continuous variables should always be examined separately in order to identify any univariate outliers.

Finally, the applications of the POSSE methods to samples with less than 50 observations will not always be very effective, particularly when the cross-tabulations result in most cells having counts of five or less. In these cases, the sparseness of the counts can lead to an unstable solution where very small changes in a table result in large changes in the contributions to the inertias.

10. Key Terms

Contribution to the inertia (Sections 5.2 and 5.4): The proportion of the inertia for a given dimension (or axis) which is due to a category. The ***correspondence*** macro (Section 8.3) produces two sets of contributions, each of which adds to one, for each dimension.

Correspondence analysis (or 'simple' correspondence analysis) (Section 5): A method for examining the associations (or correspondences) between the rows and columns of a contingency table (i.e., a two-way table).

Correspondence map (Section 5.1): A two-dimensional plot that illustrates the associations between the rows and columns of a contingency table.

Cubic clustering criterion; CCC statistic (Section 6.2): A statistical measure developed by Sarle (1983) for determining the number and reliability of groupings for a cluster analysis.

Dimension of a solution (Sections 5.1 and 5.4): The dimension of the solution for a simple correspondence analysis is defined as the minimum($r - 1, c - 1$), where r and c are the numbers of rows and columns in the table.

Discrimination measure (Section 6.1): In homogeneity analysis, a measure of the spread of the categories for a variable for a given dimension and, therefore, its ability to provide better discrimination for observations. The relative sizes of the discriminations measures can be used to determine a subset of variables with good discrimination.

Guttman scaling (Sections 4.2, 6.3 and 6.4): A special type of ordinal scale that implies a hierarchy for the outcomes for a set of binary variables.

Homogeneity analysis (Section 6): An approach to multiple correspondence analysis that extends the scaling properties of simple correspondence analysis.

Homogeneity map (Sections 6.1 and 6.2): The plot resulting from a homogeneity analysis, where the position of categories to each other in this plot indicate their correlations, while their position with respect to the origin and the axes suggest their contributions to the discrimination.

Horseshoe or arch effect (Sections 5.3 and 6.3): A perceived arch of points or labels in a map. In a simple correspondence map, this arch might be associated with an ordinal variable. In a homogeneity map, it might be associated with a series of binary variables which follow a Guttman or quasi-Guttman scale.

Inertia for a dimension (Sections 5.1, 5.2 and 6.2): A measure of variation in correspondence analysis, it is also sometimes referred to as the *principal inertia*. The principal inertias are found by taking the total inertia and then splitting it among the dimensions. (The total inertia is defined as a weighted average

of the chi-square distances (Section 5.1) between the profiles and the average profiles, using either the row or column profiles.) In simple correspondence analysis, the relative sizes of the inertias for the various dimensions indicates how much of the explained variation is captured by the dimensions. In homogeneity analysis, the relative sizes of the inertias for the various dimensions indicates the ability for a dimension to provide good discrimination for the observations.

Inertia for a table (Section 8.2): The total inertia for a table. These inertias are seen for simple tables (with one row variable and one column variable) in the printouts for the ***prelim_ca*** macro.

Profile for the row or column of a table (Section 5.1): The set or array of proportions for the counts in a given row or column, where the elements of the array add to one.

Profile for a cluster (Section 6.2): For a cluster found using homogeneity analysis, its profile is an array of values which represent the proportions of the cluster members who are classified into the various categories for the set of variables used in the cluster analysis. The components of this array will usually not sum to one.

Quality (Section 5.4): A measure of the goodness-of-fit for a category within a correspondence map. This measure is not presented numerically in the POSSE methods, but is instead tied to the sizes of the labels found in a correspondence map in order to enhance the interpretation.

Scaling and *optimal score* (Sections 5.4 and 6.1): The concept of scaling generally refers to the ability to provide an objective ordering of categories and, in some cases, a measure of the relative distances between categories. In homogeneity analysis, optimal scaling refers to the determination of scores which optimize the discrimination among the observations.

Scree plot (Section 6.1): A plot of the discrimination measures of a set of variables against their order, as ordered from largest to smallest. The elbow found in the plot is sometimes used to determine the number of variables which are providing good discrimination for the observations for a specified dimension.

Subset homogeneity analysis (Section 6.2): A technique that can provide more consistency between various possible sub-analyses. This is available for the homogeneity analysis which is performed using the ***classification*** macro.

11. Concluding Remarks

The key numerical quantities and where they are found using the POSSE macros are as follows:

(a) After the *inertias for the dimensions* are divided by the total inertia, the result is expressed as percentages of the total inertia for the various dimensions in a correspondence map when using either the ***prelim_ca*** or ***correspondence*** macros. These percentages are found along the axes of the maps.
(b) The *inertias for tables* are found in the printout for the ***prelim_ca*** macro. This quantity represents the total inertia for a simple table with one row variable and one column variable (unlike the more complicated stacked and stratified tables which are discussed in Section 4.3 and analyzed using the ***correspondence*** macro).
(c) The *contributions to the inertias* are found in the printout for the ***correspondence*** macro. This quantity is the proportion of the *inertia for a dimension* (defined in (a) above) which is due to a single category of a variable.

(d) The *discrimination measures* are found in printouts for the ***classification*** macro. This is a measure of the spread over all the categories of a single variable for a specified dimension of a homogeneity analysis.
(e) A measure of the goodness-of-fit called the *quality* can be seen in a correspondence map by relating it to the size of the label for a category. This feature is available in the maps provided by both the ***prelim_ca*** and ***correspondence*** macros.

This guide describes the application of exploratory methods used to assess the systematic variation among variables before having to settle questions about the random component of the analysis (e.g., whether the error follows a normal or lognormal distribution). There are other approaches which have similar objectives. These include methods in the area of data mining, such as those presented by D'Enza et al.[40]. Also, a method for screening variables using correspondence analysis is illustrated in Crichton and Hinde[41], and correspondence analysis has also been used in conjunction with cluster analysis by Lebart[42].

In Section 7.2 we discussed a method for choosing the number of categories when deriving categorical variables from quantitative ones. The paper by Pasta[43] offers some additional guidance for such derivations. Also, note that an early paper by Cox[44] has shown that, for a normally-distributed outcome, there is little loss of information when categories with equal frequencies are specified instead of using an optimal approach which specifies categories with unequal frequencies. This work by Cox provides some justification for using the ranking procedure found in the ***data_prep*** macro to derive categorical variables.

We have not discussed the computational engine for correspondence analysis, which is called the 'singular value decomposition' or SVD, a mathematical method used in many areas of science and technology. Books and articles by Michael Greenacre and others[10, 16, 27-29] provide some details and discussion for applications of the SVD method. We have also not provided any details in this guide for the multivariate analysis of quantitative data, but the POSSE methods are connected with two standard multivariate methods: (a) principal component analysis and (b) canonical correlation analysis[18-19, 25]. Principle component analysis can be used to explore the associations among a single set of quantitative variables, and homogeneity analysis often has that same goal for a single set of categorical variables. Although simple correspondence analysis also has some connections with principal component analysis, it has a clearer connection with canonical correlation. While principal component analysis is focused on the associations within one set of variables, canonical correlation analysis is focused on the associations between two sets of variables. Simple correspondence analysis is used in the POSSE methods to analyze two-way tables, and so it can be viewed as a categorical analogue of canonical correlation for two sets of categorical variables, namely, the row variables and the column variables. All of these techniques can be also referred to as dimension-reducing methods, because the objective is to provide a simplification for data with many variables by focusing on the relationships found for a smaller number of dimensions. This becomes particularly useful when we are dealing with 10 or more variables, because in many cases two or three dimensions can convey much of the information in the data.

As the reader moves on to study the examples in Appendix C, some practice with submitting the macros is recommended. Appendix A describes some opportunities for practice. During this process, users might want to refer back to Section 8, to the definitions of key terms in Section 10, and to the quick reference guides in Tables 7 through 13.

IV. APPENDIX A: Accessing and Using the POSSE Macros

This appendix provides information on downloading and invoking the POSSE macros, as well as practicing with the macros using some available datasets.

The POSSE macros can be downloaded at the following:

https://github.com/POSSE-Macros/POSSE-Materials

Once you have inserted the directory information, you can include the five POSSE macros within a SAS session using the following submission.

```
/*-----------------------------------------------------*
 | Include the files containing the macros. |
 *-----------------------------------------------------*/
        filename dataprep 'DIRECTORY_INFO...\data_prep.sas';
        filename classify 'DIRECTORY_INFO...\classification.sas';
        filename prelim 'DIRECTORY_INFO...\prelim_ca.sas';
        filename correspond 'DIRECTORY_INFO...\correspondence.sas';
        filename tabulate 'DIRECTORY_INFO...\tabulation.sas';
        %include dataprep classify prelim correspond tabulate;
```

The above submission is also included in two additional SAS files that can be downloaded at the same time as the macros. These files are named 'posse_templates_long' and 'posse_templates_short'. The first file includes all of the SAS code to invoke the five macros along with some notes to assist the user. The second file includes just the SAS code without any notes. These files can be adapted for your personal analyses using the POSSE macros.

To provide some opportunities for practice with the macros, the folder entitled 'example_programs' contains SAS programs which reproduce results for 11 of the 13 examples found in Appendix C. We recommend that users first start a SAS session before opening and running the programs.

V. APPENDIX B: Example Data Sets

We have tried to use datasets for the examples in Appendix C which are not overly demanding, and yet rich enough to illustrate the various features for the POSSE methods. We are grateful to those researchers and organizations which have made their data available. Note that, because of potential copyright issues, the data sets described in Sections B.2 and B.4 are not available for download.

B.1 Intensive Care Unit (ICU) Data

The intensive care unit (ICU) data are featured in all editions of the textbook *Applied Logistic Regression* by Hosmer and Lemeshow[45], including the more recent editions which were published in 2000 and 2013. Their sample of 200 subjects was taken from a larger study which was reported by Lemeshow et al.[46]. The data for the sample are included in an appendix of the textbook.

The aim of the study was to examine the ICU experience of patients in order to determine those factors or predictors which are related to survival. Table B.1 contains the variable names and brief descriptions for the data set named 'icucat', which was derived from the original dataset for use in a POSSE analysis. The response variable is the vital status of each subject (i.e., survived or died) at discharge. Among the other variables are the vital sign measurements of systolic blood pressure and heart rate, along with five binary outcomes for the initial blood-gas test results, which indicate the levels of oxygen, PH, bicarbonate, carbon dioxide and creatinine. Because adverse health outcomes are known to be associated with either low levels of oxygen, pH, and bicarbonate, or high levels of carbon dioxide and creatinine, the presence of these various low or high levels are classified as '1' for the five binary variables, and classified as zero otherwise. Another factor is the reported level of consciousness at admission. The original variable has the three levels of consciousness labeled as 'conscous', 'stupor' and 'coma'. However, due to the small number of observations, we have followed the practice of previous analyses and combined the second and third categories to create a binary variable that is classified as '1' to indicate unconsciousness, and zero otherwise. Similarly, categories of race have been combined to yield the two categories of white and non-white.

Table B.1. Variable names and brief descriptions for the ICU dataset named 'icucat' which is derived from the data from Hosmer and Lemeshow[45]. Note that some variable names have been changed from those found in the original dataset.

ICUCAT Dataset: 200 Observations / 22 Variables	
status : the response variable of vital status (survived or died) at discharge	
agecat : age categories	*uncons* : unconscious or not at admission
bloodpress : systolic blood pressure at admission	*fracture* : fracture or no fracture at admission
heartrate : heart rate at admission	*gender* : female or male
surgery : had surgery or medical service	*race* : non-white or white
cancer : admission partly due to cancer	*ID* : subject identification number for the study
renal : history of renal (kidney) failure	*bldoxy* : low oxygen (< 60) from bloodwork
infect : probable infection at admission	*bldph* : low pH (≤ 7.25) from bloodwork
cpr : CPR performed prior to admission	*bldbic* : low bicarbonate (< 18) from bloodwork
previcu : previous admission in last six months	*bldco* : high carbon dioxide (> 45) from bloodwork
emergency : elective or emergency admission	*bldcret* : high creatinine (> 2) from bloodwork
coma: 0=conscious, 1=stupor, 2=coma (*uncons* is derived from this by combining the last two categories)	

As mentioned in the main text of this users' guide, all the variables used by the POSSE macros must be numeric and categorical. Three of the original predictor variables – *age*, *systolic* and *hrtrate* – are continuous or quantitative. They measure age in years, systolic blood pressure and heart rate. In Table B.1, the categorical version of these variables are named *agecat*, *bloodpress* and *heartrate*. The derivations of these categorical variables can be seen in Sections C.2 and C.3 of Appendix C.

The original data set named 'icu' is available at the POSSE website, and can also be downloaded from several other websites. A data dictionary can also be downloaded with the data by inserting the title of the Hosmer and Lemeshow text at the following website.

http://wiley.mpstechnologies.com/wiley/BOBContent/searchLPBobContent.do

B.2 Diabetes Data

The diabetes data are analyzed in Chapter 6 of the book *Multivariate Dependencies* by Cox and Wermuth[47]. Their analysis looks at the glucose control for 68 patients who have had diabetes for less than 25 years. Because their response variable is the measured glycosylated hemoglobin (GHb, an estimate of the glucose concentration), a lower value indicates better control of diabetes.

The other variables include five quantitative variables and the two binary variables of gender (1=male, 2=female) and education (1=less than 13 years of formal schooling, 2=at least 13 years of formal schooling). The five quantitative variables include the duration of illness and four patient scores which are intended to measure (a) the patient's knowledge about diabetes, (b) the patient's sense of fatalism, (c) the patient's sense that systematic, social forces control his or her future, and (d) the patient's self-reliance or the sense that he or she has control over the future.

The ***data_prep*** macro was applied to the original data set to create a new data set named *glucat* with only categorical variables. Some preliminary results, not shown here, suggested that the choice of three categories was generally sufficient to capture the relevant information for the new categorical response variable named *control* and for the five quantitative predictor variables.

Table B.2 lists the variables used in the POSSE analysis, where the variable names used in the original data set are also seen in parentheses. In addition, after the new categorical response variable of *control* was derived from the quantitative response variable named *y*, we reversed the order for the new response variable, so that a higher category is associated with lower measured GHb. In other words, a higher categorical response of '3' indicates relatively better control of diabetes, whereas a response of '1' indicates relatively poor control.

(Note that, because of potential copyright issues, the data sets and SAS code for the analysis of this example are not available for download at the POSSE website.)

Table B.2. Variable names and brief descriptions for the dataset named 'glucat'. Note that the variable names have been changed from those found in the diabetes data used by Cox and Wermuth[47]. The original variable names are shown in parentheses. The response variable of *control* and the first five variables have been derived from quantitative variables. Each was given three categories.

GLUCAT Dataset: 68 Observations / 9 Variables	
control (y): the response variable which indicates the patient's control over diabetes, where 1=low control, 2=average control and 3=high control	
knowledge (x) : knowledge about diabetes	*duration (w)* : duration of illness
fate (z) : degree of fatalism	*education (a)* : amount of formal schooling
system (u) : sense of powerlessness	*gender (b)* : female or male
reliance (v) : self-reliance	*ID* : subject identification number for the study

Table B.3. Variable names and brief descriptions for two datasets from the seafood study (Ortega et al.[48]). The dataset named 'symptoms' contains a subset of 20 symptom (0/1) variables which were collected from 107 workers during both the initial and follow-up surveys, along with two binary variables which indicate age and exposure categories. The second dataset named 'fev1cat' contains variables with ordinal categories for FEV1, which measures the total volume of air exhaled in the first second during lung-function testing.

SYMPTOMS Dataset: 107 Observations / 23 Variables. (excluding ID variable)	
survey : 1=January survey, 2=March survey	
wheeze : chest wheezing	*coldair* : symptoms with cold air
wheezecond : wheeze besides during a cold	*workdust* : symptoms with work dust
clearcgh : wheeze clears after a cough	*wokecgh* : woken by cough
normbetw : normal breathing between wheezing	*cghmorn* : cough in the morning
shortbreath : shortness of breath	*cghlater* : cough later in the day
chesttight : chest tightness	*cghoften* : cough most days or nights
freqsympts : frequent chest symptoms	*phlgmorn* : phlegm in the morning
heavyexer : symptoms with heavy exercise	*phlglater* : phlegm later in the day
exerbreath : shortness of breath with exercise	*stuffnose* : stuffy nose or drainage
afterexer : wheezing after exercise	*blocknose* : blocked or runny nose
agecat : 1= <35 years, 2= ≥35 years	*exposure* : 1=low exposure, 2=high exposure
FEV1CAT Dataset: 99 Obs. / 2 Vars. (excluding ID variable)	
firstfev : 1=Low FEV1, 2=Ave. FEV1, 3=High FEV1 as measured in January 1998	
secondfev : 1=Low FEV1, 2=Ave. FEV1, 3=High FEV1 as measured in March 1998	

B.3 Occupational Asthma Data

The two data sets described in Table B.3 are examined from a study at a seafood-processing plant (Ortega et al.[48]). This study conducted surveys in January and March of 1998. The purpose of the study was to determine the relationship between work exposures and the onset of asthma and other respiratory conditions. Questionnaire data were collected on both occasions from 107 workers. The first data set named 'symptoms' contains the responses to a subset of questions that were asked of the workers during both the initial questionnaire in January 1998 and the follow-up questionnaire in March. This data set also includes two binary variables that include categories of age and exposure.

The surveys also included pulmonary function testing. A second data set from the study contains the FEV1 outcomes for both surveys for 99 of the 107 workers. FEV1, which is a widely-used measure of lung function, is the total volume of air forcefully exhaled in the first second of a lung function maneuver.

Example C.11 demonstrates how categorical responses have been derived and saved to the dataset named 'fev1cat'.

Variable names and brief descriptions for the two data sets are found in Table B.3.

B.4 Coronary Risk Factor Data

The Muscatine Coronary Risk Factor study (Woolson and Clarke[49], Fitzmaurice et al.[50]) surveyed 4,856 school-age children in Muscatine, Iowa during the years 1977-1981. The purpose of the study was to assess the association of coronary disease with possible risk factors, including obesity. The study contains five birth cohorts, that is, five groups of children about the same age.

The investigation of birth cohorts is especially important when, for example, we have subjects in a longitudinal study who can be about the same age but born in different decades. We cannot assume that a 75-year-old born in 1920 and a 75-year-old born in 1940 have had the same experience with respect to access to health care and health education. In the Muscatine study, the five birth cohorts were defined as those who in 1977 are 5–7 years, 7–9 years, 9–11 years, 11–13 years, and 13–15 years. Since the study is of relatively short duration, we might expect any cohort effects to be small. Although there is no reason to believe that the experience of the cohorts is different, we can still explore the possibility of the cohorts being associated with other effects during an informal preliminary analysis. (See Fitzmaurice et al.[50] for a formal modeling approach to assess the cohort effects.)

Ideally, each subject in the study would have had assessments at three time points for whether his or her weight was normal or obese. However, because there were many missing values for this study, instead of two classifications for obesity, there were three possible classifications: 'Normal', 'Obese' and 'Missing'.

Table B.4(a)-(b). (a) The sample sizes for the various age cohorts and survey years for the Muscatine study (Woolson and Clarke[49], Fitzmaurice[50]), and (b) the variable names and description for the dataset named 'muscatine'.

