Compositional Data Analysis in Population Studies

Christopher D. Lloyd,* Vera Pawlowsky-Glahn,† and Juan José Egozcue‡

*School of Geography, Archaeology, and Palaeoecology, Queen's University Belfast
†Departament Informàtica i Matemàtica Aplicada, Universitat de Girona
‡Departament Matemàtica Aplicada III, Universitat Politècnica de Catalunya

Many analyses of population data are based on percentages or proportions. Such data are referred to as *compositional* and complete compositions typically sum to 100 (if expressed as percentages) or one (if expressed as proportions). In geography, the vast majority of analyses of such data make use of standard statistical approaches. There are many warnings, however, both in the geographical literature and elsewhere, that such standard approaches should not be employed directly in the analysis of such data. Aitchison (1986) proposed a methodology for the analysis of compositional data based on taking log-ratios and then applying adapted standard approaches to the transformed data. This article highlights some problems associated with using standard approaches in the analysis of compositional data in population studies and outlines some key log-ratio-based approaches for transformation and the proper analysis of compositional data. A case study is used to demonstrate the application. *Key Words: census data, log-ratios, percentages, proportions.*

许多人口数据是基于百分比或比例的分析基础上的。这些数据被称为成分和完整的成分，通常总和到100（如果以百分比表示）或1（如果以比例表示）。在地理学中，对这些数据的绝大多数分析，使用标准的统计方法。但是，无论是在地理还是其他文献中，多次提醒这样的标准方法不应该被直接用于这些数据的分析。艾奇逊（1986）提出了一个成分数据分析的方法，该方法以对数比值为基础，把适应的标准方法应用到转换后的数据。本文着重介绍了在人口研究中与成分数据分析里使用的标准方法相关的一些问题，并概述了一些转换和正确分析成分数据的关键对数比值法。个案研究被用于演示的应用程序。*关键词：人口普查数据，对数比，百分比，比例。*

Muchos de los análisis de datos de población se basan en porcentajes o proporciones. Tales datos son referidos como *composicionales*, y las composiciones completas típicamente suman 100 (si se expresan como porcentajes) o uno (si se expresan como proporciones). En geografía, la vasta mayoría de los análisis de tales datos hacen uso de enfoques estadísticos estándar. Se han expresado, sin embargo, numerosas alertas, tanto en la literatura geográfica como de otro tipo, en el sentido de que tales enfoques estándar no deberían emplearse directamente en el análisis de aquellos datos. Aitchison (1986) propuso una metodología para el análisis de datos composicionales a partir de la toma de *log-ratios* para luego aplicar enfoques estándar adaptados a los datos transformados. Este artículo resalta algunos problemas que se asocian con el uso de enfoques estándar en el análisis de datos composicionales en estudios de población y esquematiza algunos enfoques claves basados en log-ratio para la transformación y el análisis apropiado de datos composicionales. Se utilizó un estudio de caso para demostrar la aplicación. *Palabras clave: datos censales, log-ratios, porcentajes, proporciones.*

The absolute number of people in a given class (e.g., those aged under thirty) in a given area is likely to be of less interest than the proportion or percentage of people in that class. Proportions or percentages form the core of many quantitative analyses in population studies, geography, and cognate disciplines (Evans and Jones 1981). Sets of values that sum to some fixed value like 100 are termed *compositions*. Such data are closed; that is, they have a lower and an upper limit (e.g., 0 and 1, or 0 and 100). In other words, compositional data have a restricted space—the possible values can vary from 0 to 100, 0 to 1, or another constant (Pawlowsky-Glahn and Egozcue 2006). This restricted sample space is called the *simplex*. Percentages and proportions can be obtained from the closure of a set of counts. For example, three population subgroup counts (e.g., counts of persons by religion) of 1,065, 2,528, and 1,631 expressed as percentages are 20.35 percent, 48.49 percent, and 31.16 percent. Subcompositions might also be of interest; for example, a subset of groups of people in an area can be expressed as a percentage of the total population of this subset of groups (e.g., people in a given religion subdivided by employment status, with percentages for members of that religion summing to 100). The issue of subcompositions is discussed in more detail later. As Wrigley (1973) and Evans and

Annals of the Association of American Geographers, 102(6) 2012, pp. 1251–1266
Initial submission, January 2011; final acceptance, April 2011
Published by Taylor & Francis, LLC.

Jones (1981) have argued in specifically geographical contexts, there are several reasons why compositional data should not be analyzed using standard statistical approaches. Despite these and many other warnings, conventional analyses of closed data are commonplace in population studies and throughout the physical and social sciences. It is the focus of this article to consider some potential problems that could arise in the analysis of compositional data using standard approaches and to illustrate some methods that overcome these problems. The substantive focus is on the analysis of census data, in which context counts are frequently expressed as percentages. Obvious examples include percentages of people by age groups, housing tenure, or occupational classification.

Compositions contain information on *relative* frequencies, not *absolute* frequencies (Aitchison 1986). This recognition formed the basis of a series of developments by the statistician John Aitchison. To overcome the problems associated with direct analysis of compositional data, Aitchison proposed transformations based on log-ratios that can be applied to compositional data. The log-ratio breaks the sum constraint and opens up the data—some forms of log-ratio transformation allow statistical analyses without the constraints imposed by analyzing compositional data directly. Log-ratios are familiar in log-odd calculation and in logistic regression (Aitchison, Kay, and Lauder 2004). The results of Aitchison's research were summarized in a monograph published in 1986 that describes some of the problems with the analysis of compositional data using standard statistical approaches and solutions to some of these problems. Introductions to some key principles are provided by Pawlowsky-Glahn and Egozcue (2006) and Egozcue (2009).