(a) Sample Sizes by Age Cohorts and Years						
	Age Cohort					**Totals**
	5-7	**7-9**	**9-11**	**11-13**	**13-15**	
1977	935	1014	1025	937	945	4856
1979	935	1014	1025	937	945	4856
1981	935	1014	1025	937	945	4856
Totals	2805	3042	3075	2811	2835	14568

(b) MUSCATINE Dataset: 14,568 Observations / 4 Variables	
obesity: the response variable which indicates a subject's weight classification: 1=normal, 2=obese and 3=missing	
age: middle age for a subject's cohort at the time of one of the three surveys	
cohort : age cohort	*sex* : female or male

B.5 Cancer Information Data

The cancer information data comes from a classic study which has been analyzed by a number of prominent statisticians. The data, first presented in the paper of Lombard and Doering[51], are given by the cross-tabulation in Table B.5. The five binary (0/1) variables include a person's knowledge about cancer,

rated as poor or good. This knowledge is analyzed in Appendix C as the response variable with respect to four explanatory variables that measure (a) whether or not a subject read newspapers, (b) whether or not he or she was considered a solid, or thorough, reader, (c) whether he or she attended lectures, and (d) whether he or she listened to the radio. Example C.1 in Appendix C demonstrates how the table is converted to a SAS dataset for a POSSE analysis.

Table B.5. Cross-tabulation of the five binary variables for the Lombard and Doering[51] study with 1729 subjects.

		No Radio				**Radio**			
		No Solid Reading		**Solid Reading**		**No Solid Reading**		**Solid Reading**	
		Knowledge		**Knowledge**		**Knowledge**		**Knowledge**	
		Poor	Good	Poor	Good	Poor	Good	Poor	Good
No Newspaper	**No Lectures**	393	84	83	67	50	13	16	16
	Lectures	10	2	8	3	3	4	3	1
Newspapers	**No Lectures**	156	75	177	201	59	35	67	102
	Lectures	6	7	18	27	4	8	8	23

B.6 X-Ray Classification Data

The X-ray classification data set named 'ratings' is from unpublished data from the National Institute for Occupational Safety and Health (NIOSH). It contains x-ray image ratings by nine doctors at two different time points using a classification system for pneumoconioses[30] (i.e., silicosis and other dust-related lung diseases). This classification system has been developed by the International Labour Organization (ILO) and includes a classification of 'profusion' which corresponds to the grading by a doctor of the frequency of small shadows in chest X-rays. This assessment is composed of 12 ordered categories. Because of small frequencies, the bottom two categories are collapsed into one category, and the top two categories are combined as well, resulting in 10 ordered categories, which range from the lowest profusion category of paired zeros '0/0' to the highest profusion category of paired threes '3/3'.

The scaling example in Section 5.3 used classifications by nine doctors for over 200 images. Of those images, there were 129 images with repeated classifications at a second time point by the same nine doctors. Variable names and brief descriptions for this data are found in Table B.6.

Table B.6. The variable names and description for the unpublished NIOSH data set named 'ratings'.

RATINGS Dataset: 1,161 Observations / 3 Variables	
Ordinal Categories for the Two Response Variables: '0/0' '0/1' '1/0' '1/1' '1/2' '2/1' '2/2' '2/3' '3/2' '3/3'	
Variable	**Description**
ratingone	response for the first round of the trial
ratingtwo	response for the second round of the trial
rater	raters 1 – 9
image	number identifying an x-ray image

VI. APPENDIX C: Examples

As Mallows and Tukey[9] have stated, exploratory data analysis "is necessarily an iterative process, in which one performs many tentative analyses, following up promising leads, discarding others, and searching always for strong, stable and meaningful patterns in the data." Some of the examples that follow will demonstrate that the POSSE methods are more open-ended than a standard statistical method like the t-test, which can be described in terms of a step-by-step closed process. The diversity of challenges that are possible near the beginning of an analysis necessitate a variety of approaches which might rely on different features of the macros, or use the macros in different combinations. We have tried to present a collection of examples which will suggest an array of options for the exploration of data.

To provide some opportunities for practice with the macros, the folder entitled 'example_programs', available for download, contains SAS programs which reproduce results for 11 of the 13 examples found in Appendix C. We recommend that users first begin a SAS session before opening and running the programs.

C.1: Converting a Table to a SAS Data Set

In Section B.5, the frequencies for a cancer information study[51] are found within Table B.6. To use the information in a POSSE analysis, the following SAS code can be applied to convert the table to a SAS data set named 'cancer_information', where the number of observations is equal to the sum of the frequencies.

```
data canc_info;
input radio papers lectures reading @;
 do knowledge = 0 to 1;
   input count @;
   output;
 end;
datalines;
0 0 0 0 393  84
0 0 0 1  83  67
0 0 1 0  10   2
0 0 1 1   8   3
0 1 0 0 156  75
0 1 0 1 177 201
0 1 1 0   6   7
0 1 1 1  18  27
1 0 0 0  50  13
1 0 0 1  16  16
1 0 1 0   3   4
1 0 1 1   3   1
1 1 0 0  59  35
1 1 0 1  67 102
1 1 1 0   4   8
1 1 1 1   8  23
;

data canc_info;
  set canc_info;
    do i=1 to count;
```

```
    output;
   end;
drop count i;

data canc_info;
length id 8;
 set canc_info;
id = _n_;

data cancer_information;
 set canc_info;
run;
```

The following SAS code recreates the cross-tabulation of Table B.5 using the new SAS data set.

```
proc format;
 value radio   0='No Radio' 1='Radio';
 value reading  0='No Solid Reading' 1='Solid Reading';
 value papers  0='No Newspaper' 1='Newspapers';
 value lectures 0='No Lectures' 1='Lectures';
 value knowledge 0='Poor' 1='Good';
run;

proc tabulate data=cancer_information format=6.;
class radio reading papers lectures knowledge;
tables (papers=''*lectures=''),
   (radio=''*reading=''*knowledge='Knowledge')*(n=' ') / misstext=' ' rts=15;
format radio radio. reading reading. papers papers. lectures lectures. knowledge knowledge.;
run;
```

C.2: Determining the Categories for a Derived Categorical Variable

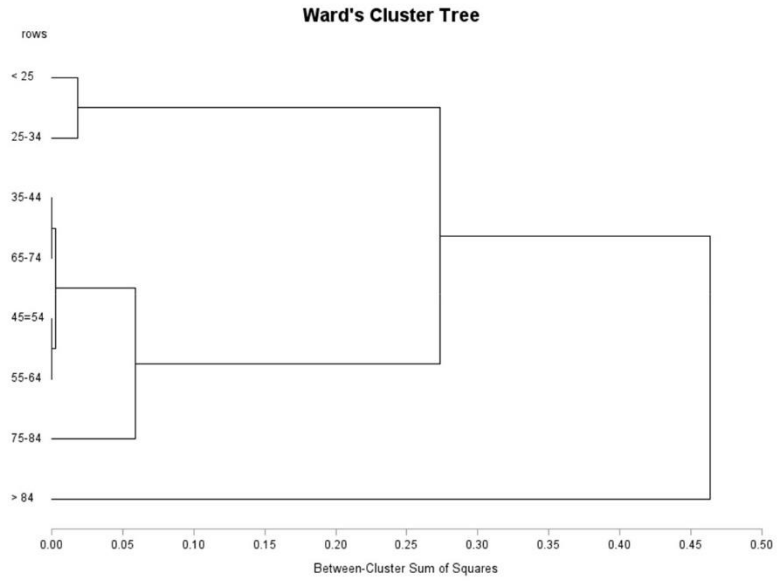
In the original dataset for the intensive care unit (ICU) example[45-46], which is described in Section B.1 of Appendix B, the age variable is quantitative. Therefore, a categorical age variable must be derived before applying a POSSE analysis. This section will describe methods for determining the categories for the new variable.

In the solution manual by Cook[52], the survival of the ICU patients at discharge is cross-tabulated with the eight age categories which are represented by the rows of Table C.1. However, with only 200 observations, the cross-tabulations of the eight age categories with other variables will result in tables that are sparse, that is, tables with many empty cells as well as cells with counts of five or less. This will make interpreting the results of the POSSE analysis more difficult. In such cases, the clustering procedure in the ***data_prep*** macro and described in Section 7 of the text can be used to explore whether we should collapse any of these age categories represented by the rows of Table C.1, where those row results are defined by their association with the response variable of *status*, the column variable of Table C.1. (See also Section 7, where the clustering results of Figure 9 are used to collapse some of the rows of Table 6, and where the row results are defined by their association with the columns of Table 6.)

Table C.1: The cross-tabulation for the vital status at discharge versus ages, grouped into eight categories defined by Cook[52], together with the mortality for each age category.

	Survived	Died	Totals	Mortality
Age (years)				
< 25	24	2	26	8%
25-34	8		8	0%
35-44	9	2	11	18%
45=54	20	5	25	20%
55-64	31	8	39	21%
65-74	41	9	50	18%
75-84	21	9	30	30%
> 84 yrs.	6	5	11	45%
Total	160	40	200	20%

Figure C.1. The cluster tree and cluster history from the submission which analyzes the eight-category variable for age, which is represented by the rows of Table C.1.



Root-Mean-Square Total-Sample Standard Deviation	0.241552

Cluster History					
Number of Clusters	**Clusters Joined**		**Freq**	**R-Square**	**Between Cluster Sum of Squares**
7	**3**	**6**	2	1.00	103E-7
6	**4**	**5**	2	1.00	0.0001
5	**CL7**	**CL6**	4	.996	0.0029
4	**1**	**2**	2	.974	0.0185
3	**CL5**	**7**	5	.902	0.0586
2	**CL4**	**CL3**	7	.567	0.2732
1	**CL2**	**8**	8	.000	0.4635

The ordinal variable named *agegrp* is based on the rows of Table C.1. The following macro submission performs the clustering procedure. The SAS code which follows the macro submission adds some labels that have been created using PROC FORMAT to the rows of the cluster tree seen in Figure C.1.

```
%data_prep(
   rawdata=icu,
   cluster=yes,
   response=status,
   predvar=agegrp);

 ods html;
 ods graphics on;
 proc tree data=_tree_ horizontal pages=1;
 id rows;
 title "Ward's Cluster Tree";
 format rows agegrp.;
 run;
```

The 'R-Square' column of the cluster history below Figure C.1 suggests that, with respect to the outcome of survival, using four categories will still retain more than 95% of the information which is contained in the original eight categories. (See Section 7.1 of the main text to review the guidelines for using and interpreting the clustering results.) In addition, the cluster tree of Figure C.1 illustrates that the four categories could be specified by choosing cut-points at 35, 75, and 85 years old. (Note that the precise ranges for the resulting categories will be <35, 35-74, 75-84, and >84.) After submitting the following macro code, this new categorical age variable, which is named *agecat*, is derived from the quantitative variable of *age*.

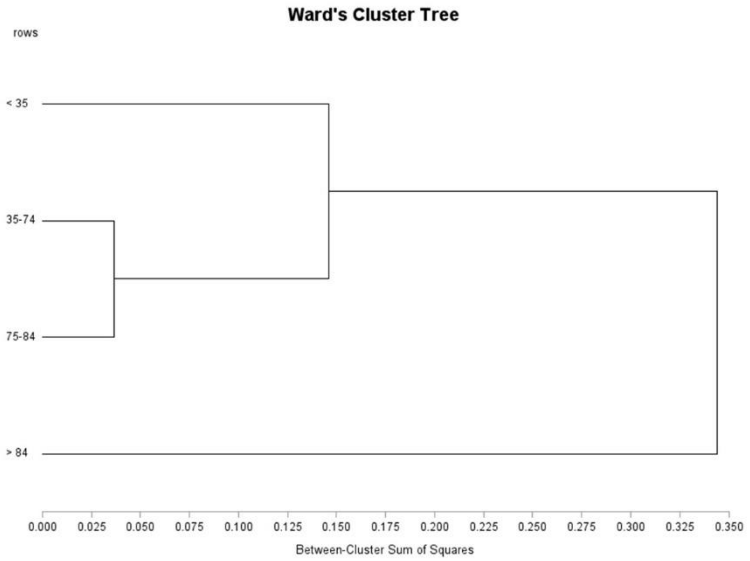
```
%data_prep(
   rawdata=icu,
   contvar=age,
   catname=agecat,
   firstlevel=1,
   numcutpts=3,
   cutpoint1=35,
   cutpoint2=75,
   cutpoint3=85,
   cluster=yes,
   response= status,
   predvar=agecat);
```

To examine another criterion for choosing the number of categories, we can examine the results from this macro submission which creates the four-category variable *agecat*. This criterion, mentioned in Section 7.1, uses the 'Root-Mean-Square Total-Sample Standard Deviation' in the SAS output. This root-mean-square estimate indicates the overall within-cluster variation for the four-category solution. Selected output from the macro submission is given in Figure C.2, and it indicates that the estimated root-mean-square is about 0.296 for the four-category age solution.

The R^2 statistics in the cluster history for either Figure C.1 or Figure C.2 also indicate that a three-category solution would still capture at least 90% of the explained variability. When the macro is submitted interactively while varying the cut-points (output not shown), the iterations determine two cut-points at 30 and 70 years of age (i.e., with insertions of 'cutpoint1=30,' and 'cutpoint2=70,' in the macro). Figure C.3 illustrates that the root-mean-square for this solution is reduced to about 0.188. Therefore, this variable with three categories (i.e., defined by the two cut-points) can be chosen and saved as the final *agecat* variable for the POSSE analysis. Note that the cluster history in Figure C.3 indicates that the number of age categories should not be reduced any further, since the R^2 results indicate that a reduction to two categories (i.e., by combining the two highest categories) would retain only about

80% of the explained variation which is found for the three-category age solution. However, the choice of cut-points is often robust. For example, it can be demonstrated that choosing the neighboring cut-points at, for instance, 35 and 75 years of age would work just as well in the POSSE analysis, whose modest aim is to determine the most unusual patterns for the data and not to provide the precise results and conclusions we would expect for formal analyses. Of course, in any subsequent modelling, a researcher would normally use the original quantitative *age* variable.

Figure C.2. The cluster tree and cluster history from the submission for the four-category variable for age.

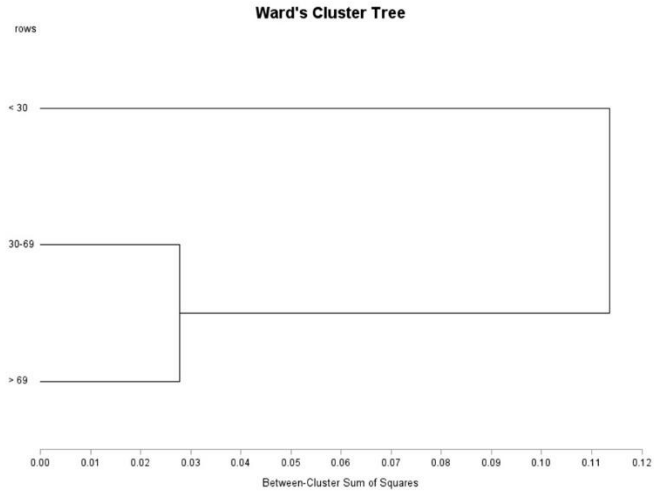


Root-Mean-Square Total-Sample Standard Deviation	0.296233

Cluster History

Number of Clusters	Clusters Joined		Freq	R-Square	Between Cluster Sum of Squares
3	**2**	**3**	2	.931	0.037
2	**1**	**CL3**	3	.654	0.146
1	**CL2**	**4**	4	.000	0.344

Figure C.3. The cluster tree and cluster history from the submission which analyzes the final three-category variable for age, which resulted from choosing cut-points at 30 and 70 years.



Root-Mean-Square Total-Sample Standard Deviation	0.188021

Cluster History

Number of Clusters	Clusters Joined		Freq	R-Square	Between Cluster Sum of Squares
2	2	3	2	.803	0.0278
1	1	CL2	3	.000	0.1136

C.3: Creating a Categorical Data Set Using the *Data_Prep* Macro

If your dataset has only categorical variables, then the POSSE macros can be applied directly to it, unless you wish to use the methods in the previous section to possibly reduce the number of categories for some variables. Otherwise, the ***data_prep*** macro provides a relatively easy way to create a new data set containing only categorical variables. A description of its features can be found in Section 8.1. An important point to remember when using this macro to create a categorical data set is that, when deriving multiple categorical variables, the user must turn on the 'newdata=yes' feature after deriving the first categorical variable. Turning on the 'newdata=yes' feature will then direct the macro to save any new categorical variables (each of which is created using separate invocations of the macro) into the same dataset which contains the first new variable. For example, subject-matter information is used to choose cut-points at 90 and 140 milliliters of mercury (Hg) to create an ordinal variable for systolic blood pressure, and to choose cut-points at 60 and 100 beats per minute to create an ordinal variable for heart rate. The first categorical variable of *agecat* is created using two cut-points at 30 and 70 years of age.

```
%data_prep(
   rawdata=icu,
   contvar=age,
   catname=agecat,
   firstlevel=1,
   numcutpts=2,
   cutpoint1=30,
   cutpoint2=70);
```

We can then submit the following, which creates the variable *bloodpress* with categories representing

levels of blood pressure from the quantitative variable named *systolic*. Because of the entry 'newdata=yes', the macro ignores the 'rawdata=' entry and uses the data set which was created during the previous macro submission.

```
%data_prep(
   rawdata=icu,
   newdata=yes,
   contvar=systolic,
   catname=bloodpress,
   firstlevel=1,
   numcutpts=2,
   cutpoint1=90,
   cutpoint2=140);
```

The final macro submission can be seen below. It creates the categorical variable *heartrate* from the quantitative variable named *hrtrate*. At the same time, it outputs the variables (the newly-created ones and other variables from the original data set) to a new data set named 'icucat' along with the ID variable.

```
%data_prep(
    rawdata=icu,
    newdata=yes,
    contvar=hrtrate,
    catname=heartrate,
    firstlevel=1,
    numcutpts=2,
    cutpoint1=60,
    cutpoint2=100,
    outdata=icucat,
    id=id,
    savevars= status gender surgery cancer renal infect
          cpr previcu emergency fracture bldoxy
          bldph bldco bldbic bldcret uncons coma
          race agecat bloodpress heartrate);
```

(Because adverse health outcomes are known to be associated with either low levels of oxygen, pH, and bicarbonate, or high levels of carbon dioxide and creatinine, the presence of these various low or high levels are classified as '1' (and zero otherwise) for the five binary variables of *bldoxy*, *bldph*, *bldco*, *bldbic* and *bldcret*.)

C.4: Screening Variables

A POSSE analysis might often begin with a general screening of the data for associations and patterns by performing some preliminary correspondence analyses using the ***prelim_ca*** macro. In some cases, the objective will be to reduce the number of variables used in subsequent analysis. We will demonstrate this screening here using the ICU data described in Section B.1. As noted in Sections 8.2 and 8.3 of the main text, it is necessary to create a format data set that contains the labels for the categories of the explanatory variables and the covariates prior to invoking either the ***prelim_ca*** or ***correspondence*** macros. For the present example, the following code creates labels and saves them to a data set called 'othrfmt1'. Formats are not required for response variables, but if the user wishes to provide a format for the response variable named *died*, then it can be given in a separate PROC FORMAT statement without using the 'cntlout=' option, or, as is done here, included with the formatting for the other variables.

```
proc format cntlout=othrfmt1;                    *** required formats ***;
value status 0='Surv' 1='Died';
value gender 0='Male' 1='Female';    value previcu 0='NoPrev' 1='Prev';
value surgery 0='NoSurg' 1='Surg';   value emergency 0='NoEmer' 1='Emer';
value cpr 0='NoCPR' 1='CPR';         value uncons 0='Cons' 1='UnCons';
value cancer 0='NoCanc' 1='Canc';    value fracture 0='NoFr' 1='Fract';
value race 0='NonWh' 1='White';      value agecat 1='Age1' 2='Age2' 3='Age3';
value renal 0='NoRen' 1='Renal';     value infect 0='NoInf' 1='Infect';
value bldoxy 0='HiOxy' 1='LoOxy';    value bldph 0='NoPH' 1='LoPH';
value bldco 0='LoCO2' 1='HiCO2';     value bldbic 0='NoBIC' 1='LoBIC';
value bldcret 0='LoCr' 1='HiCreat';
value bloodpress 1='LoBP' 2='MedBP' 3='HiBP';
value heartrate 1='LoHR' 2='MedHR' 3='HiHR';
run;
```

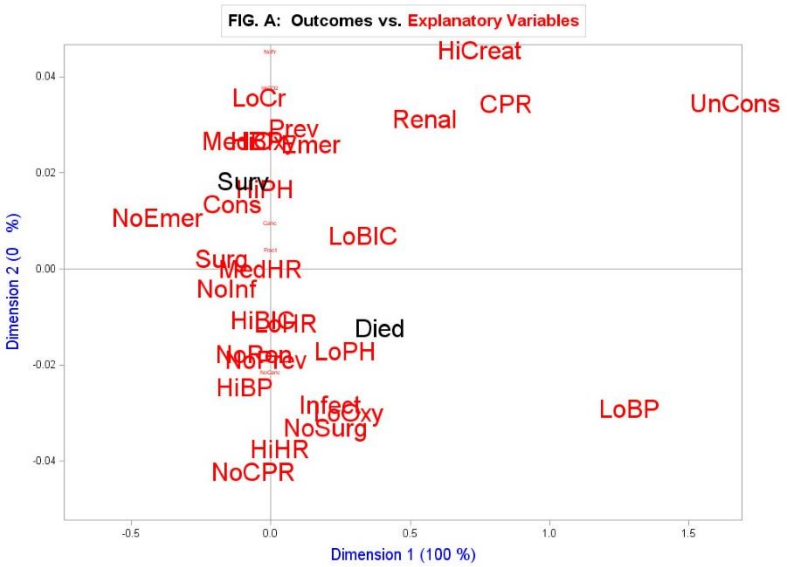
Once these labels are created, the ***prelim_ca*** macro is submitted using the code below, where the format data set is identified in the 'fmtothr=' entry of the macro. This submission performs three correspondence analyses between (a) the response and explanatory variables, (b) the response and covariates, and (c) the explanatory variables and covariates. As mentioned in Sections 8.2 and 8.3 in the text, a simple binary response variable will result in a one-dimensional solution in correspondence analysis, which then requires the 'onedim=yes' insertion seen below. However, this applies only to the correspondence analyses for (a) and (b). The solution for the third analysis (c) has a dimension of six, one less than the total number of categories for the covariates.

```
%prelim_ca(
    dataset=icudat,
    response=status,
    fmtresp= status,
    explanvars=surgery cancer renal infect cpr previcu emergency fracture
          bldoxy bldph bldco bldbic bldcret uncons bloodpress heartrate,
    covars=gender agecat race,
    fmtothr=othrfmt1,
    id=id,
    onedim=yes);
```

Results for the correspondence analyses for (a) and (c) from the above macro submission are illustrated here in Figures C.4 and C.5. Figure C.4 contains the results for the first correspondence analysis (a). Since the figure represents a one-dimensional solution, the points represented by the labels should actually lie along the x-axis, but they have been randomly spread out vertically in order to avoid the overplotting of labels. The map in Figure C.4 suggests that the explanatory factors of consciousness and blood pressure are contributing the most to the dimension. (See Sections 5.2, 5.4 and 9.1 to review the guidelines for interpreting correspondence maps.) The table below Figure C.4 lists the inertias for the tables which cross-tabulate the response variable *died* with the 16 explanatory variables listed in the macro submission. The higher inertias for these factors of consciousness and blood pressure confirm what is suggested by the figure. To help interpret these results, we use the ***tabulation*** macro to produce Tables C.2(a) and C.2(b). The first table is constructed using the following submission.

```
%tabulation(
    dataset=icudat,
    response= status,
    fmtresp= status,
    secondvar=uncons,
    id=id);
```

Figure C.4. Output for the first of three correspondence analyses from the ***prelim_ca*** macro submission for the data from the ICU study[45-46]. Because this is a one-dimensional map, all the inertia is in the first dimension. Although all the labels would ideally be plotted on the horizontal axis, a small random increment has been added to the vertical coordinate of the labels in order to avoid overplotting.



Inertias for Contingency Tables Associated with Fig. A

	OUTCOMES
	status
EXPLANATORY VARIABLES	
bldbic	0.009
bldco	0.000
bldcret	0.030
bldoxy	0.007
bldph	0.005
bloodpress	**0.124**
cancer	0.000
cpr	0.050
emergency	0.059
fracture	0.000
heartrate	0.001
infect	0.033
previcu	0.001
renal	0.032
surgery	0.034
uncons	**0.225**

Table C.2(a) illustrates that mortality was about 87% (13 of 15 subjects) for those who were unconscious at admission, versus about 15% for those who were conscious. The second table is constructed using the following submission.

```
%tabulation(
   dataset=icudat,
   response= status,
   fmtresp= status,
   secondvar=bloodpress,
   id=id);
```

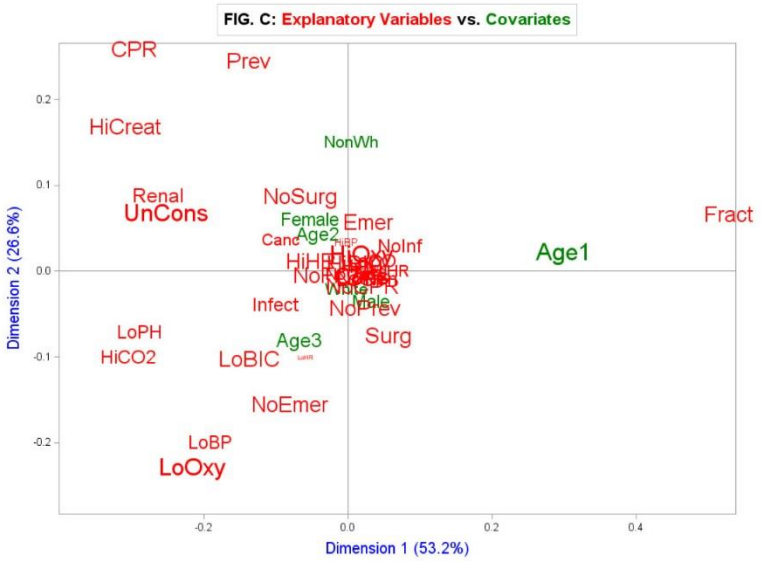
Table C.2(b) illustrates that the mortality is about 70% (10 of 14) for the subjects with low blood pressure, in contrast to about 16% for the other two categories for blood pressure. (This second result also suggests that the relationship of blood pressure with mortality is not linear.)

Table C.2(a)-(b). The cross-tabulation of the vital status versus (a) level of consciousness and (b) the blood-pressure categorical variables which are derived from the quantitative variables found in the ICU data[45-46].

(a)	Survived	Died	Total
Conscious	158 (85 %)	27 (15 %)	185
UnConscious	2 (13 %)	13 (87 %)	15
Total	160 (80 %)	40 (20 %)	200

(b)	Survived	Died	Total
Low Blood-Pressure	4 (29 %)	10 (71 %)	14
Mid Blood-Pressure	89 (84 %)	17 (16 %)	106
High Blood Pressure	67 (84 %)	13 (16 %)	80
Total	160 (80 %)	40 (20 %)	200

Figure C.5. Partial output for the third of three correspondence analyses from the ***prelim_ca*** macro submission for the data from the ICU study[45-46]. The table of inertias has been edited to include only those columns (i.e., the explanatory variables) with the largest inertias.



Inertias for Contingency Tables Associated with Fig. C

	EXPLANATORY VARIABLES									
	bldco	**bldph**	**bloodpress**	**cpr**	**emergency**	**fracture**	**infect**	**previcu**	**renal**	**surgery**
COVARIATES										
agecat	0.028	0.012	0.017	0.019	0.023	**0.075**	0.023	0.021	0.019	0.023
gender	0.014	0.007	0.005	0.016	0.014	0.001	0.001	0.006	0.004	0.006
race	0.006	0.010	0.012	0.001	0.003	0.003	0.011	0.019	0.007	0.005

To conserve space, the table of inertias for correspondence analysis (c) has been edited, so that the table below Figure C.5 contains only the results for the ten explanatory variables with the highest inertias with

respect to the covariates. Figure C.5 suggests an association between age and the presence of fracture. In the table of inertias below Figure C.5, the inertia for the age and fracture variables is substantially larger than the other inertias. To interpret this result, Table C.3 is constructed using the following submission.

```
%tabulation(
   dataset=icudat,
   response=fracture,
   fmtresp=fracture,
   secondvar=agecat,
   id=id);
```

Table C.3 illustrates that the proportion of fracture was about 24%, 7% and 2% for the three age categories. However, since there was little evidence of fracture being related to survival, this result might be ignored if a researcher later examines models for predicting mortality.

Table C.3. The cross-tabulation, together with the row percentages, for the presence of fracture versus the categorical age variable.

Age Category (years)	No Fracture	Fracture	Total
Age1 (< 30)	22 (76 %)	7 (24 %)	29
Age2 (30 – 69)	100 (93 %)	7 (7 %)	107
Age3 (≥ 70)	63 (98 %)	1 (2 %)	64
Total	185 (93 %)	15 (8 %)	200

C.5: Examining the Scaling for a Set of Variables

In this section we examine the scaling results from the ICU data (Section of B.1), in order to detect and describe possible dependencies among the variables *cpr*, *uncons*, *emergency*, *surgery*, *cancer* and *previcu*. Section 3.2 of the main text introduces the concept of Guttman scaling, and Section 6.3 describes a special application of homogeneity analysis which can indicate this scaling by examining the order of variable levels along the first dimension of a map. The first part of this analysis uses the following macro submission, where the 'sub=all' entry causes a homogeneity analysis to be performed on all the variables listed in the 'var=' line. As mentioned in Section 8.4 of the main text, the variable names listed in the ***classification*** macro should be no more than seven letters in length, so the variable *emergency* has been renamed here as *emerg* in a data step not shown here.

```
%classification(
   data=icudat,
   var=surgery cancer cpr previcu emerg uncons,
   sub=all,
   noprint=yes);
```

The output from this submission includes the map seen in Figure C.6. We first observe the ordering of the positive outcomes along the first dimension of the map. The ordering of categories from left to right is first *cpr1*, and then *uncons1*, *emerg1*, *previcu1*, *surgery1*, and finally *cancer1*. However, note that the two levels of the *previcu* variable are primarily aligned with the second-dimensional axis in the map. The discrimination measures (not shown in our output here) for the *previcu* variable are 0.018 and 0.337 for the first and second dimensions, which are, respectively, the smallest and largest values for those dimensions. Because the *previcu* variable is contributing primarily to the second dimension and because

this scaling analysis is focused on the contributions to the first dimension, we can exclude the *previcu* variable from the rest of the analysis. When we ignore this variable, the other categories suggest a 'horse-shoe' or 'arch' effect in Figure C.6.

Figure C.6. The first map from the homogeneity analysis, which analyzes the scaling for six variables from the ICU study[45-46].

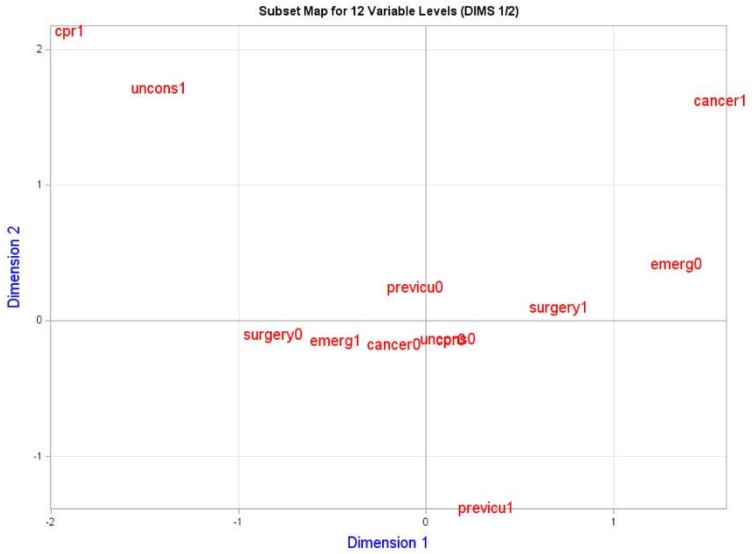


Table C.4. The various profiles and their frequencies for five variables determined and reported at admission in the ICU data[45-46], where a blank cell represents a negative outcome for a variable. The rows of the table are ordered by the first-dimensional coordinates for the profiles in the homogeneity map of Figure C.6.

cpr	uncons	emerg	surgery	cancer	Frequency	Cumulative Frequency
1	1	1			4	4
1		1			7	11
1	1	1	1		1	12
	1	1			5	17
1		1		1	1	18
		1			73	91
	1	1	1		4	95
		1	1		48	143
					1	144
		1		1	2	146
	1		1		1	147
			1		36	183
		1	1	1	2	185
			1	1	15	200

To further examine the remaining five variables, we generate Table C.4 for the frequencies for the 200 profiles of zeros and ones, ordered by their first-dimensional coordinates. For this submission, seen below, it is necessary to list the remaining five variables in the 'haclust=' line in the order in which they

were identified above. This new submission then saves the results of the cluster analysis to a temporary data set named 'tempclust', and the final entry of 'printfreqs=yes' creates the output of Table C.4 from this temporary data set. (Note that the blank cells in the Table C.4 represent the zeros in the data set.)

```
%classification(
    data=icudat,
        haclust=cpr uncons emerg surgery cancer,
        nclust=2, /* this and 'out=' entries are necessary to use the 'printfreqs=' option */
        id=id,
        allbin=yes,
        out=tempclust,
        printfreqs=yes);
```

Table C.4 illustrates that the variables of *cpr* (whether CPR is performed prior to admission) and *emergency* (whether the type of admission was an emergency or not) follow a Guttman scale, where all patients who received CPR had an emergency admission, but not vice versa. Further examination of Table C.4 illustrates that there were four profiles which involved a single subject; these profiles are seen directly below along with their subject ID and their vital status at discharge. For instance, subject #597 is associated with the row of all zeros (i.e., blank cells) in Table C.4.

Profiles Involving a Single Subject						
ID=**331**	CPR	Unconscious	Emergency	Surgery	No Cancer	Died
ID=**766**	CPR	Conscious	Emergency	No Surgery	Cancer	Survived
ID=**597**	No CPR	Conscious	Non-Emergency	No Surgery	No Cancer	Survived
ID=**208**	No CPR	Unconscious	Non-Emergency	Surgery	No Cancer	Died

When we remove the rows in Table C.4 with frequencies of only one or two subjects, then a more pronounced pattern emerges for the remaining 192 subjects, as seen in Table C.5. A comparison of the two tables indicates that, when subject #208 is removed from the data, the variables of *emergency* and *uncons* also follow a Guttman scale, where an unconscious person is always considered an emergency, but not vice versa. In other words, these two variables can be said to follow a quasi-Guttman scale for the full dataset. Table C.6 indicates a stratified cross-tabulation with mortality for the *emergency*, *uncons* and *cpr* binary variables. The bolded and underlined cell entry in the table is from subject #208.

Table C.5. The subset of profiles from Table C.4 which have frequencies greater than two. This reduces the number of subjects from 200 to 192.

cpr	uncons	emerg	surgery	cancer	Frequency
1	1	1			4
1		1			7
	1	1			5
		1			73
	1	1	1		4
		1	1		48
			1		36
			1	1	15

As suggested in Section 3.2, it might be useful to combine variables which follow a Guttman or quasi-Guttman scale. A new omnibus variable is temporarily created from the three binary variables identified above. Its categories are represented by the rows of Table C.7. As you can see, this new variable is

ordinal with respect to the mortality percentages in the far-right column of the table.

Table C.6. A stratified cross-tabulation for three binary variables with the vital status. The bolded and underlined cell entry belongs to subject #208 of the ICU study[45-46].

		Non-Emergency		Emergency		Totals
		Survived	Died	Survived	Died	
No CPR	**Conscious**	51	1	101	24	177
	Unconscious		**1**	2	7	10
CPR	**Conscious**			6	2	8
	Unconscious				5	5
Totals		51	2	109	38	200

Table C.7. The cross-tabulation for the vital status versus an omnibus variable which is created from the three binary variables of *emergency*, *cpr*, and *uncons*.

Omnibus Variable	Survived	Died	Totals	Mortality
Non-Emergency	51	1	52	2%
Emergency	101	24	125	19%
Emergency/CPR	6	2	8	25%
Emergency/Unconscious	2	7	9	78%
Emergency/Unconscious/CPR		5	5	100%
Subject #208 (Non-Emergency/Unconscious)		1	1	100%
Totals	160	40	200	20%

(Technical notes: Tables C.6 and C.7 indicate that the inability of the prior CPR to restore consciousness at admission is a perfect predictor of mortality. However, if we use this fact in a logistic regression model, it leads to the calculation of an infinite parameter and a 'quasi-completeness' condition for which SAS issues a warning. An alternate approach such as that by Firth[53] can be utilized in this case. This option is available in the SAS procedure PROC LOGISTIC[37]. The Guttman scaling also has implications for the modeling. For example, the *cpr* and *emergency* variables follow a Guttman scale. If we wish to know the effect of the two factors acting together, the corresponding interaction term *cpr*emergency* is determined by whether or not CPR and an emergency admission both occurred. However, because of the dependency implied by the Guttman scaling, the effects associated with CPR are aliased or confounded with the effects for the *cpr*emergency* interaction.)