Problems with the Analysis of Compositional Data

The analysis of compositional data using, for example, conventional correlation and regression is common in the population studies literature. As an example, Catholics (%; x) in Northern Ireland can be plotted against Protestants and other Christians (%; y). Given the regression equation $y = 80.733 - 0.855x$, for an area with 99 percent Catholics, the predicted percentage of Protestants and other Christians would be –3.912 percent; clearly this is nonsensical. The predictions should be constrained to the sample space (in this case 0 and 100) and a solution is detailed later.

Many researchers have identified problems with the application of conventional statistical methods to the analysis of compositional data. One problem that has received much attention is spurious correlation, and research on this topic dates back to the seminal paper of Pearson (1897). Parts of a composition are always positive and are usually represented as constrained to some constant. Because compositional data are not free to vary independently like unconstrained data, there are implications in that at least one covariance (and, therefore, correlation coefficient) must be negative. In this case, none of the correlation coefficients is free to range from –1 and +1. Spurious correlations are induced between components of a composition because the data are closed, with a bias toward negative correlations. Pawlowsky-Glahn and Egozcue (2006) give the example of a two-part composition summing to a constant, whereby the correlation between the two elements must be –1. Distinguishing between these spurious effects and the effects that can be attributed to natural processes is not possible (Pawlowsky-Glahn and Egozcue 2006). In addition to problems with multivariate analyses, Filzmoser, Hron, and Reimann (2009) argued that univariate statistical methods are not appropriate for the direct analysis of raw compositional data (see Lloyd 2012 for a population studies example).

Another particular problem with compositional data relates to the concept of subcompositional coherence. Aitchison (1986) demonstrated that the covariance (and thus correlation coefficient) between variables can change markedly and with no apparent pattern, as parts of an original five-part composition are removed to form new four-part or three-part compositions. As an example, a composition and a subcomposition are given here. The values in the subcomposition are expressed as percentages of the first three parts of the whole composition:

x_1	x_2	x_3	x_4	s_1	s_2	s_3
10	10	10	70	33.3	33.3	33.3
20	10	10	60	50	25	25
30	30	20	20	37.5	37.5	25

Although x_1 and s_1 represent the same part, as do x_2 and s_2, the correlation coefficient for x_1, x_2 is 0.866, whereas for s_1, s_2 it is –0.839. Log-ratios ensure subcompositional coherence: If the ratio is between the same two parts, then the value will be equal irrespective of the total number of parts in the composition or subcomposition. For example, $x_1/x_2 = s_1/s_2$ (taking the second case, $20/10 = 2$ and $50/25 = 2$). Accounts of problems

with the analysis of compositional data using conventional methods are provided by, among others, Chayes (1971), Evans and Jones (1981), Aitchison (1986), and Rock (1988).

Subcompositional coherence can be also violated when using inappropriate distances between compositions. For instance, ordinary Euclidean distance between rows 1 and 2 in the preceding example yields $d_x^2 = (20 - 10)^2 + (10 - 10)^2 + (10 - 10)^2 + (60 - 70)^2 = 200$ and $d_s^2 = (50 - 33.3)^2 + (25 - 33.3)^2 + (25 - 33.3)^2 = 416.67$. Therefore, we get a larger distance in the three-part subcomposition than in the four-part composition. This is counterintuitive because a sound projection must reduce distances; Aitchison (1986) called this principle *subcompositional dominance*. This example prevents the use of distances based on differences of parts.

Wrigley (1973) and others have suggested the use of the logit transform prior to analysis of compositional data. The logit transform of a percentage value x_i is given by:

$$y_i = \ln \frac{x_i}{100 - x_i}. \quad (1)$$

The logit transform is a valid way to analyze a two-part composition. Provided the composition is closed, the first part determines the second as $x_2 = 100 - x_1$. The logit transform has serious problems, however, when there are more than two parts, as the denominator is the amalgamation of all the parts except the one considered. When balances are used (as is done later) the implied transformation can be viewed as a generalization of the logit transform to compositions of more than two parts.

Equation 1 is termed a log-ratio, and such approaches, building on the work of Aitchison (summarized in his monograph of 1986), form the basis of the analysis of compositional data. $\ln(x_1/x_2) = -\ln(x_2/x_1)$ and $\ln(x_1/x_2) = \ln(x_1) - \ln(x_2)$ and dealing with log-ratios is more straightforward than dealing with ratios (Evans and Jones 1981; Aitchison 1986).

Compositional Data Analysis

Three-part compositions can be visualized using ternary diagrams. Figure 1 gives an example of a ternary diagram using data from the 2001 Census of Population in Northern Ireland. The data are for wards ($n = 582$) and refer to religion using three groups: Catholics, Protestants and other Christians, and other religions and no religion. The data are described in more detail later.