C.6: Inspecting an Outlier

In the previous section we examined the ICU data (Section B.1) and found a dependency between being an emergency admittance and being unconscious, with the exception of subject #208. In this section we will define a new omnibus variable in order to further examine this outlier. This new variable named *condition* is defined as follows:

```
data icudat;
 set in.icucat;
** new variable **;
if (emergency=0 and uncons=0)      then condition=1;
```

```
else if (emergency=1 and uncons=0) then condition=2;
else if (emergency=1 and uncons=1) then condition=3;
                                        else condition=4;
run;
```

The fourth category of the new variable *condition* is associated with just subject #208, and the labels assigned to the categories of *condition* are as follows:

```
value condition 1='NoEmerg' 2='EmergCons' 3='EmergUncons' 4='Subj208';
```

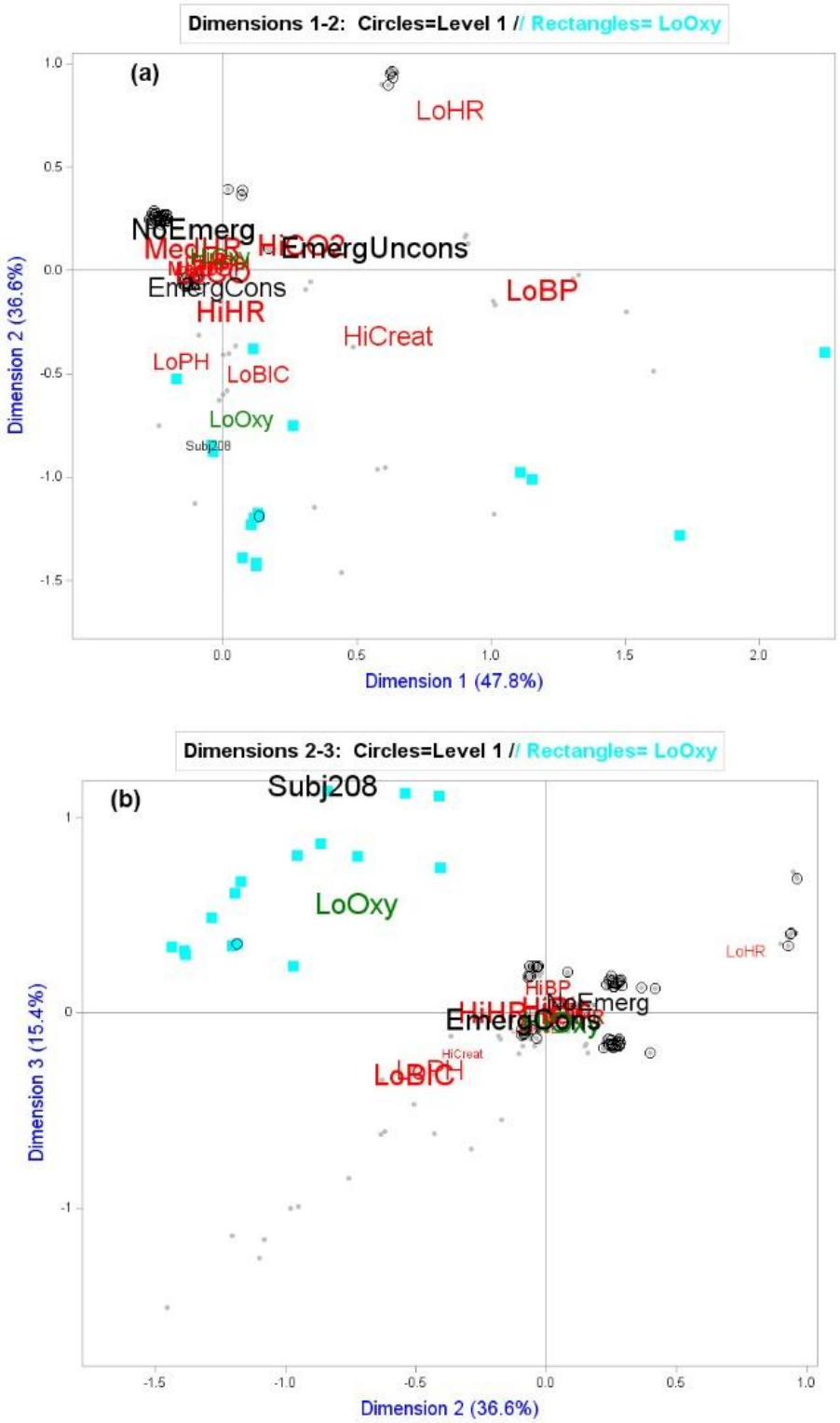
The outcomes for any single variable for subject #208 are not unusual, so this observation is not a univariate outlier. It is the combination of factors that makes it unusual, so it can be characterized as a multivariate outlier. Subject #208 was classified as a non-emergency at admission, but his or her condition was serious enough to require surgery and eventually resulted in death. This patient's level of consciousness was originally classified as 'stupor' and all five of the patients with the condition of 'stupor' died, in contrast to eight of ten with the condition of 'coma'. Therefore, the condition of 'stupor' appears to be just as critical as the condition of 'coma' with respect to survival. We would, therefore, expect the admission of subject #208 to be classified as an emergency, but cannot determine whether an error in the data entry or some other reason led to his or her classification as a non-emergency. To further analyze this, we will regard the outcome for subject #208 as a missing value in order to investigate what other value for the variable for *condition* we might substitute for it. This process is usually referred to as 'imputation'. The results for a correspondence analysis can be used to informally impute our missing value into one of the three other categories of *condition* using the vital signs and bloodwork variables.

One strategy for imputation would be to choose a radius in the map about a missing value, which would then be imputed using the most common value for the observations within the radius with non-missing values. Section 9.11 of Greenacre[25] uses a similar approach to make predictions, but his correspondence analysis is two-dimensional. That means that all of the inertia is contained in a two-dimensional correspondence plot. However, the correspondence analysis for the following submission is three-dimensional, so here we examine results for three dimensions.

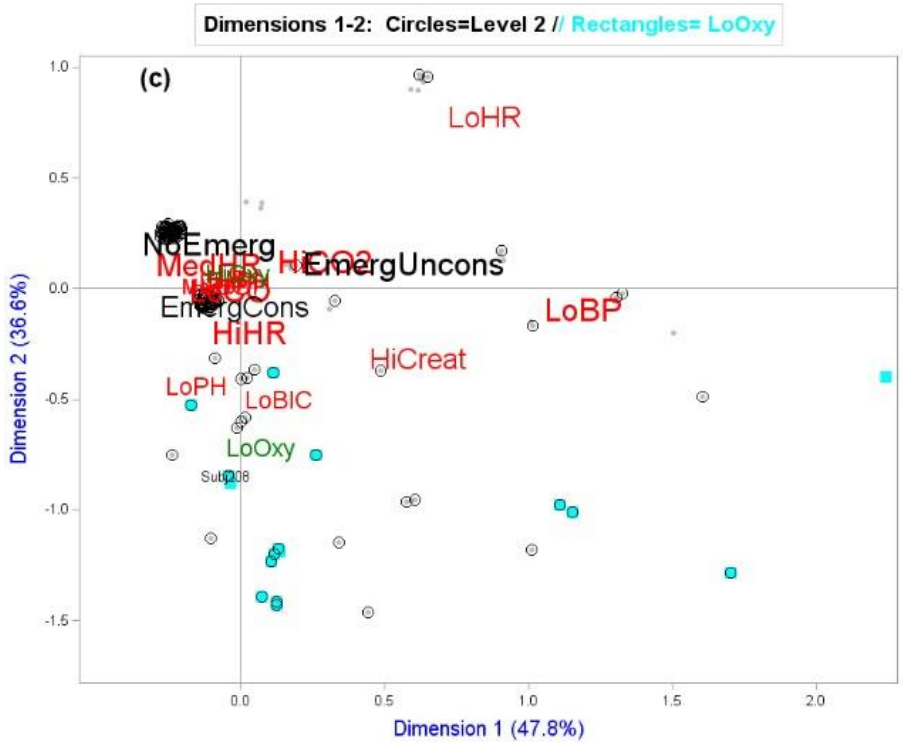
```
%correspondence(
   dataset=icudat,
   response=condition,
   fmtresp=condition,
   explanvars=bloodpress heartrate bldph bldco bldbic bldcret,
   covars=bldoxy,
   fmtothr=othrfmt1,
   id=id,
   showobs=yes,
   highlightobs=LoOxy,
   circlelevel=1);   /* Insert '2' here and resubmit the macro to produce Figures C.7(c)-(d) */
```

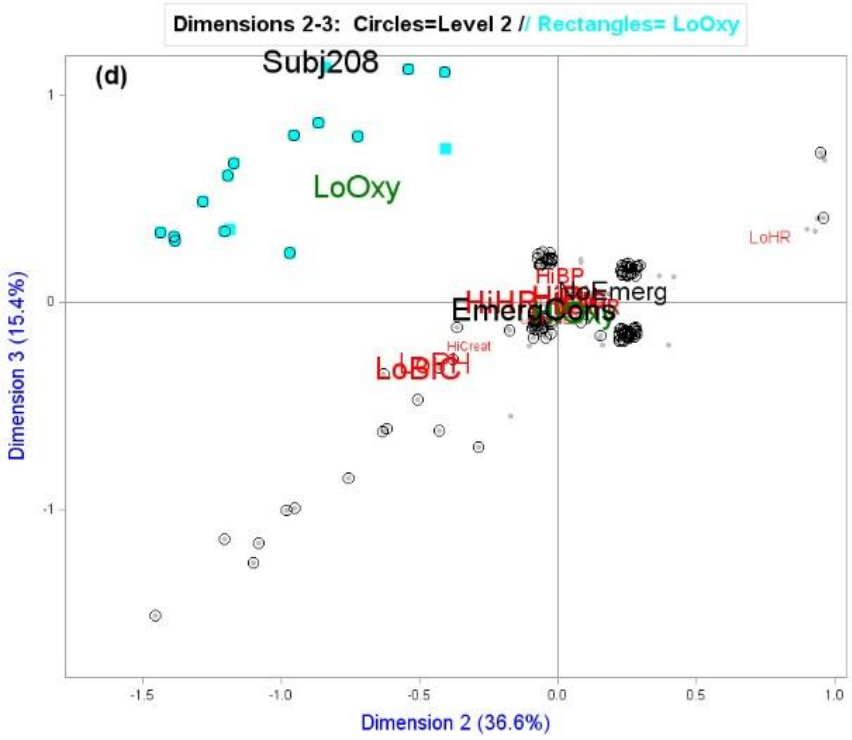
The submission above creates the correspondence map of Figures C.7(a)-(b). A second submission with the entry of 'circlelevel=2' creates the other correspondence maps in Figures C.7(c)-(d) in the second row. All the figures plot the observations ('showobs=yes') in the correspondence maps. The circled observations in Figures C.7(a)-(b) represent the observations for the first category of the response variable, those conscious patients with a non-emergency admission ('circlelevel=1'). The circled observations in Figures C.7(c)-(d) represent the observations for the second category of the response variable, those conscious patients with an emergency admission ('circlelevel=2'). The solid squares in all the figures represent those patients, including subject #208, who had low blood oxygen levels at admission.

Figures C.7(a)-(d). The maps for the correspondence analysis which uses the new *condition* variable as the response, where the fourth level of the variable is represented by the missing value of subject #208 of the ICU study[45-46]. Maps in the same column are identical except for the fact that the circled points represent patients classified into the first category of the response variable in the upper figure, whereas they represent patients classified into the second category in the lower figure. The solid squares in all figures represent patients who have low oxygen levels. (See next page for Figures C.7(c)-(d).)



Figures C.7(c)-(d).





The point representing subject #208 in Figures C.7(a)-(b) is generally far from the circled observations, which represent patients that have been classified as non-emergencies. However, the point representing subject #208 in Figures C.7(c)-(d) is in the same general vicinity of the circled observations, which represent patients who are classified as emergencies. Since the objective measurements of vital signs and blood gases were presumably used to determine the type of admission, then subjected #208 would likely be imputed into the category of patients who are classified as emergencies. Therefore, the correspondence results suggest that subject #208 is misclassified as a non-emergency, assuming that the objective measurements for subject #208 are accurate.

Note that in Figure C.7(a), the circled point below that for subject #208 represents subject #597, who was also among the four unusual observations identified in the previous section. Please also note that these two points are easier to see on a computer monitor than on a printed copy of Figure C.7.

(Technical Note: For those who are interested, Audigien et al.[54] have provided a rigorous method for using correspondence analysis for imputation.)

C.7: Deriving a Cluster Variable

In this section we continue examining some variables from the ICU data (Section of B.1), in order to create a cluster variable from a subset of the variables. We remind the reader that, because adverse health outcomes are known to be associated with either low levels of oxygen, pH, and bicarbonate, or high levels of carbon dioxide and creatinine, the presence of these various low or high levels have been classified as '1' for the five binary variables, and classified as zero otherwise. Therefore, clusters with relatively unhealthy outcomes will be distinguished by having relatively large proportions (e.g., 0.70 – 1.00) of their subjects who are classified as '1' for one or more of these binary variables.

One possible objective for the ICU study would be to determine the contributions to predicting mortality for the objective measures of vital signs (heart rate and blood pressure) and the five results from the initial bloodwork. During the initial analysis, we submit the following macro to perform a homogeneity analysis on the seven variables. Figure C.8 illustrates the resulting scatterplot for the discrimination measures for the first two dimensions.

```
%classification(
    data=icudat,
    var=oxy ph co bic creat bp hr);
```

Because Figure C.8 indicates that the largest discrimination measures belong to the bloodwork variables, we will analyze them separately from the vital signs. A subset homogeneity analysis is then performed for the five bloodwork variables using the following macro submission.

```
%classification(
    data=icudat,
    var=oxy ph co bic creat bp hr,
    sub=oxy ph co bic creat);
```

The output includes Figure C.9, where the relatively close proximity of three outcomes in the lower right corner of the map suggest relatively frequent co-occurrences for high levels of carbon dioxide and low levels of oxygen and pH. A cluster analysis can characterize subjects with respect to a set of variables, and suggest whether certain outcomes tend to co-occur more often for some groups. This, in turn, might

also suggest what variable interactions should be included in a statistical model. (More will be said about this in Section C.8.)

Figure C.8. The scatterplot for the discrimination measures for the vital signs and bloodwork variables from the ICU study[45-46].

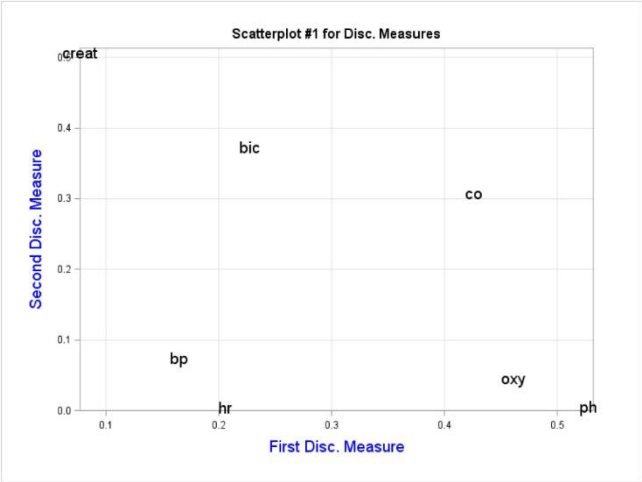
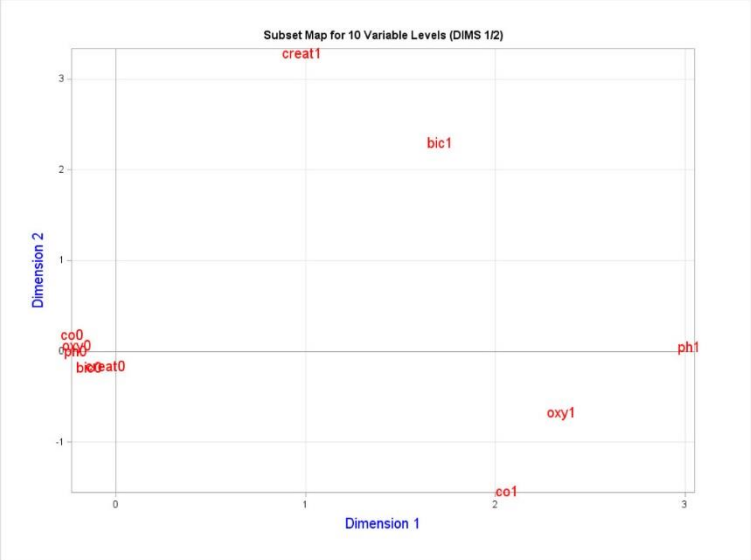


Figure C.9. The map for the subset homogeneity analysis for the bloodwork variables from the ICU study[45-46].

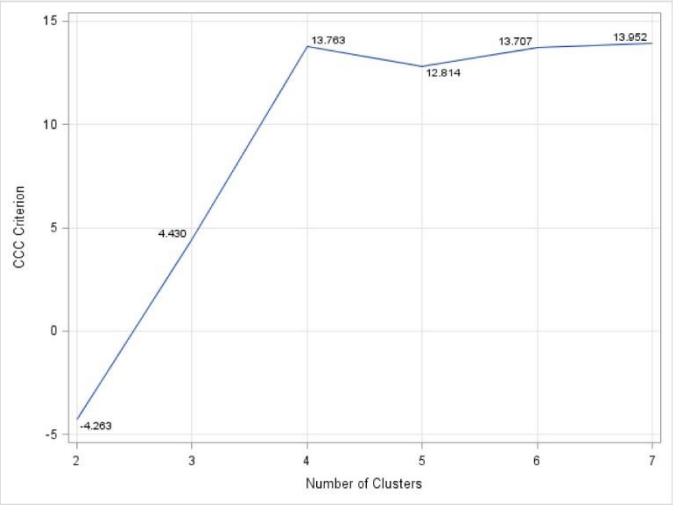


To perform a cluster analysis for the bloodwork variables, we first use the 'fitclust=yes' and 'haclust='options in the ***classification*** macro, as seen in the submission below. This produces the calculated CCC statistics in Figure C.10 for a series of two to seven clusters.

```
%classification(
    data=icudat,
    haclust=oxy ph co bic creat,
```

```
fitclust=yes,
id=id);
```

Figure C.10. The plot of the CCC statistic for a series of cluster sizes for the bloodwork variables from the ICU study[45-46].



As mentioned in Section 6.2, a CCC statistic of at least 2.0 usually indicates well-defined clusters. We generally recommend picking the cluster number associated with either the largest overall CCC value or the largest local maximum, where the local maxima are indicated by the peaks in figures displaying the CCC values. The one and only local maximum indicated by Figure C.10 is for the four-cluster solution with a peak CCC value at about 13.7, which is larger than either the three-cluster or five cluster solutions on either side of it. (Even though the seven-cluster solution has a slightly higher CCC value, it would result in some clusters having very few members.) The macro submission below saves the information for the four-cluster solution to a dataset called 'bloodclusters' and produces the bloodwork profiles for the four clusters seen in Table C.8.

```
%classification(
    data=icudat,
    haclust=oxy ph co bic creat,
    nclust=4,
    id=id,
    allbin=YES,
    out=bloodclusters);
```

Table C.8. The bloodwork profiles for the four clusters determined using homogeneity analysis.

Blood-Gases	clust1 n=10	clust2 n=169	clust3 n=11	clust4 n=10
Low Blood BIC	0.00	0.00	1.00	0.40
High Blood CO2	1.00	0.05	0.09	0.00
High Blood Creatinine	0.00	0.00	0.00	1.00
Low Blood Oxygen	0.70	0.02	0.36	0.10
Low Blood pH	0.80	0.00	0.27	0.20

In Table C.8, the profile for the first cluster of ten subjects does indicate the co-occurrences for the low pH, low oxygen and high carbon dioxide measurements that are suggested by the map in Figure C.9. However, the last two clusters are dominated by a single blood-gas outcome, and the second cluster is identified with normal blood gas measurements at admission. The cluster assignments are saved as a cluster variable named *bloodcluster* and we use the profiles from the table to characterize and label the four categories as follows:

1='HiCO2' 2='Norm' 3='LoBIC' 4='HiCreat'

The ***tabulation*** macro is used to produce the frequencies in Table C.9, where we can see that the estimated mortality in the bottom row is higher for the third and fourth clusters. Three of the cluster sizes are small and, therefore, any cross-tabulations with other variables will produce sparse tables. In cases where the cluster sizes are larger, we might decide to use the cluster variable as the response variable in a simple correspondence analysis to explore its relationship to the other variables. It should also be emphasized that the interpretation of results from cluster analysis can benefit from subject-matter experts. For example, the first cluster might be connected with a condition called respiratory acidosis.

Table C.9. The cross-tabulation for the vital status versus the four levels of the cluster variable, together with the percentage of mortality for each cluster.

	HiCO2	Norm	LoBIC	HiCreat	Total
Survive	8	140	7	5	160
Died	2	29	4	5	40
Total	10	169	11	10	200
Mortality	20%	17%	36%	50%	20%

(Technical note: For a choice of four clusters, the clusters found using the advanced method of latent class analysis[55] are similar to those found here. The largest difference is for their fourth clusters, which is identified with subjects with low oxygen using latent class analysis. The fourth cluster for the POSSE analysis is identified with subjects with high creatinine.)

C.8: Exploring Main Effects and Interactions

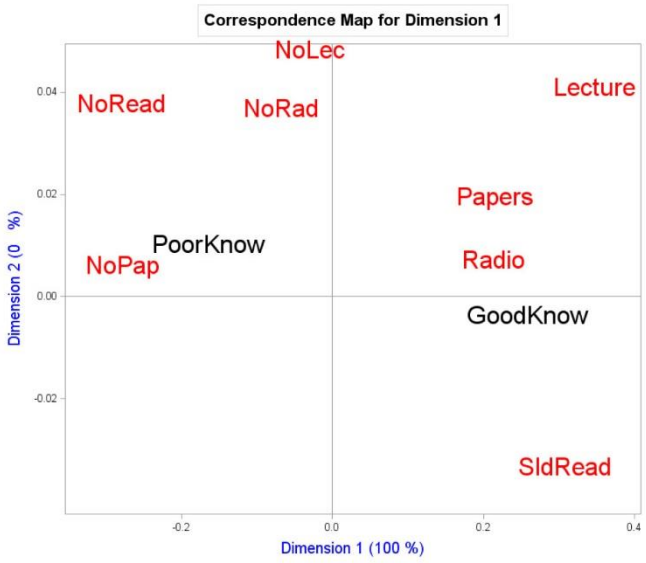
This example explores the main effects and interactions connected with Table B.5. The analysis is applied to the data set that is created in Section C.1. It examines the response variable named *knowledge*, which rates knowledge about cancer as either poor or good, and its relation to four other binary explanatory variables. These variables are named (and described) as (a) *papers* (whether or not a person read newspapers), (b) *reading* (whether or not a person was considered a solid or thorough reader), (c) *lectures* (whether or not a person attended lectures), and (d) *radio* (whether or not a person listened to the radio).

A simple, one-dimensional correspondence analysis with *knowledge* as the response variable is first performed using the following macro submission.

```
%correspondence(
   dataset=cancer_information,
   response=knowledge,
   fmtresp=knowledge,
   explanvars=radio reading papers lectures,
```

```
fmtothr=othrfmt1,
id=id,
onedim=yes);
```

Figure C.11. Output for the one-dimensional correspondence analysis with the response variable of *knowledge* for the cancer information study[51].



Contributions of Outcomes to Inertia

Variable Levels	Contrib. to 1st Dim.
GoodKnow	0.614
PoorKnow	0.386

Contributions of Other Variables to Inertia

Variable Levels	Contrib. to 1st Dim.
NoRad	0.020
Radio	0.064
NoRead	**0.239**
SldRead	**0.265**
NoPap	**0.198**
Papers	**0.154**
NoLec	0.005
Lecture	0.055

An examination of the inertias below the map of Figure C.11 suggests that, among the explanatory variables with possible associations with the 'GoodKnow' response, solid reading is the most effective main effect for knowledge of cancer, followed by being a newspaper reader. The other factors of radio and lectures are less effective. These conclusions are consistent with those of Winsor[56].

The next part of the analysis follows the reasoning given by Melamed et al.[57], who discuss the duality that often exists between the clustering of observations and the interactions among variables. They develop a rigorous method for determining interactions using cluster analysis. Here we illustrate an informal approach using the POSSE methods. The following macro submission produces Figure C.12, which displays the CCC statistics. (The length for variable names for the ***classification*** macro should be

no more than seven, so the data step before the macro submission assigns temporary shortened names to the variables.)

```
data cancer_info;
 set cancer_information;
array x[5] knowledge lectures papers radio reading;
array y[5] know lec pap rad read;
do i=1 to 5;
 y[i]= x[i];
end;
keep id know lec pap rad read;
run;

%classification(
    data=cancer_info,
    haclust=know lec pap rad read,
    fitclust=yes,
    id=id);
```

Figure C.12. The plot of the CCC statistic for a series of cluster sizes for the five variables for the cancer information data[51].

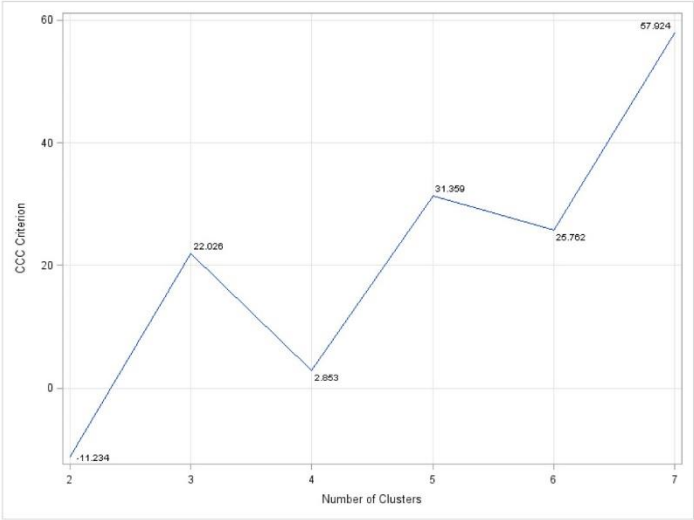


Figure C.12 illustrates that the results for the CCC statistics have several peaks. Sarle[31] has stated that there can be several peaks when the data structure is hierarchical, and an analysis similar to the one in Section A.5 would reveal, for example, that those who attend lectures are generally either solid readers or readers of newspapers. In cases where there are multiple peaks, we suggest choosing the number of clusters associated with the largest local maximum, where the local maxima are indicated by the peaks in the CCC values in Figure C.12. A further result (not shown here) for a submission of the macro with the entry 'nclust=8' yields a CCC statistic of about 59.86. Therefore, since Figure C.12 illustrates that the CCC statistic for seven clusters is about 59.9, the solution for seven clusters also appears to be a local maximum. However, it also has two relatively small clusters with only 26 and 31 subjects, respectively, so it will be more difficult to generalize this solution. For this reason, we choose the solution for five clusters, which is generally consistent with the solution for seven clusters. Using the following macro submission, we produce the results for five clusters, which is associated with a local maximum of about 31.

```
%classification(
    data=cancer_info,
    haclust=know lec pap rad read,
    nclust=5,
    id=id,
    allbin=yes); /* all variables are binary (which simplifies some output) */
```

The output includes Table C.10, which illustrates the profiles for the five clusters. The first two clusters, which are by far the largest clusters, demonstrate that low and high proportions of good knowledge are associated, respectively, with low and high proportions of newspaper readers and solid readers. This implies that we might expect a substantial *papers*reading* two-way interaction effect for a model with the response variable of *knowledge*. The much smaller third cluster would suggest a less important three-way *papers*reading*lectures* interaction. Although these informal results are no substitute for the rigorous methods of Goodman[58], they are generally consistent with his findings.

Table C.10. The profiles for the five-cluster solution for the variables from the cancer information study[51].

Condition	clust1 n=716	clust2 n=622	clust3 n=86	clust4 n=256	clust5 n=49
Good Knowledge	0.12	0.72	0.70	0.25	0.31
Attends Lectures	0.00	0.00	1.00	0.00	1.00
Reads Newspapers	0.22	0.89	0.87	0.63	0.53
Listens to Radio	0.00	0.16	0.27	1.00	0.63
Solid Reader	0.12	0.88	0.92	0.39	0.24

C.9: Assessing General Associations

(Note: Because of potential copyright issues, the data set and SAS code for this example are not available for download at the POSSE website.)

This section will use the diabetes data set[47] described in Section B.2 in order to look for the linear and quadratic associations among variables. This is somewhat equivalent to a correlation analysis near the beginning of a data analysis, although the POSSE analysis is not just looking at pairwise comparisons of the variables, but examining them in a global approach, and is also not restricted to a search for linear associations.

The data contains information concerning the glucose control for 68 patients with less than 25 years of diabetes. The POSSE data set named 'glucat' contains the variable of *control* with three categories (low/average/high glucose control) and the binary variables of *gender* (1=male, 2=female) and *education* (1=less than 13 years of formal schooling, 2=at least 13 years of formal schooling). The categorical variable of *duration*, which indicates the duration of illness, has three categories (short/average/long). The last four variables with three categories (low/average/high) represent the outcomes for some patient scores. These four variables are (a) *knowledge* (the patient's knowledge about diabetes), (b) *fate* (the patient's sense of fatalism), (c) *system* (the patient's sense that systematic, social forces control his or her future), and (d) *reliance* (the patient's self-reliance or the sense that he or she has control over the future).

We begin the analysis with the following submission of the ***classification*** macro, which performs a homogeneity analysis. As noted in Section 8.4, the variable names used in this macro are required to be no more than seven letters in length, so we first used the ARRAY statement in a data step to assign some temporary variable names.

```
data glucat;
 set glucat;
array x[8] control knowledge fate system reliance duration education gender;
array y[8] cont know fate sys rel dur edu gend;
do i=1 to 8;
 y[i]= x[i];
end;
keep cont know fate sys rel dur edu gend;
run;

%classification(
    data=glucat,
    var=cont know fate sys rel dur edu gend);
```

Among the output are the two scatterplots for the discriminations measures found in Figures C.13(a)-(b). We will first discuss the results for the third dimension. The second scatterplot of Figure C.13(b) suggests that the discrimination measures for the third dimension are dominated by the *gender* variable and the response variable of *control*. This result is illustrated by the relatively weak association indicated by Table C.11, where the row percentages reveal that 40% of the females are in the high-control category, in contrast to about 24% of the males.

Figures C.13(a)-(b). Scatterplots using the diabetes data[47] for the first-dimensional versus the second-dimensional discrimination measures, and also for the second- versus the third-dimensional discrimination measures.

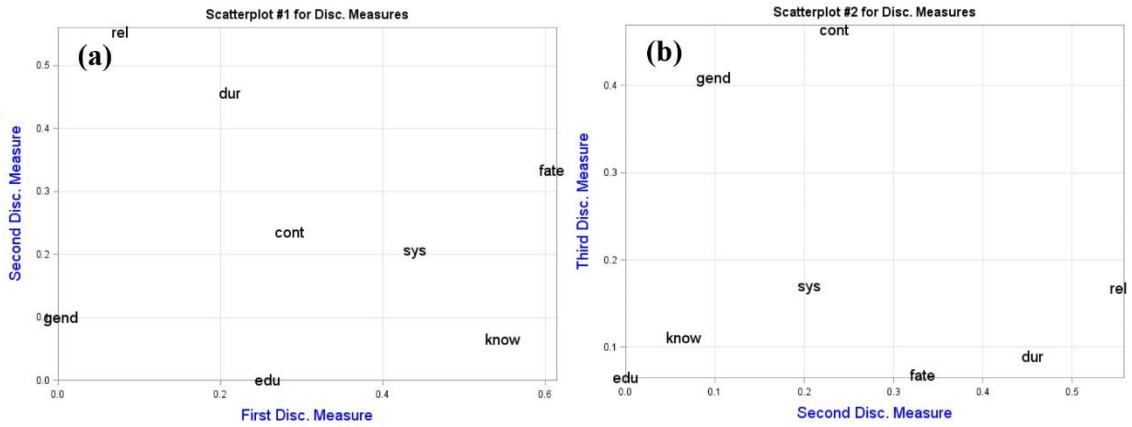


Table C.11. The cross-tabulation of *gender* with the response variable of *control* with the addition of the row percentages for the diabetes data[47].

	Low Control (cont1)	Ave. Control (cont2)	High Control (cont3)	Total
Male	13 (39%)	12 (36%)	8 (24%)	33
Female	10 (29%)	11 (31%)	14 (40%)	35
Total	23	23	22	68

To examine the other results more closely, we produce the homogeneity map of Figure C.14 using the following macro submission.

```
%classification(
    data=glucat,
    var=cont know fate sys rel dur edu gend,
    sub=cont know fate sys dur edu);
```

As can be seen in the scatterplots in Figures C.13(a)-(b), the discrimination measures for the second dimension (on the vertical or y-axis) are dominated by the *reliance* and *duration* variables. It can be helpful to visualize the projection of the labels for these variables onto the second (vertical) axis of Figure C.14. If the original quantitative variables were positively correlated, then we would expect the labels for the lowest to highest categories of the variables to be in same order along the second axis. If the original quantitative variables were negatively correlated, then we would expect the labels for the lowest to highest categories to be in reverse order along the second axis. From the bottom to the top of the second axis of Figure C.14, we find the first category, second category, and third category of the *reliance* variable, in that order. However, the order for the *duration* variable is the second, the first, and then the third category. This suggests a more complicated connection for the variables. In order to investigate this association, we produce Table C.12. This table suggests a quadratic trend where the degree of self-reliance is larger for those with shorter duration of illness, declines for those with average duration, and then begins to rise somewhat for those with longer duration of illness.

Figure C.14. Homogeneity map for the first two dimensions for the diabetes data[47].

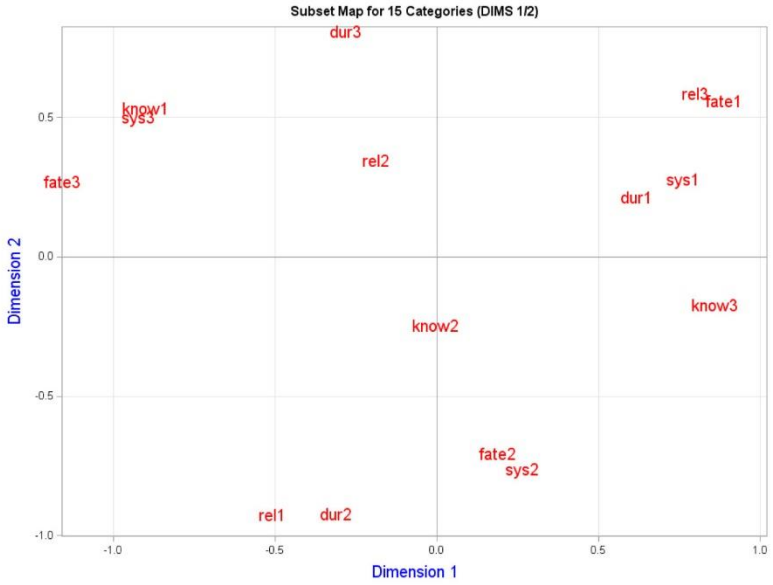


Table C.12. The cross-tabulation of the *reliance* and *duration* variables with the addition of row percentages for the diabetes data[47].

	Low Self-Reliance (rel1)	Ave. Self-Reliance (rel2)	High Self-Reliance (rel3)	Total
Short Duration (dur1)	6 (27%)	6 (27%)	10 (46%)	22
Ave. Duration (dur2)	13 (54%)	8 (33%)	3 (13%)	24
Long Duration (dur3)	3 (14%)	12 (55%)	7 (32%)	22
Total	22	26	20	68

Looking again at the scatterplot in Figure C.13(a), we see that the discrimination measures for the first dimension are dominated by the *fate*, *knowledge* and *system* variables. When we visualize the projection of these variables along the first axis of Figure C.14, we see that the *fate* and *system* variables follow the same order, which suggests a positive correlation for the original quantitative variables. However, the order of the *knowledge* variable is opposite to the other two variables, which suggests that the original quantitative variable is negatively correlated with the other two quantitative variables. Tables C.13(a)-(c) list the cross-tabulations for these variables.

Tables C.13(a)-(c). The cross-tabulations for variables which are loading on the first dimension for the homogeneity analysis for the diabetes data[47]. The approximate row percentages are added.

(a)	fate1	fate2	fate3	Total
sys1	10 (48%)	9 (43%)	2 (10%)	21
sys2	8 (35%)	11 (49%)	4 (17%)	23
sys3	4 (17%)	5 (21%)	15 (63%)	24
Total	22	25	21	68

(b)	know1	know2	know3	Total
fate1	3 (14%)	9 (41%)	10 (45%)	22
fate2	5 (20%)	10 (40%)	10 (40%)	25
fate3	12 (57%)	8 (38%)	1 (5%)	21
Total	20	27	21	68

(c)	know1	know2	know3	Total
sys1	4 (19%)	8 (38%)	9 (43%)	21
sys2	4 (17%)	10 (43%)	9 (39%)	23
sys3	12 (50%)	9 (38%)	3 (13%)	24
Total	20	27	21	68

(Technical note: Our results using various correspondence analyses are, for the most part, consistent with the conclusions of Cox and Wermuth[47], who examine partial correlations prior to modeling. The main difference for the models fitted with the response of measured glucose level is the inclusion in our model of some quadratic terms, such as the inclusion of the squared terms for *knowledge* and *duration*, as well as some additional interactions. These additions lead to an increase in the R^2 value from 34% to 41%. The adjusted R^2 values, which takes into account the number of parameters, are approximately 30% and 34%, respectively, for the two models. Therefore, a model that includes quadratic effects still represents a slight improvement, even when we consider the increase in the number of parameters.

The POSSE approach to the analysis and the approach taken by Cox and Wermuth are not mutually exclusive, and both can provide useful information for fitting a model. For this analysis, we also found it helpful to perform regression diagnostics on the models. For example, when the 'hat' values (see Neter et al.[59]) are calculated to indicate the 'leverage' or influence of the observations, the results for both our model and the model of Cox and Wermuth identify two observations with unusually small values of *x (knowledge)*. These two values are very influential and largely responsible for the significant p-values due to this explanatory variable. In other words, the modeling diagnostics for this sample suggest that the relationship of knowledge with diabetic control is less stable than, for instance, the relationship of education with diabetic control. This is a reminder that no one method (including the POSSE methods) is a panacea for data analysis, and that every stage of the analysis can provide an important contribution.)

C.10: Exploring Missing Values

(Note: Because of potential copyright issues, the data set and SAS code for this example are not available for download at the POSSE website.)

One of the advantages of a categorical analysis is that missing values can be simply represented as an additional category for a variable, which makes it relatively easy to investigate whether the pattern of missing values is random or not. We will demonstrate this with data from the Muscatine Coronary Risk Factor study[49-50] which surveyed over 4,800 school-age children in Muscatine, Iowa during the years 1977-1981. The data is described in Section B.3. The general purpose of the study was to assess the association of coronary disease with possible risk factors, including obesity. The study contains five birth cohorts, that is, five groups of children about the same age. Note that, although this study involves repeated measures, this fact is not used in the POSSE analysis here.

The investigation of birth cohorts is especially important when, for example, we have subjects in a longitudinal study who can be about the same age but born in different decades. For example, we cannot assume that a 75-year-old born in 1920 and a 75-year-old born in 1940 have had the same experience with respect to access to health care and health education. In the Muscatine study, the five birth cohorts are defined as those who in 1977 are 5–7 years, 7–9 years, 9–11 years, 11–13 years, and 13–15 years. Since the study is only five years long, we might expect any cohort effects to be small, but we can explore the possibility of the cohorts being associated with other effects during an informal preliminary analysis.