Pawlowsky-Glahn and Egozcue (2006) provide an introduction to basic operations in compositional data analysis, and a very brief summary of some of these operations is provided here. Perturbation (indicated by $\oplus$) and powering (indicated by $\odot$) are important operations in compositional data analysis (Pawlowsky-Glahn and Egozcue 2006). Perturbation and powering are both demonstrated with reference to examples in a population studies context. Taking two compositional vectors $\mathbf{x} = [20.35, 48.49, 31.16]$ and $\mathbf{y} = [10, 60, 30]$ then $\mathbf{x}$ is perturbed by $\mathbf{y}$ as follows:

$$\mathbf{z} = \mathbf{x} \oplus \mathbf{y} = C[20.35 \cdot 10, 48.49 \cdot 60, 31.16 \cdot 30]$$
$$= C[203.50, 2909.40, 934.80] = [5.03, 71.88, 23.09],$$

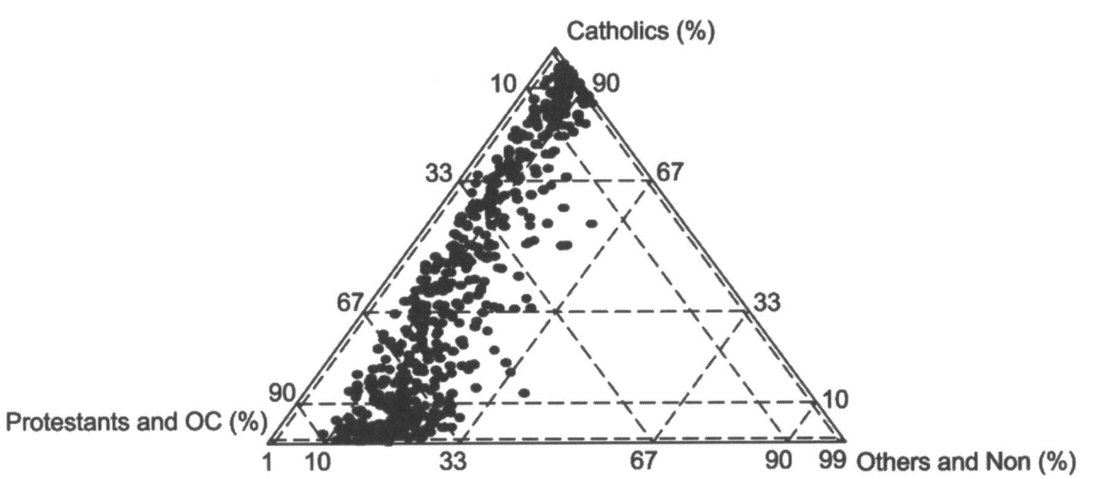

Figure 1. Ternary diagram for religion in Northern Ireland.

where C[.] indicates closure or normalization to 100 (sometimes to one), and **y** indicates the proportion of each group remaining after perturbation. Perturbation allows exploration of change on a composition. For a specific example, **y** is the percentage by religion and **x** is the percentage of people who smoke for each religion, then **x** is perturbed **y** giving the percentage of religions in the population of smokers. In this simple example, change in the composition following change in each part is computed. Alternatively, the calculations could be done with the raw counts, rather than in the simplex (i.e., 10 percent of the number of Catholics, 60 percent of the number of Protestants and other Christians, and 30 percent of others). In terms of the example, the scenario might be explored that each population group decreases (or increases) by a fixed amount over several time periods (e.g., years) and the effect on the overall composition could be considered by perturbing **z** by **y** several times, obtaining a new **z** vector at each stage. Perturbing a composition **x** by itself is given by:

$$\begin{aligned}\mathbf{x} \oplus \mathbf{x} &= [20.35,48.49,31.16] \oplus [20.35,48.49,31.16] \\ &= C[20.35^2, 48.49^2, 31.16^2].\end{aligned}$$

This process could be completed any number of times (i.e., a larger power used; even a noninteger power could be used). Other core concepts including the Aitchison inner product, Aitchison norm, Aitchison distance, Aitchison geometry, and data centering were introduced by Pawlowsky-Glahn and Egozcue (2006), and interested readers are referred to that book chapter. More detailed accounts were given by Pawlowsky-Glahn and Egozcue (2001) and Egozcue and Pawlowsky-Glahn (2006).

The remainder of this section defines different forms of log-ratios used to transform compositional data. Aitchison (1986) introduced the additive log-ratio (alr) and the centered log-ratio (clr), and Egozcue et al. (2003) developed the isometric log-ratio (ilr) transform. Outputs from the alr and clr transforms are subject to some restrictions in their treatment by standard methods, whereas ilr transformed data can be analyzed directly using standard univariate or multivariate statistical methods. Log-ratios can easily be back-transformed and results such as (back-transformed) predicted values obtained through regression will be restricted to the sample space (e.g., 0–100 for percentages). The following section defines the alr, clr, and ilr transforms and outlines some advantages and disadvantages of each.

The alr transform is also known as the *generalized logistic transformation* (Aitchison 1986). Aitchison noted that the transform is a sensible starting point as a simple transformation that was already used in other contexts. The alr transform is given by:

$$\mathbf{y} = \text{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right). \tag{2}$$

In the case of a two-part composition, this is equivalent to the logit transformation given in Equation 1. One limitation of this transform is that it is not symmetric, as the Dth component is the divisor and there are $D-1$ alr values. Returning to one particular problem outlined earlier—that of predicting values outside of the sample space using regression—predictions in terms of log-ratios can be back-transformed to proportions and the back-transformed predicted values are constrained to be within the sample space (e.g., 0–100 for percentages).

Pawlowsky-Glahn and Egozcue (2006) provided a summary of the restrictions of the different forms of log-ratios. An example of the limitations of the alr transform is provided by Pawlowsky-Glahn and Olea (2004) in a discussion about spatial prediction using kriging. The authors note that cross-validation (i.e., removal of an observation and prediction of that value using neighboring values, returning the value and conducting the same procedure at all data locations) for alr transformed data cannot be conducted using Euclidean distance in the simplex (see Pawlowsky-Glahn and Olea 2004, 120–21, with an example on 146–47). In other words, if the difference between true and cross-validation predicted values is being considered then simple differences (i.e., Euclidean distances) cannot be used to compute errors. Pawlowsky-Glahn and Olea (2004) employed two strategies—they used instead the Aitchison distance and compositional Mahalanobis distance in the simplex (i.e., using the transformed values), or else standard approaches using the back-transformed data.