Each of the children in the study has assessments where he or she is classified as obese or non-obese at three time points: at the beginning of the study in 1977, at the middle in 1979, and at the end in 1981. There are also many missing values, so that the three possible outcomes for the *obesity* variable are 'Normal', 'Obese' and 'Missing'. In order to assess the cohort variable with respect to these outcomes, we explore whether the factors of cohort and age interact in their association with the obesity classification using the following submission of the ***correspondence*** macro. (Note that the age variable is defined here as the middle age for the subject's cohort. Also, in the following submission, the variable of *obesity*, which would normally be regarded as a response variable, is inserted here as the explanatory variable. This illustrates the fact that, during a preliminary exploratory analysis, strict distinctions between the various types of variables can be relaxed.)

```
data muscatine;
 set in.muscatine;
if (y=.) then y=2;
response = y + 1;
sex = gender + 1;
if (baseage=6)     then cohort=1;
else if (baseage=8)  then cohort=2;
else if (baseage=10) then cohort=3;
else if (baseage=12) then cohort=4;
            else cohort=5;
keep id response cohort sex age;

proc format cntlout=othrfmt1;
value response 1='Norm' 2='Obese' 3='Miss';
value sex 1='Male' 2='Female';
value age 6='Age6' 8='Age8' 10='Age10' 12='Age12' 14='Age14' 16='Age16' 18='Age18';
value cohort 1='Co1' 2='Co2' 3='Co3' 4='Co4' 5='Co5';
run;
```

```
%correspondence(
    dataset=muscatine,
    response=cohort,
    fmtresp=cohort,
    explanvars=response,
    covars=sex,
    fmtothr=othrfmt1,
    id=id,
    stratavar=age,
    fmtstrata=age,
    noplot=two);
```

The output from this submission is found in Figure C.15, which indicates that the first dimension captured over 95% of the variation. The 'Missing' category is contrasted with the other categories along this dimension. The following ***tabulation*** macro is used to derive the frequencies and percentages found in Table C.14.

```
%tabulation(
    dataset=muscatine,
    response=response,
    fmtresp=response,
    secondvar=age,
    thirdvar=cohort,
    id=id);
```

Table C.14 indicates that missing classifications are more likely to occur for the youngest children of the lowest birth cohort and the oldest children of the highest birth cohort. Although we might not interpret this result as a cohort effect in the sense given above, it demonstrates that missing values are more common at the beginning and end of this study. Knowing this could help in the planning of future studies.

Figure C.15. The stratified correspondence analysis of the Muscatine data[49-50].

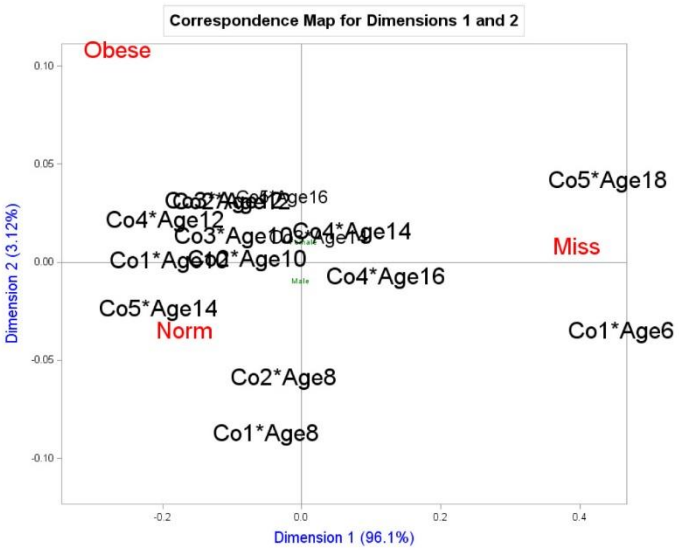


Table C.14. The percent of missing outcomes are shown for the ages and birth cohorts in the Muscatine study[49-50].

	Cohort 1			Cohort 2			Cohort 3			Cohort 4			Cohort 5		
Ave. Age At Testing	**6**	**8**	**10**	**8**	**10**	**12**	**10**	**12**	**14**	**12**	**14**	**16**	**14**	**16**	**18**
Sample Size	935	935	935	1014	1014	1014	1025	1025	1025	937	937	937	945	945	945
Percentage Missing	62%	28%	20%	30%	27%	26%	26%	25%	34%	20%	37%	40%	19%	31%	61%

C.11: Data Preparation for Multiple Quantitative Responses (repeated measures)

In order to make the results more consistent with the results found for other methods for repeated measurements, the derivation of categorical variables from multi-response quantitative variables is performed differently using the ***data_prep*** macro. Here we will demonstrate the approach by deriving the data set named 'fev1cat', which comes from the seafood study[48] which is described in Section B.3 of Appendix B. The original data set named 'fev1' has two quantitative response variables named *janfev* and *marchfev*, which respectively measure the FEV1, a widely-used measure of lung function, for 99 seafood workers at the two time points of January 1998 and March 1998. In this example, we will use the uniform ranking method in the ***data_prep*** macro to create three categories from each variable. The basic idea here is that, because we are specifically interested in the change from the first to the second response, we will want these variables to be ranked together or jointly. As seen in the code below, this is carried out by first splitting the data in two, where each half contains one of the two responses. This results in a dataset which then has twice as many observations. We then create a new variable named *score* which represents all the responses for the two variables. The macro then applies the ranking procedure to the variable *score*, and then the two halves of the data are recombined by matching on the ID variable, in order to create the two categorical variables named *firstfev* and *secondfev*, which are then saved in the SAS data set named 'fev1cat.'

```
data splitdata;
 set in.fev1;  ** must be sorted by ID variable **;
survey=1; score = janfev; output;
survey=2; score = marfev; output;
keep id score; run;

%data_prep(
    rawdata=splitdata,
    contvar=score,
    id=id,
    catname=catscore,
    firstlevel=1,
    ranklevs=3,
    outdata=pooled,
    savevars=catscore);

proc transpose data=pooled out=transpool prefix=survey;
by id;
var catscore;

data fev1cat;
 set transpool (rename=(survey1=firstfev survey2=secondfev));
drop _name_;
run;
```

If there had been responses at three time points with another FEV1 measurement in April 1998, the data set would then have been split into three parts by adding another 'output' line with 'survey=3' and 'score=aprilfev' in the initial data step which precedes the macro submission. In addition, another component 'survey3=thirdfev' would then have been added when renaming the variables in the final data step.

The new categorical variables *firstfev* and *secondfev* can then be used in a POSSE analysis. The following macro submission creates Table C.15 using the new data set. The diagonal frequencies in the shaded cells represent the paired responses that agree for the two surveys. In general, there appears to be little change in the FEV1 measurement from the first survey to the second survey.

```
proc format;
value firstfev 1='LoFEV1' 2='MidFEV1' 3='HiFEV1';
value secondfev 1='LoFEV1' 2='MidFEV1' 3='HiFEV1';
run;

%tabulation(
   dataset=fev1cat,
   response=firstfev,
   fmtresp=firstfev,
   secondvar=secondfev,
   id=id,
   perc=no,
   rowresp=yes);
```

Some readers might wonder why we don't base the analysis on the ranking of the differences between the responses at the two time points. However, that approach would not have told us whether a change of +1 (i.e., an increase of one category from the first to the second time point) is associated with a change from the first to the second category, or from the second to the third category, or both.

Tables C.15. The cross-tabulation of the first FEV1 in January 1998 versus the second FEV1 in March 1998 for the seafood study[48].

		March Testing			
		LoFEV1	**MidFEV1**	**HiFEV1**	**Total**
January Testing	**LoFEV1**	29	5		34
	MidFEV1	3	25	4	32
	HiFEV1		5	28	33
	Total	32	35	32	99

C.12: Exploring the Bias in Square Tables (repeated measures)

One approach to repeated measurements would be to include a time variable in a simple correspondence analysis, which is the approach used in the analysis of time series data in Chapter 9 of the Greenacre[25] text. A more complex approach to matched pairs of outcomes is given in Chapter 22 of the later Greenacre[28] text. The method there provides for the analysis of square tables, where the rows and column categories represent the same outcomes at two different time points. The approach introduced and used here will also be applied to square tables for matched pairs of responses.

In this section, we revisit the example using the ILO ordinal classification system, which was presented

in Section 5.3 and Figure 5. The presentation in Section 5.3 involves one set of classifications for over 200 x-ray images by nine doctors. Of these images, 129 images were later classified a second time by the same nine doctors. The data described in Section B.6 contains matched pairs of classifications for each image and doctor combination for these 129 x-ray images. One way to present the results of such repeated classifications is in a square table like Table C.16(a). The frequencies along the main diagonal of the table represent those paired classifications that are identical, or classified into the same category, at the two time points. In assessing any systematic changes for the two classifications, the focus will be on the off-diagonals. Counts above the diagonal represent pairs where the second classification is higher than the first, while counts below the diagonal represent pairs where the second classification is lower than the first. If there are no systematic changes, we would expect the pattern of frequencies above the main diagonal to roughly mirror the pattern below.

Tables C.16(a)-(b). (a) The cross-tabulation of the first classifications (i.e., the row variable) versus the second classifications for nine doctors and 129 x-ray images (unpublished data), and (b) the frequencies for the trans-positioned cells above ('+') versus below ('-') the main diagonal for those matched cells with the largest contributions to the inertia.

(a)		Second Classification										
		(1) 0/0	(2) 0/1	(3) 1/0	(4) 1/1	(5) 1/2	(6) 2/1	(7) 2/2	(8) 2/3	(9) 3/2	(10) 3/3	Total
First Classification	(1) 0/0	287	39	28	18	2	4	1		1		380
	(2) 0/1	42	35	42	11					1		131
	(3) 1/0	28	21	73	35	4	1	1	1			164
	(4) 1/1	2	4	28	65	23	8	7	1			138
	(5) 1/2	2		9	22	17	10	8	2			70
	(6) 2/1	2		1	11	17	10	14	4	3	1	63
	(7) 2/2			1	4	12	18	25	14	3	6	83
	(8) 2/3			2		1	12	19	17	8	2	61
	(9) 3/2						3	10	9	13	3	38
	(10) 3/3						2	2	4	9	16	33
	Total	363	99	184	166	76	68	87	52	38	28	1161

(b)	1*4	1*7	1*9	2*3	2*4	2*9	3*5	4*8	5*6	6*8	7*9	7*10	9*10
+	18	1	1	42	11	1	4	1	10	4	3	6	3
-	2	0	0	21	4	0	9	0	17	12	10	2	9
Total	20	1	1	63	15	1	13	1	27	16	13	8	12

A formal test of symmetry about the diagonal of a 2-by-2 table is given by McNemar[60] and, more generally, for larger square tables by Bowker[61]. During the calculations, these tests of significance use the differences in the frequencies for the corresponding ij^{th} and ji^{th} cells above and below the diagonals (i.e., where $i \neq j$). The exploratory approach presented here is similar, insofar as it also matches the frequencies in the ij^{th} and ji^{th} cells about the diagonals before applying a correspondence analysis. As is also true for the formal tests of symmetry, the frequencies along the main diagonal are not used during the correspondence analysis.

The code which prepares the data and performs the correspondence analysis can be seen just below. The variables named *ratingone* and *ratingtwo* represent the first and second classifications. We are simply investigating whether the classifications changed over time and in what direction, so during the initial data step the data are reduced to those observations where the ratings changed from the first classification

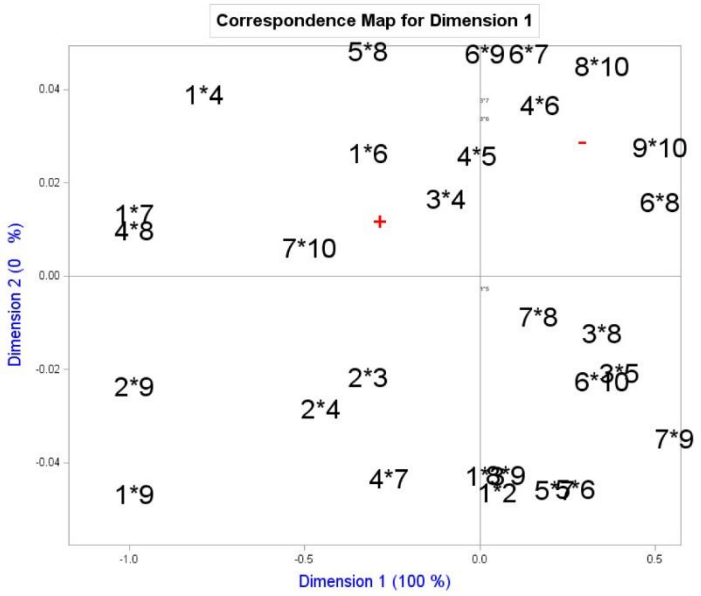
to the second. In addition, a new variable named *matchedcells* represents the results for the matched cells found above and below the diagonal, and a binary variable named *direction* indicates whether the first or second classification is larger. The label for the first level of *direction* was set to '+' to indicate that the second or follow-up classification is larger than the initial classification, and set to '-' to indicate the reverse. Also, a new ID variable named *iid* is created for the matched observations. Because of the many possible combinations for the first and second ratings, a user might decide to not assign a format for the response variable, but we have included one here in order to make the results easier to interpret.

```
data subset_ratings;
 set allratings;
if (ratingone=ratingtwo) then delete;
if (ratingtwo > ratingone) then do;
 matchedcells=100*ratingone + ratingtwo; direction=1;
end;
else do;
 matchedcells=100*ratingtwo + ratingone; direction=2;
end;
run;

data subset_ratings;
 set subset_ratings;
iid = _n_; ** ID variable for the matched cells **;
run;

proc format cntlout=othrfmt1;
value ratingone 1='0/0' 2='0/1' 3='1/0' 4='1/1' 5='1/2' 6='2/1' 7='2/2' 8='2/3' 9='3/2' 10='3/3';
value ratingtwo 1='0/0' 2='0/1' 3='1/0' 4='1/1' 5='1/2' 6='2/1' 7='2/2' 8='2/3' 9='3/2' 10='3/3';
value matchedcells 102='1*2' 103='1*3' 104='1*4' 105='1*5' 106='1*6' 107='1*7' 109='1*9' 203='2*3'
204='2*4' 209='2*9' 304='3*4' 305='3*5' 306='3*6' 307='3*7' 308='3*8' 405='4*5'
406='4*6' 407='4*7' 408='4*8' 506='5*6' 507='5*7' 508='5*8' 607='6*7' 608='6*8' 609='6*9'
610='6*10' 708='7*8' 709='7*9' 710='7*10' 809='8*9' 810='8*10' 910='9*10';
value direction 1='+' 2='-';
run;

%correspondence(
    dataset=subset_ratings,
    response=matchedcells,
    fmtresp=matchedcells,
    explanvars=direction,
    fmtothr=othrfmt1,
    id=iid,
   onedim=yes);
```

Since the variable *direction* has only two categories and is the only predictor variable, the correspondence analysis is one-dimensional, but the labels in Figure C.16 have been spread out randomly over the second dimension to avoid over-plotting. Note that the matched cells are represented in the figure by the paired-numbers labels, where the two numbers represent the category numbers which are involved in the change of classification. For example, the label '2*4' indicates the matched cells involving changes in ratings between the second and fourth categories of the ILO classification system in either direction. Therefore, it represents the combined frequency of changes from either the second to the fourth category, and from the fourth to the second category, for the corresponding cells below and above the diagonal in Table C.16. These two cells are then set apart by the *direction* variable, which indicates whether the larger classification occurred during the first classification or during the second classification. The left side of the first-dimensional axis in Figure C.16 is identified with the first category of the *direction*

variable, labeled as '+', and the right side of the axis is identified with the second category, labeled as '-'. Therefore, the numbered labels found to the left of the origin represent images where the second classifications tend to be larger than the initial classifications, whereas the labels to the right represent images where the second classifications tend to be smaller. For example, the label '1*4', found to the left and near the bottom of the figure, represents a tendency to classify some images into the lowest or first category '0/0' during the initial or first-round assessment, but to classify the same images into the fourth '1/1' category during the second-round assessment. On the other hand, the label '7*9', found on the right side of the figure, represents a tendency to classify some images into the ninth '3/2' category during the first round, but then into the seventh '2/1' category during the second round.

Figure C.16. The correspondence map for the one-dimensional correspondence analysis for the repeated classifications of 129 x-ray images by nine doctors for the matched categories (unpublished data). The first digit of the label represents the rating category for the first classification, and the rest of the label represents the rating category for the second classification.



Matched Categories with the Largest Contributions to the 1st Dimension Inertia												
1*4	**1*7**	**1*9**	**2*3**	**2*4**	**2*9**	**3*5**	**4*8**	**5*6**	**6*8**	**7*9**	**7*10**	**9*10**
0.246	0.019	0.019	0.128	0.061	0.019	0.041	0.019	0.041	0.085	0.079	0.038	0.064

To conserve space, the table below Figure C.16 contains the contributions to the inertia for those matched cells where frequencies are at least one and the contribution is at least 0.019. Table C.16(b) contains the frequencies for this subset of matched cells. The largest contribution to the inertia of 0.246 found below Figure C.16 is associated with the matched cells involving changes between the first '0/0' category and the fourth '1/1' category. Table C.16(b) indicates that there are 18 images whose initial classification is '0/0', but with a follow-up classification of '1/1', but only two images with the reverse of this. The second largest contribution of 0.128 involved changes from the second '0/1' category to the third '1/0' category, and Table C.16(b) indicates that there are 42 images whose initial classification is '0/1' with a follow-up classification of '1/0', but only 21 images with the reverse of this. Both of these results indicate tendencies towards higher second-round classifications for images which were initially classified into a lower category. On the other hand, there are also some weaker tendencies for the second classification

to be lower for some images that were initially classified into some higher categories.

Most studies involving repeated measurements generally have more interest in questions about whether the changes in the responses are related to other factors. For example, we might wonder if the results differ substantially for the nine different raters. One way to investigate this would use the same approach as above to prepare the data, followed by a somewhat more complicated correspondence analysis. Fortunately, the same results are easily obtained with the POSSE methods by using the following submission. As can be seen below, the preparation within the data step for this approach simply involves removing the observations where the matched pairs of responses are identical. We can also use the original ID variable in this macro submission.

```
data subset_ratings;
 set allratings;
if (ratingone=ratingtwo) then delete;

proc format cntlout=othrfmt1;
value ratingone 1='1' 2='2' 3='3' 4='4' 5='5' 6='6' 7='7' 8='8' 9='9' 10='10';
value ratingtwo 1='1' 2='2' 3='3' 4='4' 5='5' 6='6' 7='7' 8='8' 9='9' 10='10';
value rater 1='Rd1' 2='Rd2' 3='Rd3' 4='Rd4' 5='Rd5' 6='Rd6' 7='Rd7' 8='Rd8' 9='Rd9';
run;

%correspondence(
   dataset=subset_ratings,
   response=ratingone,
   fmtresp=ratingone,
   explanvars=rater,
   fmtothr=othrfmt1,
   id=id,
   stratavar=ratingtwo,
   fmtstrata=ratingtwo);
```

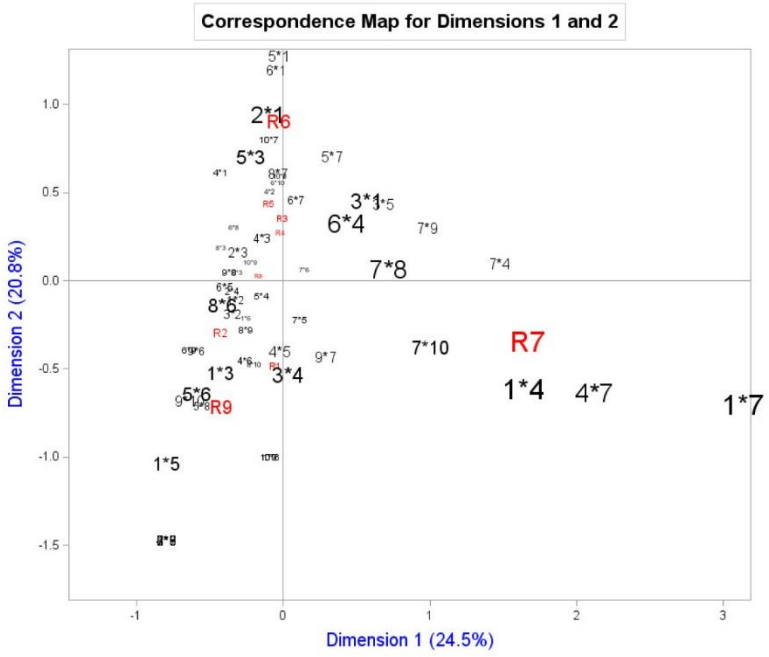
Figures C.17 is the resulting correspondence map, where the label '1*4' represents the images which were classified into the first category during the first round and into the fourth category during the second round. Note that, unlike in Figure C.16, the *direction* variable is not necessary here, because the reverse of every label is also found in the map. For example, the label '4*1', which represents the images classified into the fourth category during the first round and the first category during the second round, can be found in the upper-left quadrant. Figure C.17 suggests that the '1*4' combination of responses is especially associated with the seventh rater, who is represented by the 'R7' label in the maps. The label '2*1' represents images which were classified into the second category during the first round and into the first category during the second round. This set of responses is associated with the sixth rater, who is represented by the 'R6' label.

These results are further interpreted by stratifying Table C.16(a) by the rater variable using the following macro submission.

```
proc format;
value ratingone 1='0/0' 2='0/1' 3='1/0' 4='1/1' 5='1/2' 6='2/1' 7='2/2' 8='2/3' 9='3/2' 10='3/3';
value ratingtwo 1='0/0' 2='0/1' 3='1/0' 4='1/1' 5='1/2' 6='2/1' 7='2/2' 8='2/3' 9='3/2' 10='3/3';
value rater 1='R1' 2='R2' 3='R3' 4='R4' 5='R5' 6='R6' 7='R7' 8='R8' 9='R9';
run;

%tabulation(
   dataset=allratings,
   response=ratingone,
```

```
 fmtresp=ratingone,
 secondvar=ratingtwo,
 thirdvar=rater,
 id=id,
perc=no,
rowresp=yes,
range1=1 to 10,
range2=1 to 10,
range3=1 to 9);
```

Figure C.17. The map for the correspondence analysis which examined the factor of *rater* for the repeated x-ray classifications (unpublished data). Unlike Figure C.16, there is no *direction* variable, but the direction of the change in classification is signified by the order of the paired categories.



A portion of the results produced by the macro is found in Table C.17, which contains only the results for the sixth and seventh raters. The upper portion of Table C.17 reveals that there are 16 images classified by the sixth rater with an initial classification into the second category of '0/1' and a follow-up classification into the lowest category of '0/0', but there is only one image with the reverse of this. The lower portion of Table C.17 reveals that the seventh rater initially classified 10 images into the lowest category, but then subsequently classified them into the fourth '1/1' category, but there are no images with the reverse of this. Clearly, follow-up classification results differ substantially for these two raters, when compared to other raters.

Table C.17. Portion of table created using the ***tabulation*** macro containing the results for Raters #6 and #7. A zero (in red) has been inserted in the results for Rater #7 to indicate the transpositioned cell corresponding to the large off-diagonal frequency of 10.

		0/0	0/1	1/0	1/1	1/2	2/1	2/2	2/3	3/2	3/3
R6	0/0	50	1	1							
	0/1	16	1	6							
	1/0	5	1	7		1					
	1/1	1		6	1						
	1/2	1		3	2			1			
	2/1	1			2	1	1	1			
	2/2					1	3	3	2		
	2/3			1			1	3	2		
	3/2							1		1	
	3/3										1
R7	0/0	57			10			1			
	0/1										
	1/0	6			4	1					
	1/1	0		1	11	2		5			
	1/2				1			1			
	2/1				2			1			
	2/2				2	2	2	6	4	1	2
	2/3							1	1		
	3/2							2		1	
	3/3										2

C.13: Examining the Changes in Cluster Profiles (repeated measures)

This section applies a cluster analysis to questionnaire data from 107 workers from the seafood-processing study[48] described in Section B.3. The purpose of the study is to determine the relationship between work exposures and the onset of asthma and other respiratory conditions. Unfortunately, there is no universally-accepted case definition for asthma[62]. The application of the POSSE methods here might be regarded as an attempt to formulate 'empirical' case definitions (i.e., based only on the data). In practice, such an approach can benefit from the input of subject-matter experts.

Occupational health surveys often have many variables. This particular survey also included pulmonary function testing and bloodwork. However, the analysis here only involves a subset of questions which were asked of the workers during both the initial questionnaire in January 1998 and the follow-up questionnaire in March of the same year, focusing on the changes in responses that occurred from the initial survey to the follow-up survey. The names and brief descriptions for the questionnaire variables are found in Section B.3 and reproduced in Table C.18. The goal of the analysis will be to use the variables found in Table C.18 to formulate symptom clusters, and then use the resulting symptom clusters in a correspondence analysis.

Table C.18. Variable names and brief descriptions for the subset of 20 symptom variables which were collected during both the initial and follow-up surveys of the seafood study[48]. The 15 variables in bold are retained after the first stage of the homogeneity analysis and the 12 checked variables are retained after the second stage.

Symptom Variable Names and Descriptions	
√ - ***wheeze*** : chest wheezing	*coldair* : symptoms with cold air
√ - ***wheezecond*** : wheeze besides than during a cold	*workdust* : symptoms with work dust
clearcgh : wheeze clears after a cough	√ - ***wokecgh*** : woken by cough
√ - ***normbetw*** : normal breathing between wheezing	***cghmorn*** : cough in the morning
√ - ***shortbreath*** : shortness of breath	√ - ***cghlater*** : cough later in the day
√ - ***chesttight*** : chest tightness	√ - ***cghoften*** : cough most days or nights
√ - ***freqsympts*** : frequent chest symptoms	***phlgmorn*** : phlegm in the morning
√ - ***heavyexer*** : symptoms with heavy exercise	√ - ***phlglater*** : phlegm later in the day
exerbreath :	√ - ***stuffnose*** : stuffy nose or drainage
afterexer : wheezing after exercise	***blocknose*** : blocked or runny nose

Section C.11 describes how, when we wish to derive the response variables from repeated quantitative ones, we will want to jointly rank the multiple responses. The same underlying principle is followed for this example, where we first perform a joint cluster analysis over the two time points of the study, and then afterwards merge the clusters for the time points by the ID variable before analyzing the changes over time. The process we use in this analysis to find clusters is described in Section 6.2. Prior to the cluster analysis, the first and second survey data are pooled using the following code. Note that shortened variable names are also assigned prior to the macro submission.

```
data symptoms;
 set in.symptoms; ** data for both surveys **;
iid = _n_;
run;

data sympt;
 set symptoms;
array x[20] wheeze wheezecond clearcgh shortbreath normbetw afterexer chesttight freqsympts
     heavyexer coldair workdust exerbreath wokecgh cghmorn cghlater cghoften
     phlgmorn phlglater stuffnose blocknose;
array y[20] wh whc clcgh shbr norb exer chtgh frsymp hexer cair wdust
     exerb woke cghm cghl cgho phm phl stnos blnos;
do i=1 to 20;
 y[i]= x[i];
end;
keep iid wh whc clcgh shbr norb exer chtgh frsymp hexer cair wdust
    exerb woke cghm cghl cgho phm phl stnos blnos;
run;
```

Following the process described in Section 6.2, the first step in the clustering process is to examine the scree plots (Section 6.1) in order to identify the variables with the largest discrimination measures.

```
%classification(
     data=sympt,
     var=wh whc clcgh shbr norb exer chtgh frsymp hexer cair
        wdust exerb woke cghm cghl cgho phm phl stnos blnos);
```

Figures C.18(a)-(c) are the three scree plots that are produced from this submission. Dotted lines have been inserted to indicate the number of variables with larger discrimination measures which are chosen from each of the three dimensions. After inserting 'print=yes' along with 'ndim1=9', 'ndim2=6' and 'ndim3=4' and resubmitting the macro, a list of 15 unique variables is printed. Therefore, during this first step, the number of variables is reduced from 20 to the 15 variables with relatively larger inertias. These 15 variables are seen in bold in Table C.18.

Figure C.18(a)-(c). The scree plots for the three dimensions for the seafood worker study[48]. The horizontal dotted lines have been inserted to indicate the number of variables which are retained during the first stage of the cluster analysis.

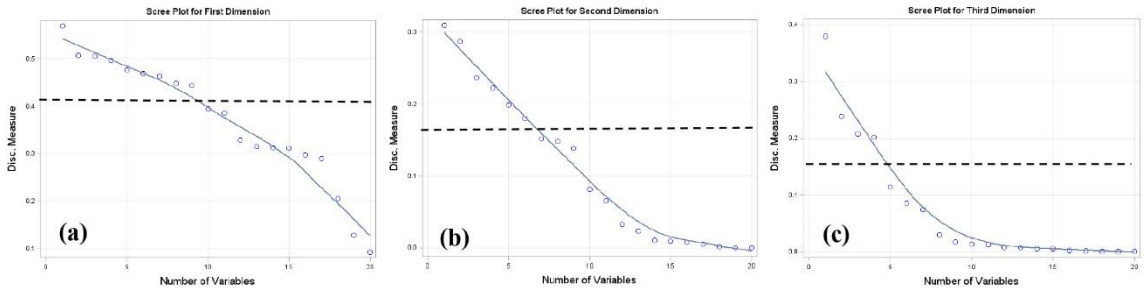
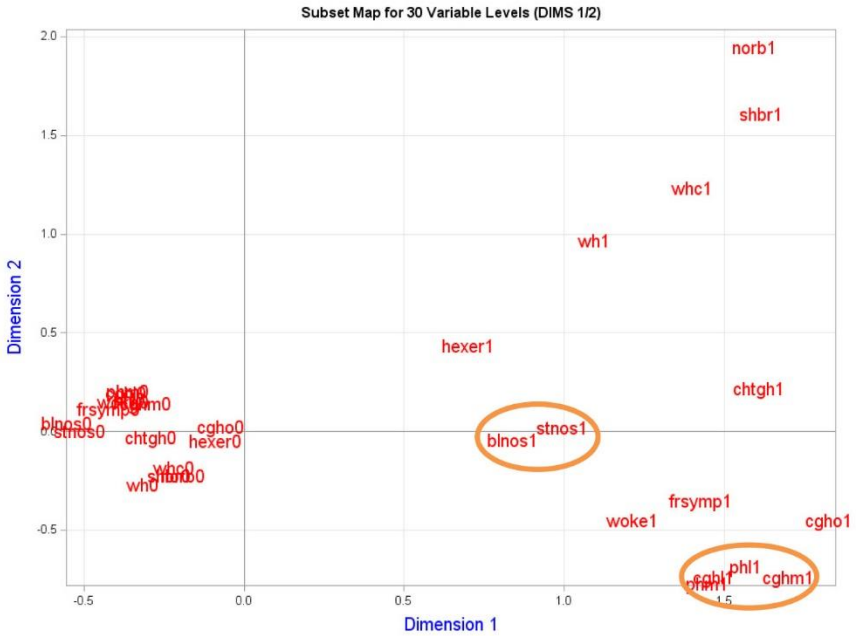


Figure C.19. The map is shown which is produced during the second stage of the homogeneity analysis of the seafood study[48]. The 'blnos1' and 'stnos1' labels along the first axis represent the positive responses for the *blocknose* and *stuffnose* variables, and their close proximity indicates that they tend to co-occur. Because the discrimination measures for the *stuffnose* variable were larger, it is chosen from the two variables for the subsequent cluster analysis. A similar approach is followed for the other cluster of variable categories which is circled in the lower portion of the plot.

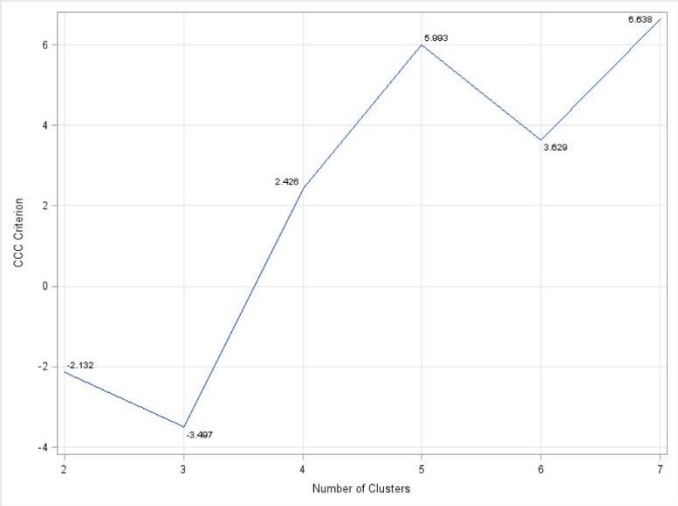


During the second step, the remaining 15 variables are then included in a subset homogeneity analysis in order to identify any redundancies among the variables. This part of the analysis is performed by inserting the shortened names for the 15 variables in the 'sub=' entry, as seen below. The output includes Figure C.19. The clusters of labels which are circled in Figure C.19 identify those variable categories

which were somewhat correlated. After examining the discrimination measures in the output (not shown here) for the correlated categories, we then reduced the number of variables from 15 to 12 variables. These remaining 12 variables are checked in Table C.18. (Note that another possible choice in the case of redundant variables is to find a cluster variable to represent them, as demonstrated in Section C.7. It might also be instructive to see how different choices impact the clustering which follows.)

```
%classification(
   data=sympt,
   var=wh whc clcgh shbr norb exer chtgh frsymp hexer cair wdust
      exerb woke cghm cghl cgho phm phl stnos blnos,
   sub=blnos cghl cghm cgho chtgh frsymp hexer norb phm phl shbr stnos wh whc woke,
   noprint=yes);
```

Figure C.20. The plot of the CCC statistics for the series of clusters using the symptom variables from the seafood study[48].



As seen in the next submission below, the shortened names for the remaining 12 variables are then inserted into the 'haclust=' entry of the macro, while the 'fitclust=yes' option is used to calculate the CCC statistic for various cluster sizes. Figure C.20 plots the CCC statistics against the number of clusters, where a local maximum or peak value of 5.99 for the CCC statistics is associated with a choice of five clusters.

```
%classification(
   data=sympt,
   haclust=cghl cgho chtgh frsymp hexer norb phl shbr stnos wh whc woke,
   fitclust=yes,
   id=iid);
```

Finally, the macro submission below saves the clustering information for a choice of five clusters into the dataset named 'tempclust', and the additional formatting below the macro creates the row labels before printing Table C.19.

```
%classification(
   data=sympt,
   haclust=cghl cgho chtgh frsymp hexer norb phl shbr stnos wh whc woke,
   nclust=5,
```

```
    id=iid,
    allbin=yes,
    out=tempclust);

proc format;
value $var_ 'cghl1'='Cough Later in Day' 'cgho1'='Cough Often' 'chtgh1'='Chest Tight'
    'frsymp1'='Frequent Chest Symptoms' 'hexer1'='Symptoms with Heavy Exercise'
    'norb1'='Normal Between Wheezing' 'phl1'='Phlegm Later in Day'
    'shbr1'='Shortness of Breath' 'stnos1'='Stuffed Nose or Drainage' 'wh1'='Chest Wheezing
    'whc1'='Wheeze Without Cold' 'woke1'='Woken by Cough';

proc print data=_proportions_ split='_' double;
id condition;
format _numeric_ 5.2 condition $var_.;
label condition='Condition';
run;
```

Table C.19. The five cluster profiles are shown with the column labels which are assigned prior to the correspondence analysis. (Note that the clusters are re-ordered before applying the subsequent correspondence analysis.)

Condition	'Cold' (n=26)	'Norm' (n=138)	'Other' (n=4)	'Asth' (n=21)	'HvyEx' (n=25)
Cough Later in Day	0.88	0.03	0.50	0.52	0.12
Cough Often	0.00	0.00	1.00	0.10	0.08
Chest Tight	0.46	0.02	0.50	0.71	0.04
Frequent Chest Symptoms	0.85	0.03	0.75	0.67	0.24
Symptoms with Heavy Exercise	0.04	0.00	1.00	0.19	0.60
Normal Between Wheezing	0.00	0.01	0.00	0.95	0.04
Phlegm Later in Day	0.77	0.03	0.75	0.52	0.08
Shortness of Breath	0.08	0.01	0.00	1.00	0.08
Stuffed Nose or Drainage	0.85	0.15	1.00	0.86	0.32
Chest Wheezing	0.27	0.04	0.75	0.86	0.56
Wheeze Without Cold	0.15	0.00	0.75	0.67	0.32
Woken by Cough	0.77	0.07	1.00	0.52	0.28

Note that, because the data from the 107 subjects are pooled for the two rounds, the total sample size for the five clusters in Table C.19 is 214, or twice the number of subjects. In order to characterize the clusters, labels are added in Table C.19. For example, the fourth cluster, labeled as 'Asth', is identified with asthma-like symptoms, whereas the third cluster, labeled as 'Cold', is identified with those with cold-like symptoms. It should also be noted that, prior to the saving the cluster information in the data set named 'symptclusters', the cluster numbers have been reordered as follows.

```
1='Norm'   2='HvyExer'   3='ColdSympt'   4='AsthLike'   5='Other'
```

Once we have a cluster determination for each subject for both first and second survey responses, we can then explore the changes in the cluster classifications over time. The following code is used to combine the cluster determinations for each subject.

```
data one;
 set in.symptclusters;

data survey1;
 set one;
if (survey=1);
clustone = symptcluster;
keep id clustone agecat exposure;

data survey2;
 set one;
if (survey=2);
clusttwo = symptcluster;
keep id clusttwo;

data all;
 merge survey1 survey2;
by id;
run;
```

The ***tabulation*** macro is then applied to the merged data set to produce Table C.20(a), which displays the overall results, and Table C.20(b), which displays the matched cells with relatively large differences in frequencies.

Table C.20(a)-(b). (a) The cross-tabulation of the cluster assignments for the initial survey versus the cluster assignments for the follow-up survey for the 107 workers in the seafood study[48]. (b) The frequencies for matched cells above and below the diagonal of the cross-tabulation. A zero (in red) has been inserted in the first table to indicate the transpositioned cell corresponding to the large off-diagonal frequency of 14. (Note that the clusters in Table 19 have been reordered here and in the other results that follow.)