The inverse of the alr transform, the generalized additive logistic transformation (agl) is given by:

$$\begin{aligned}\mathbf{x} &= \text{agl}(\mathbf{y}) \\ &= \left(\frac{\exp(y_1)}{1 + \sum_{i=1}^{D-1} \exp(y_i)}, \ldots, \right. \\ &\qquad \left. \frac{\exp(y_{D-1})}{1 + \sum_{i=1}^{D-1} \exp(y_i)}, \frac{1}{1 + \sum_{i=1}^{D-1} \exp(y_i)}\right).\end{aligned} \tag{3}$$

The clr transform is given by:

$$\mathbf{y} = \text{clr}(\mathbf{x}) = \left(\ln\frac{x_1}{g(\mathbf{x})}, \ldots, \ln\frac{x_D}{g(\mathbf{x})}\right) = \ln\frac{\mathbf{x}}{g(\mathbf{x})}, \tag{4}$$

where $g(\mathbf{x})$ is the geometric mean across the composition x, which is computed as $\left(\prod_{i=1}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right]$.

A key reason for using the clr transform is that there is symmetry in the components (Aitchison 1986, 78–79; Pawlowsky-Glahn and Olea 2004, 38). Unlike with alr transformed variables (as detailed earlier), ordinary distances can be computed using clr-transformed variables (Pawlowsky-Glahn and Egozcue 2006). The clr transform is useful for generation of biplots (Aitchison and Greenacre 2002; Pawlowsky-Glahn and Egozcue 2006), and this is illustrated later. A major restriction is that clr coefficients sum to zero and the covariance and correlation matrices are singular as their determinant is zero (Pawlowsky-Glahn and Egozcue 2006). Pawlowsky-Glahn and Olea (2004) described some problems in using cokriging (a method for spatial prediction that makes use of secondary variables to try to increase the accuracy of predictions) with clr coefficients (27, 39).

The inverse clr transformation is given by:

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{y}) = \left(\frac{\exp(y_1)}{\sum_{i=1}^{D}\exp(y_i)}, \ldots, \frac{\exp(y_D)}{\sum_{i=1}^{D}\exp(y_i)}\right). \tag{5}$$

The ilr transform is given by Egozcue et al. (2003), Egozcue and Pawlowsky-Glahn (2005), Pawlowsky-Glahn and Egozcue (2006). An example of ilr transformation is

$$\mathbf{y} = \text{ilr}(\mathbf{x}) = (y_1, \ldots, y_{D-1}) \in \Re^{D-1}, \tag{6}$$

where

$$y_i = \frac{1}{\sqrt{i(i+1)}}\ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i}\right) \text{ for } i = 1, \ldots, D-1.$$

As an illustration, following this example, the ilr transform for a five-part composition (with parts x_1, x_2, x_3, x_4, and x_5) is given by:

$$y_1 = \frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2},\ y_2 = \frac{1}{\sqrt{6}}\ln\frac{x_1x_2}{x_3^2},$$
$$y_3 = \frac{1}{\sqrt{12}}\ln\frac{x_1x_2x_3}{x_4^3},\ y_4 = \frac{1}{\sqrt{20}}\ln\frac{x_1x_2x_3x_4}{x_5^4}.$$

The outputs of the alr and ilr transforms are referred to as *coordinates*. The inverse ilr transformation is explicitly given by (Filzmoser and Hron 2008):

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{y}) = \left(\frac{\exp(z_1)}{\sum_{i=1}^{D}\exp(z_i)}, \ldots, \frac{\exp(z_D)}{\sum_{i=1}^{D}\exp(z_i)}\right), \tag{7}$$

where

$$z_i = \sum_{j=i}^{D}\frac{y_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}}y_{i-1} \text{ for } i = 1, 2, \ldots, D,$$

with $y_0 = y_D = 0$.

Aitchison (2008) provided a criticism of the use of ilr coordinates. Several authors have noted problems, in some cases, with the interpretation of ilr coordinates. Egozcue et al. (2003) and Egozcue and Pawlowsky-Glahn (2005) have developed a form of ilr coordinates called *balances*. Balances represent the relative variation in two groups of parts and they could have straightforward interpretations. Balances are particular cases of ilr coordinates and therefore they can be analyzed using standard multivariate statistical approaches. Pawlowsky-Glahn and Egozcue (2006) give a numerical example of the computation of balances. Additive aggregation of components, called *amalgamation* (for avoiding zeros, as discussed later, or simply to make analyses more straightforward) results in a loss of information and should only be done where the aggregated values are meaningful. Balances offer a means of analyzing simultaneously variation within groups of parts and between groups of parts (Egozcue and Pawlowsky-Glahn 2005; see also Filzmoser and Hron 2009). The general equation for computing balances is

$$y_i = \sqrt{\frac{rs}{r+s}}\ln\left(\frac{\left(\prod_{+} x_j\right)^{\frac{1}{r}}}{\left(\prod_{-} x_k\right)^{\frac{1}{s}}}\right) \text{ for } i = 1, \ldots, D-1. \tag{8}$$