(a)		Clusters for the Follow-Up Survey					
		Norm	HvyEx	Cold	Asth	Other	Total
Clusters for the Initial Survey	Norm	51	3	14	9		77
	HvyEx	8	3	5	3		19
	Cold	0		2			2
	Asth	2		2	1	1	6
	Other			1	2		3
	Total	61	6	24	15	1	107

(b)	Norm~HvyEx	Norm~Cold	Norm~Asth	HvyEx~Cold	Total
+ (above diagonal)	3	14	9	5	31
– (below diagonal)	8		2		10
Total	11	14	11	5	41

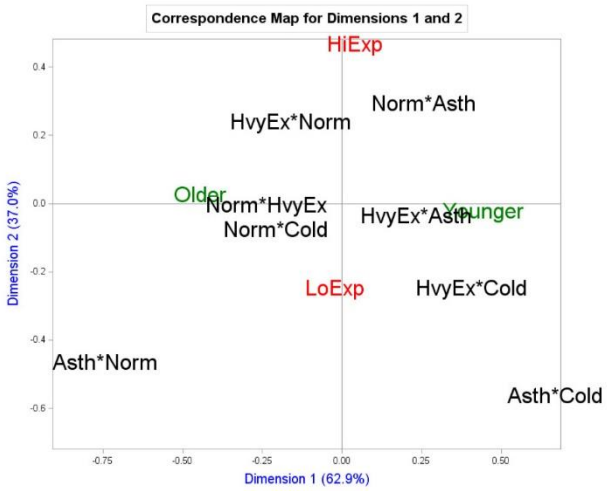
We next examine whether these changes might be related to two other factors. Of the 107 workers, 28 workers or about one-quarter of the workers, are identified as working in areas with higher levels of organic matter. To assess the effect of age, a binary age variable is derived using the median age of 35 years old. We then perform a correspondence analysis with the explanatory binary variables of exposure and age. Also, because of the small number of workers in the fifth cluster, this cluster is removed in the data step prior to the correspondence analysis.

```
data suball;
 set all;
if (clustone=clusttwo) then delete;
if (clustone=5 or clusttwo=5) then delete; ** removes fifth cluster from analysis **;


proc format cntlout=othrfmt1;
value clustone 1='Norm' 2='HvyEx' 3='Cold' 4='Asth' 5='Other';
value clusttwo 1='Norm' 2='HvyEx' 3='Cold' 4='Asth' 5='Other';
value agecat 1='Younger' 2='Older';
value exposure 1='LoExp' 2='HiExp';
run;

%correspondence(
   dataset=suball,
   response=clustone,
   fmtresp=clustone,
   explanvars=exposure,
   covars=agecat,
   fmtothr=othrfmt1,
   id=id,
   stratavar=clusttwo,
   fmtstrata=clusttwo);
```

Results from the correspondence analysis are seen in Figure C.21. The first dimension primarily contrasts the results for the older versus the younger workers, whereas the second dimension contrasts the high versus low exposure areas. The largest contributions to the first dimension are associated with two older workers in low exposure areas who reported asthma-like symptoms in January which apparently were resolved by March. The highest contribution to the second dimension is associated with changes from a normal condition to having asthma-like symptoms. Although Tables C.21 and C.22 indicate that small frequencies are involved, they provide some interpretation for the results. For example, in Table C.22, 5/20 or 25% of those higher-exposed workers who were normal in January developed asthma-like symptoms by March, but this was true for only 4/57 or about 7% of the lower-exposed workers.

(Technical Note: For a choice of five clusters, the clustering results above are similar to those using the advanced method of latent transition analysis[55]. Results from the latent transition analysis (Table C.23) indicate a pattern of frequencies which is similar to those found in Table C.20(a).)

Figure C.21. The correspondence map for the changes in cluster classifications with respect to the factors of age and exposure for the analysis of the seafood study[48].



Contributions of Outcomes to Inertia

Variable Levels	Contrib. to 1st Dim.	Contrib. to 2nd Dim.
Asth*Cold	0.196	0.237
Asth*Norm	0.241	0.162
HvyEx*Asth	0.036	0.002
HvyEx*Cold	0.182	0.115
HvyEx*Norm	0.045	0.168
Norm*Asth	0.130	0.285
Norm*Cold	0.133	0.032
Norm*HvyEx	0.037	0.000

Contributions of Other Variables to Inertia

Variable Levels	Contrib. to 1st Dim.	Contrib. to 2nd Dim.
LoExp	0.002	0.346
HiExp	0.003	0.649
Younger	0.498	0.002
Older	0.498	0.002

Table C.21. Cross-tabulations of the initial versus the second cluster assignments, stratified by the two age categories.

<35 Years Old		Norm	HvyEx	Cold	Asth	Other	Total
Younger (< 35 years)	Norm	20	1	5	6		32
	HvyEx	3	2	4	2		11
	Cold			2			2
	Asth			2	1	1	4
	Other				2		2
Total		23	3	13	11	1	51
>35 Years Old		Norm	HvyEx	Cold	Asth	Other	Total
Older (≥ 35 years)	Norm	31	2	9	3		45
	HvyEx	5	1	1	1		8
	Cold						
	Asth	2					2
	Other			1			1
Total		38	3	11	4		56

Table C.22. Cross-tabulations of the initial versus the second cluster assignments for workers in low-exposure and high-exposure areas.

Low Exposure		Norm	HvyEx	Cold	Asth	Other	Total
Lower Exposure Areas	Norm	41	2	10	4		57
	HvyEx	4	3	4	2		13
	Cold			1			1
	Asth	2		2	1		5
	Other			1	2		3
Total		47	5	18	9		79
High Exposure		**Norm**	**HvyEx**	**Cold**	**Asth**	**Other**	**Total**
Higher Exposure Areas	Norm	10	1	4	5		20
	HvyEx	4		1	1		6
	Cold			1			1
	Asth					1	1
	Other						
Total		14	1	6	6	1	28

Table C.23. The cross-tabulation of the cluster assignments using the advanced method of latent transition analysis for the 107 workers in the seafood study[48]. The results are generally similar to those found in Table C.20(a). (Note that the latent transition clusters numbers have been re-ordered to correspond to the cluster numbers found using the POSSE methods.)

		Clusters for the Follow-Up Survey					
		Norm	HvyEx	Cold	Asth	Other	Total
Clusters for the Initial Survey	Norm	52	5	16	9		82
	HvyEx	6	4	6	4		20
	Cold			1			1
	Asth					1	1
	Other			1	2		3
	Total	58	9	24	15	1	107

VII. APPENDIX D: Review of Chi-Square Statistic

The chi-square distribution (sometimes written as χ^2) is one of the most important in applied statistics. For example, it supplies the theoretical basis for the analysis of variance. It is also essential for many of the techniques for analyzing count or frequency data. As Mirkin[63] has shown, the chi-square statistic has a variety of uses and interpretations. The one that we focus on here is called the chi-square test of independence.

Table D.1. An example of independent or non-associated row and column variables (χ^2=0.62, p=0.999), where the row variable with categories 'r1-r4' is cross-tabulated with the column variable with categories 'c1-c4'.

	Column Variable				Totals
	c1	c2	c3	c4	
Row Variable					
r1	5	5	4	4	18
r2	2	3	2	3	10
r3	4	3	3	3	13
r4	7	6	5	6	24
Totals	18	17	14	16	65

To present a more intuitive understanding for the test of independence, we first examine Table D.1, which provides an example of independence or non-association between the row and column variables. The row and column totals give us the 'marginal' results, whereas the counts within the interior cells of the table give us the 'conditional' results. If the row and column variables are independent, we should be able to provide good predictions of the conditional results (in the interior of the table) by using the marginal results. Under the assumption of independence, these predictions or expected values are estimated using the following equation:

$$RowProportion \times ColumnProportion \times TotalCount = ExpectedValue$$

For example, for the second cell in the first row, the expected value would be as follows:

$$\frac{17}{65} \times \frac{18}{65} \times 65 \approx 4.71$$

The chi-square statistic is calculated by (a) squaring the difference between the actual count and the expected value for each cell in the table, (b) dividing each squared difference by its expected value, and (c) summing the results over all the cells. If we identify each cell by its i^{th} row number and j^{th} column number, then the formula for the summation over all the cells can be written as follows:

$$\chi^2 = \sum_{i,j} \frac{(Actual - Expected)^2}{Expected}$$

Applying this formula to the cells of Table D.1 results in the following chi-square estimate:

$$\chi^2 = \frac{(5-4.98)^2}{4.98} + \frac{(5-4.71)^2}{4.71} + \cdots + \frac{(6-5.91)^2}{5.91} \approx 0.62$$

There are good predictions for all the cells using the marginal counts. This accounts for the low chi-square value of about 0.62 for Table D.1, with a p-value of almost one (p=0.999) based on nine degrees of freedom. Therefore, in this case we would accept the null hypothesis that the row and column variables are independent and not associated.

Table D.2. An example of an extremely significant association between the row and column variables (χ^2=195, p<0.001). (Technical note: The chi-square value of 195 is, in fact, the maximum possible value of the chi-square for a table with this total count and dimension, and is equal to N, the total count, multiplied by the min($r-1$, $c-1$), where r and c are the numbers of rows and columns: $65 \cdot 3 = 195$.)

	Column Variable				**Totals**
	c1	c2	c3	c4	
Row Variable					
r1		18			18
r2	10				10
r3			13		13
r4				24	24
Totals	10	18	13	24	65

Table D.2 is an example of the extreme opposite situation. In this case we have perfect predictions. This can be seen without any calculations by noting that we know the outcomes for the row variable once we know the outcomes for the column variable, and vice versa. Each variable is completely determined by the other. In this case, the estimated chi-square is 195 with a p-value less than 0.001. Therefore, we would reject the hypothesis that the row and column variables are independent.

Table D.3. An example of a significant association between the row and column variables (χ^2=84.2, p<0.001).

	Column Variable				**Totals**
	c1	c2	c3	c4	
Row Variable					
r1	4	13	1		18
r2	1	7	2		10
r3		2	10	1	13
r4			2	22	24
Totals	5	22	15	23	65

In practice, we would probably never see an association as strong as that found in Table D.2, but we might observe one like Table D.3, which has an estimated chi-square of 84.2 with a p-value less than 0.001. An examination of this table reveals that some combinations of the row and column categories tend to co-occur much more often. This tendency is usually what is referred to when we say that the row and column variables are 'correlated.'

When we are examining an association, there are generally two attributes of interest: (1) a measure of the strength of the association, and (2) a measure of how real the association is. For example, in regression analysis we often rely on the size of the regression coefficient to indicate how strong the relationship is, whereas we look at the p-value to indicate whether the relation is real or not, in other words, whether this relation could have occurred by chance alone. Generally speaking, the chi-square statistic is a measure of whether or not the association is real (i.e., is statistically significant). However, an attractive property of the chi-square statistic is that it can be split into disjoint components, and this feature is used in correspondence analysis when the inertia is derived from the chi-square in order to assess the relative strength of the contributions to the inertias for various dimensions and the categories of variables.

References

1. Cochran, W.G. The planning of observational studies of human populations. *J R Stat Soc Ser A Stat Soc* 1965; 128:234-266.
2. Rosenbaum, P.R. *Observational Studies* (Second Edition). New York: Springer-Verlag; 2002.
3. Nenadić, O. and Greenacre, M.J. Correspondence analysis in R, with two- and three-dimensional graphics: The **ca** package. *J Stat Softw* 2007; 20:1-13.
4. Tukey, JW. Nonparametric statistical data modeling: comment. *J Am Stat Assoc* 1979; 74:121-122.
5. Tukey, J.W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
6. Daniel, C. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics* 1959; 1:311-341.
7. Wilk, M.B. and Gnanadesikan, R. Graphical methods for internal comparisons in multiresponse experiments. *Ann Math Stat* 1964; 35:613-631.
8. Johnson, E.G. and Tukey, J.W. Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In: Mallows, C.L., editor. *Design, Data and Analysis*. New York: Wiley; 1987. p. 171-244.
9. Mallows, CL and Tukey, JW. An overview of techniques of data analysis, emphasizing its exploratory aspects. In: de Oliviera, T. and Epstein, B., editors. *Some Recent Advances in Statistics*. New York: Academic Press; 1982. p. 111-172.
10. Weller, S.C. and Romney, A.K. *Metric Scaling: Correspondence Analysis*. London: Sage Publications; 1990.
11. ter Braak, C.J.F. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 1986; 67:1167-79.
12. Greenacre, M. and Pardo, R. Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research* 2007; 35:193-218.
13. Michailidis, G. and de Leeuw, J. The Gifi system of descriptive multivariate analysis. *Stat Sci* 1998; 13:307-336.
14. Gifi, A. *Nonlinear Multivariate Analysis*. New York: Wiley; 1990.
15. Hill, M.O. Correspondence analysis: a neglected multivariate method. *J R Stat Soc Ser C Appl Stat* 1974; 23:340-354.
16. Greenacre, M. and Hastie, T. The geometric interpretation of correspondence analysis. *J Am Stat Assoc* 1987; 82:437-447.
17. Bergman, L.R. and Magnusson, D. A person-oriented approach in research on developmental psychology. *Dev Psychopathol* 1997; 9:291-319.
18. Kendall M. *Multivariate Analysis*. New York: MacMillan; 1980.
19. Morrison, D.F. *Multivariate Statistical Methods* (Third Edition). New York: McGraw-Hill; 1990.
20. Carroll, J.D., Green, P.E., and Chaturvedi, A. *Mathematical Tools for Applied Multivariate Analysis*. London: Academic Press; 1997.
21. Izenman, A.J. *Modern Multivariate Statistical Techniques*. New York: Springer; 2008.
22. Fleiss, J.L. and Zubin, J. On the methods and theory of clustering. *Multivariate Behav Res* 1969; 4:235-250.
23. Chatfield, C. and Collins, A.J. *Introduction to Multivariate Analysis*. Boca Ration, FL: Chapman & Hall; 1981.
24. Nishisato, S. *Multidimensional Nonlinear Descriptive Analysis*. London: Chapman & Hall; 2006.
25. Greenacre, M. *Theory and Applications of Correspondence Analysis*. London: Academic Press; 1984. (Available at http://www.carme-n.org/ .)
26. Clausen, S-E. *Applied Correspondence Analysis*. London: Sage Publications; 1998.
27. Greenacre, M. Correspondence analysis in medical research. *Stat Methods Med Res* 1992; 1:97-117.
28. Greenacre, M. *Correspondence Analysis in Practice* (Second Edition). Boca Ration, FL: Chapman & Hall; 2007.
29. Friendly, M. *Visualizing Categorical Data*. Cary, NC: SAS Institute; 2000.
30. International Labour Office. *Guidelines for the Use of the ILO International Classification of Radiographs*

of Pneumoconioses, Revised Edition 2000. Geneva: International Labour Office; Occupational Safety and Health Series, Vol. 22, 2002.
31. Sarle, W.S. *Cubic Clustering Criterion* (SAS Technical Report A-108). Cary, NC: SAS Institute; 1983.
32. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition). New York: Springer-Verlag; 2009.
33. Kay, P. A Guttman scale model of Tahitian consumer behavior. *Southwest J Anthropol* 1964; 20:160-167.
34. Digby, P.G.N. and Kempton, R.A. *Multivariate Analysis of Ecological Communities*. London: Chapman and Hall; 1987.
35. Srole, L. *Mental Health in the Metropolis*. New York: McGraw-Hill; 1962.
36. Greenacre MJ. Clustering the rows and columns of a contingency table. *J Classif* 1988; 5:39-51.
37. SAS Institute. *SAS/STAT 9.3 Users Guide*. Cary, NC: SAS Institute; 2011.
38. Cleveland, W. *The Elements of Graphing Data*. Summit, NJ: Hobart Press; 1994.
39. Jacobsen, M. Against Popperized epidemiology. *Int J Epidemiol* 1976; 5:9-11.
40. D'Enza, A.I., Palumbo, F. and Greenacre, M. Exploratory data analysis leading towards the most interesting simple association rules. *Computational Statistics & Data Analysis* 2008; 52:3269-3281. DOI: 10.1016/j.csda.2007.10.006
41. Crichton, N.J. and Hinde, J.P. Correspondence analysis as a screening method for indicants for clinical diagnosis. *Stat Med* 1989; 8:1351-1362.
42. Lebart, L. Complementary use of correspondence analysis and cluster analysis. In: Greenacre, M. and Blasius, J., editors. *Correspondence Analysis in the Social Sciences*. London: Academic Press; 1994. p. 162-178.
43. Pasta, D.J. Learning when to be discrete: continuous vs. categorical predictors. *Proceedings of the SAS Global Forum 34* 2009; Paper 248.
44. Cox, D.R. Note on grouping. *J Am Stat Assoc* 1957; 52:543-547.
45. Hosmer, D.W. and Lemeshow, S. *Applied Logistic Regression*. New York: Wiley; 1989.
46. Lemeshow, S., Teres, D., Avrunin, J.S., and Pastides, H. Predicting the outcome of intensive care unit patients. *J Am Stat Assoc* 1988; 83:348-356.
47. Cox, D.R. and Wermuth, N. *Multivariate Dependencies (Models, Analysis and Interpretation)*. London: Chapman and Hall; 1996.
48. Ortega, H.G., Daroowalla, F., Petsonk, E.L., Lewis, D., Berardinelli, S., Jones, W., Kreiss, K. and Weissman, D. Respiratory symptoms among crab processing workers in Alaska: epidemiological and environmental assessment. *Am J Ind Med* 2001; 39:598-607.
49. Woolson, R.F. and Clarke, W.R. Analysis of categorical incomplete longitudinal data. *J R Stat Soc Ser A Stat Soc* 1984; 147:87-99.
50. Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley; 2004.
51. Lombard, H.L. and Doering, C.R. Treatment of the four-fold table by partial association and partial correlation as it relates to public health problems. *Biometrics* 1947; 3:123-128.
52. Cook, E.D. *Solution Manual to Accompany Applied Logistic Regression* (Second Edition). New York: Wiley; 2001.
53. Firth, D. Bias reduction of maximum-likelihood estimates. *Biometrika* 1993; 80:27-38.
54. Audigier, V., Husson, F. and Josse, F. MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing* 2017; 27:501-518.
55. Collins, L.M. and Lanza, S.T. *Latent Class and Latent Transition Analysis*. New York: Wiley; 2010.
56. Winsor, C.P. Factorial analysis of a multiple dichotomy. *Hum Biol* 1948; 20:195-204.
57. Melamed, D., Breiger, R.L. and Schoon, E. The duality of clusters and statistical interactions. *Sociological Methods & Research* 2013; 42:41-59.
58. Goodman, L.A. The multivariate analysis of qualitative data: interactions among multiple classifications. *J Am Stat Assoc* 1970; 65:226-256.
59. Neter, J., Wasserman, W. and Kutner, M.H. *Applied Regression Models*. Homewood, IL: Irwin; 1983.
60. McNemar, Q. Notes on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; 12:153-157.
61. Bowker, A.H. Bowker's test of symmetry. *J Am Stat Assoc* 1948; 43:572-574.
62. Broaddus, V.C., Mason, R.J., Ernst, J.D., King, T.E., Lazarus, S.C., Murray, J.F., Nadel, J.A., Slutsky, A.S.,

and Gotway, M.B. *Murray and Nadel's Textbook on Respiratory Medicine* (6th Edition). Philadelphia: Elsevier Saunders; 2015.

63. Mirkin, B. Eleven ways to look at the chi-squared coefficient for contingency tables. *Am Stat* 2001; 55:111-120.