The products $\prod_{+}$ and $\prod_{-}$ refer to parts that are coded with + or − and r and s refer to the number of

parts with positive and negative signs in the partition (as illustrated next) for the ith order (Egozcue and Pawlowsky-Glahn 2006). An example partition for six parts $(x_1, \ldots, x_6)$ can be given with:

x_1	x_2	x_3	x_4	x_5	x_6	Balance (y)	r	s
1	1	1	1	−1	−1	1	4	2
1	1	1	−1	0	0	2	3	1
1	1	−1	0	0	0	3	2	1
1	−1	0	0	0	0	4	1	1
0	0	0	0	1	−1	5	1	1

The five balances are computed, following Equation 8, with:

$$y_1 = \sqrt{\frac{8}{6}} \ln \frac{(x_1 x_2 x_3 x_4)^{\frac{1}{4}}}{(x_5 x_6)^{\frac{1}{2}}}, \quad y_2 = \sqrt{\frac{3}{4}} \ln \frac{(x_1 x_2 x_3)^{\frac{1}{3}}}{x_4},$$

$$y_3 = \sqrt{\frac{2}{3}} \ln \frac{(x_1 x_2)^{\frac{1}{2}}}{x_3}, \quad y_4 = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}, \quad y_5 = \sqrt{\frac{1}{2}} \ln \frac{x_5}{x_6}.$$

Partitions can be selected using expert knowledge or using a compositional biplot, as illustrated in the later case study. The transforms are illustrated in the Appendix, where worked examples are given.

A core tool in the analysis of compositional data is Aitchison distances. With respect to balances, these enable the exploration of the contributions of each balance. In other words, it is possible to assess the variation between balances or the contributions to balances (Egozcue and Pawlowsky-Glahn 2005).

The limitations of the different forms of log-ratios, as briefly alluded to earlier, mean that they cannot necessarily be analyzed using all standard statistical methods. The exception is ilr coordinates (and particularly balances) that can be analyzed using all standard approaches. Multivariate methods for the analysis of compositional data are introduced by Aitchison (1986). For example, Aitchison (1986) described log contrast principal component analysis, which (unlike standard principal component analysis) is suitable for compositional data.

Because logarithms cannot be taken from zero values, where there are zero values in parts of a composition some solution must be found before log-ratios can be computed. Zeros can be divided into essential zeros and rounded or trace zeros. Essential zeros are cases where there is no value to observe. Aitchison (1986) gave the example of a household spending nothing on a particular class of commodity such as "tobacco and alcohol." Trace or rounded zeros include cases where a property is below the threshold for measurement—for example, there might be no trace of a particular mineral in a rock, but that does not mean that it is completely absent. In the case of essential zeros, amalgamation is one option. That is, if it is meaningful to aggregate components such that there are no zero values after aggregation then such a step can be taken (Aitchison 1986). One should be aware, however, that after amalgamation the initial composition cannot be recovered. Another sound approach consists of modeling the whole population as two different populations: individuals presenting the essential zero and those who do not (Bacon-Shone 2003). Some strategies for modeling the zero components separately have been considered (Aitchison 1986). In the case of rounded or trace zeros, amalgamation is one option. Another possibility is to add some small value to the zero values (Aitchison 1986). Lloyd (2010b) calculated proportions from counts n_1, n_2 with $n_1 + 1$ and $n_2 + 1$ (an approach used by Mateu-Figueras and Daunis-i-Estadella 2008). In words, a value of one was added to both sets of counts and the two percentages, x_1, x_2, were calculated from the modified counts. This study categorized members of the population and zeros in this case could be considered essential—there is no individual or household in a given class in a particular area. Given the element of adjustment of 2001 UK Census data for purposes of confidentiality, however, and the uncertainty in counts associated with nonresponse to the census and the imputation of missing individuals under the One Number Census (ONC) procedure (Office for National Statistics 2001; see Williamson 2007 for a discussion about this issue), it is conceptually logical to treat these zeros as rounded or trace zeros, as we cannot be sure that individuals or households within a particular category do not exist in an area even if none are represented in the data. To assess robustness, in that study the results were compared with those obtained using other sets of zones including wards (for which there were no zero values).

Applications of Log-Ratios

The treatment of compositional data necessitates computing log-ratios, computing summary statistics, selecting appropriate methods, and analyzing the transformed data before back-transforming, if required. There are many applications of log-ratio analysis of compositional data in the academic literature. Most of these are in geological contexts, although others are found, for example, in archaeology and ecology.

Aitchison (1986) provided many case studies, and others can be found in Pawlowsky-Glahn and Olea (2004) and in the publications cited therein. Some case studies are presented in the following section. Several researchers have questioned the appropriateness of compositional data analysis on the basis that it sometimes produces results that are more difficult to interpret than analyses based on raw data (i.e., percentages or proportions). For example, see Tangri and Wright (1993) and a response by Aitchison, Barcelo-Vidal, and Pawlowsky-Glahn (2002), which argued the case for log-ratio analysis (see also Egozcue 2009). A review of recent developments and problems in compositional data analysis was given by Aitchison and Egozcue (2005). Tools for compositional data analysis are provided in the free software package CoDaPack (Thió-Henestrosa and Martín-Fernández 2005), "compositions," a package written in the R language (van den Boogaart and Tolosano-Delgado 2008), and "robCompositions," also written in R and devoted to robust statistics (Templ, Hron, and Filzmoser 2010).

In an application making use of log-ratios in a population study context, Lloyd (2010b) computed the Moran's I spatial autocorrelation coefficient for a set of ilr transformed data. For ilr(Catholics/NonCatholics) given observations for 1 km^2 cells, $I = 0.752$ (using adjacent cells); the next largest value of I was 0.269 for ilr(SGAB,C1/SGC2; i.e., approximated social grade AB × approximated social grade C1/approximated social grade C2). This suggests that there is a greater degree of spatial structuring in the population by religion than by other major population characteristics such as unemployment or housing tenure. Lloyd (2010b) also computed bivariate correlation coefficients between each of the ilr transformed variables, including those derived using the percentages of people aged less than or equal to fifteen, sixteen to twenty-nine, thirty to sixty-four, and greater than or equal to sixty-five. For the reasons detailed previously, correlations should not be computed from raw percentages. Lloyd (2010a) utilized ilr coordinates as the basis of an analysis of population data using principal components analysis. In that analysis, fourteen ilr variables were used. These were constructed from five sets of counts divided into two groups (tenure, car ownership, employment, community background, and limiting long-term illness), there was one three-part composition (qualifications), one four-part composition (age), and one five-part composition (social grade). The percentages were computed from these sets of counts—so the compositions were two-, three-, four-, or five-part compositions and the ilr variables were computed given these percentages. The raw percentages could not be analyzed directly for the reasons detailed previously and the study demonstrated that this approach enabled the effective characterization of the population of Northern Ireland in terms of a range of socioeconomic, demographic, and religious variables.

Case Study

Some of the approaches outlined in this article are illustrated with reference to a single set of data taken from the 2001 Census of Population of Northern Ireland. The case study deals with religion in Northern Ireland in 2001 by wards ($n = 582$). The counts are numbers of people by religion, divided into seven groups: Presbyterian (Pres), Church of Ireland (COI), Methodist (Meth), other Christian (OC), Catholic (Cath), other religions (OR), and no religion or none stated (NN). Proportions were calculated from counts n_1, n_2 with $n_1 + 1$ and $n_2 + 1$ following Mateu-Figueras and Daunis-i-Estadella (2008), as outlined earlier. Figure 2 shows, as an example, Catholics as a percentage of all persons and the well-known distinction between a predominantly Catholic west and a predominantly Protestant east is apparent.

Table 1 gives the compositional descriptive statistics and Table 2 contains the variation array. The clr variances are given in Table 3. In the variation array (Table 2), the means are the average of simple log-ratios between a pair of parts and the variances are sample variances of the same simple log-ratios. Table 2 indicates that Cath accounts for the largest variability with respect to other religions, as the simple log-ratios involving Cath are the largest, except for the simple log-ratio ln(Cath/NN); see also the clr variance. The component with the second largest variabilities is OR.

Table 1. Compositional descriptive statistics

	Pres	COI	Meth	OC	Cath	OR	NN
Center	0.19	0.16	0.02	0.06	0.37	0.00	0.19
Min	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.03
Max	0.67	0.50	0.22	0.19	0.96	0.03	0.37
Q25	0.08	0.08	0.00	0.02	0.10	<0.01	0.08
Median	0.22	0.15	0.02	0.06	0.34	<0.01	0.12
Q75	0.33	0.22	0.05	0.09	0.68	<0.01	0.18

Note. Pres = Presbyterian; COI = Church of Ireland; Meth = Methodist; OC = other Christian; Cath = Catholic; OR = other religion; NN = no religion or none stated.

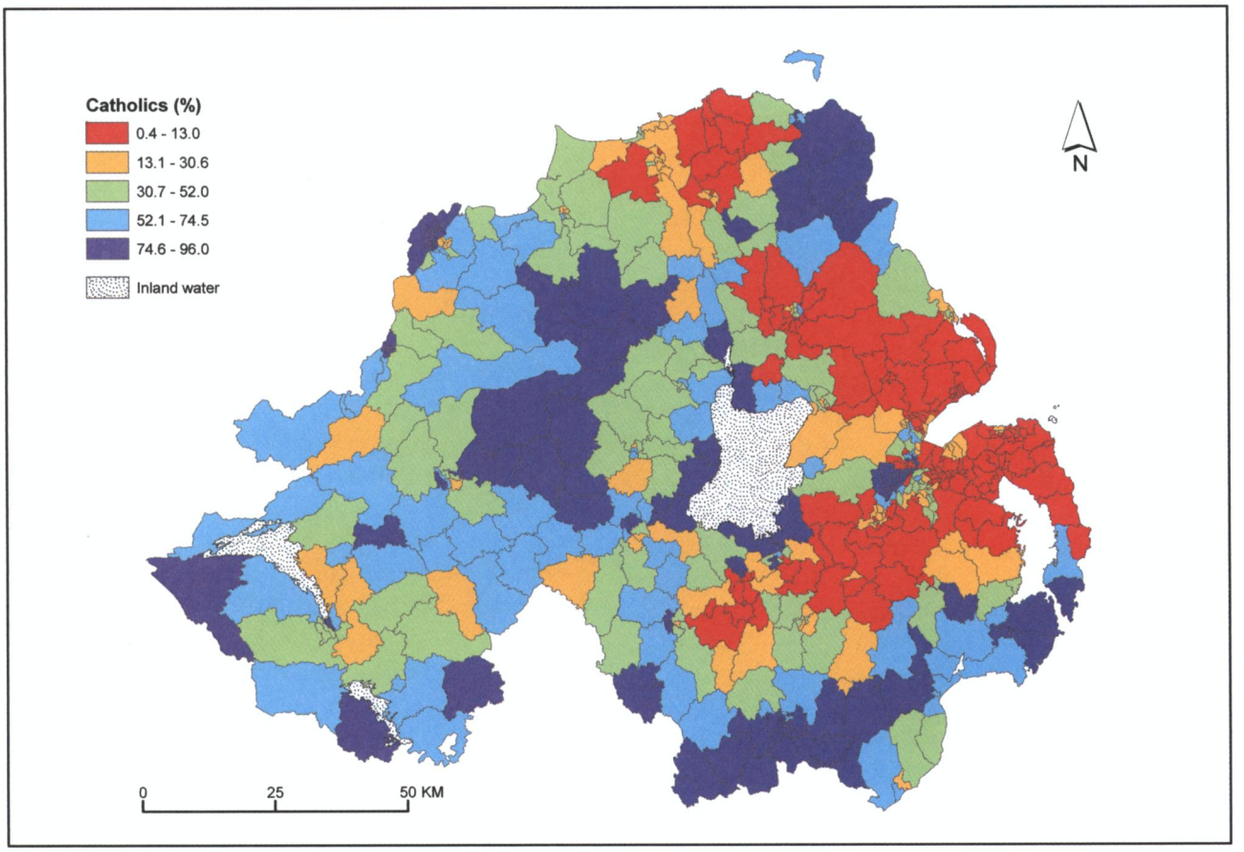

Figure 2. Catholics as a percentage of all persons by census ward, 2001. (Color figure available online.)

In this example, balances are computed with partitions constructed based on prior knowledge and the variation array and through the use of a compositional biplot. Biplots (Gabriel 1971; Udina 2005) allow the simultaneous visualization of information on variables and observations of a data matrix. The CoDaPack software includes functionality for construction of compositional biplots (Aitchison and Greenacre 2002) and this is made use of in this example. Figure 3 shows a compositional biplot for the clr transform of the seven religion categories. The lines projecting from the center are termed rays and the observations are indicated by the points. The different directions for the different rays indicate the importance of these religions relative to any other religion in distinguishing members of the population. The ray for the Catholic group can be very clearly distinguished from those for all other groups.

Table 2. Variation array

	Variances (upper triangle)						
	Pres	COI	Meth	OC	Cath	OR	NN
Pres		0.77	1.84	0.60	6.26	2.63	1.74
COI	0.19		1.17	0.56	5.19	2.09	1.32
Meth	2.20	2.00		1.21	7.42	2.74	2.03
OC	1.15	0.96	−1.05		5.21	1.90	1.07
Cath	−0.65	−0.84	−2.85	−1.80		3.31	2.71
OR	4.26	4.07	2.06	3.11	4.91		0.93
NN	0.03	−0.16	−2.17	−1.12	0.68	−4.23	
Means (lower triangle)							
Total variance: 7.52							

Note. Pres = Presbyterian; COI = Church of Ireland; Meth = Methodist; OC = other Christian; Cath = Catholic; OR = other religion; NN = no religion or none stated.

Table 3. Clr variance

Pres	COI	Meth	OC	Cath	OR	NN
0.99	0.79	1.17	0.75	2.15	0.97	0.70

Note. Pres = Presbyterian; COI = Church of Ireland; Meth = Methodist; OC = other Christian; Cath = Catholic; OR = other religion; NN = no religion or none stated.

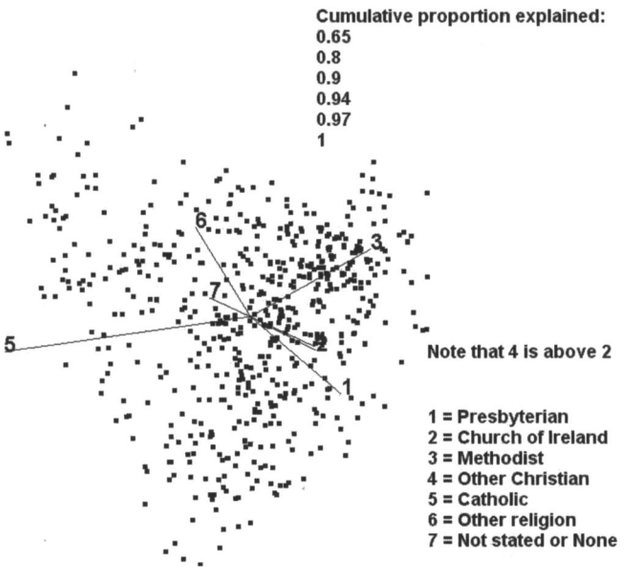

Figure 3. Compositional biplot.

The biplot reflects clearly the variability pattern of the variation array: The largest ray is the one corresponding to Cath, which is expected given the concentration of the largest variances in the array, followed by the rays corresponding to Meth, OR, and Pres. The links between rays in the biplot are informative—the link joining 4 (OC) and 5 (Catholic), extended until it meets the link joining 6 (OR) and 7 (NN), is nearly orthogonal to the latter and thus a (near) zero correlation between the corresponding balances would be expected. This is explored later.

This finding meets with expectations in that the Catholic group would be expected to be distinct from the Protestant denominations Presbyterian, Church of Ireland, Methodist, and other Christian (Protestants and related). This suggests that the choice of partition should include a Catholic–Protestant distinction. Given this information, one possible portioning scheme, which starts with a split between Protestants and Catholics and everyone else, is given in Table 4.

Table 4. Partition scheme

OR	NN	Cath	COI	Meth	Pres	OC	Balance
1	1	−1	−1	−1	−1	−1	1
1	−1	0	0	0	0	0	2
0	0	1	−1	−1	−1	−1	3
0	0	0	1	1	1	−1	4
0	0	0	1	1	−1	0	5
0	0	0	1	−1	0	0	6

Note. Pres = Presbyterian; COI = Church of Ireland; Meth = Methodist; OC = other Christian; Cath = Catholic; OR = other religion; NN = no religion or none stated.

Table 5. Correlation matrix for the balances

	B1	B2	B3	B4	B5	B6
B1	1	0.310	0.496	−0.191	0.103	0.719
B2	0.310	1	−0.059	−0.109	−0.008	0.257
B3	0.496	−0.059	1	−0.406	−0.194	0.769
B4	−0.191	−0.109	−0.406	1	0.062	−0.401
B5	0.103	−0.008	−0.194	0.062	1	−0.271
B6	0.719	0.257	0.769	−0.401	−0.271	1

This enables a split, for balance 3, between the Protestant groups and the Catholic group. Balances computed with these partitions provide the basis of the following analyses.

One way of representing the partition, along with statistical summaries, is to compute a balance dendrogram (Thió-Henestrosa et al. 2008). Figure 4 shows a balance dendrogram with respect to the example given here. As discussed by Thió-Henestrosa et al. (2008), the balance dendrogram shows the partition of the composition using vertical and horizontal links. The sample mean or center with respect to each balance is the point where the vertical bars end. The length of the vertical bars represents the proportion of the sample total variance that corresponds to each balance. The sum of these lengths represents the total variance (Thió-Henestrosa et al. 2008). For each balance, summary statistics are represented by the box plots, including the percentiles 5, 25, 50, 75, and 90, on each horizontal bar. The balance dendrogram provides a convenient graphical summary of a composition given a particular portioning scheme.

Again the balance Cath/all other Christians (B3) represents the most important part of the total variance. As the next stage of the analysis, the bivariate correlations between the balances (B) were computed and are given in Table 5.

As suggested by the biplot, the correlation between B2 and B3 is very close to zero. The largest positive correlation between the balances is with respect to B3 and B6 (0.769), suggesting a moderate positive relationship between the log-ratio of Cath to COI, Meth, Pres, and OC and the log-ratio of COI to Meth. The two balances are plotted against each other in Figure 5. The nature of this relationship suggests that the geography of the Catholic population, in terms of the share of that group, relates to the geography of COI members relative to Methodists. This issue is explored further later.

The largest negative correlation is between B3 and B4. This suggests that there is a weak linear relationship between the log-ratio of Cath to COI, Meth, Pres, and OC and the log-ratio of COI, Meth, and Pres, to OC with a correlation coefficient of −0.406.

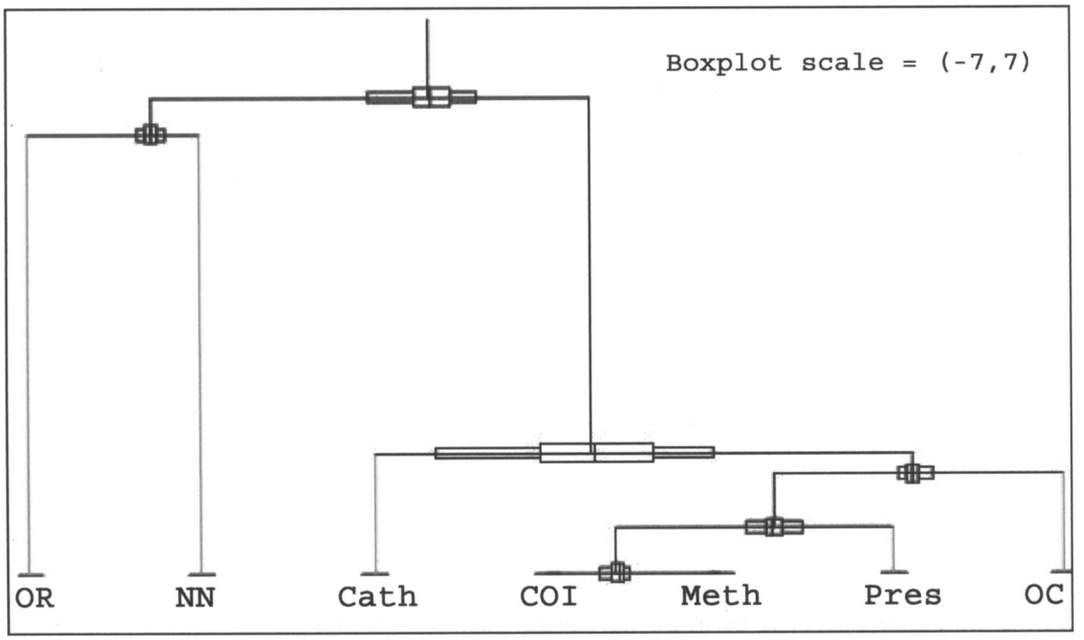

Figure 4. Balance dendrogram.

The exploration of spatial dependence is often a key part of summarizing the spatial properties of a variable. The measure of spatial autocorrelation encountered most frequently in the spatial analysis literature is the I coefficient proposed by Moran (1950; Cliff and Ord 1973). The weights, w_{ij}, between locations i and j are often row-standardized (i.e., they sum to one), as in the present analysis, and in this case Moran's I can be given by:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (9)$$

where the values y_i have the mean $\bar{y}$ (the use of y indicates, following Equation 1, that the variables are ilr transformed for this study). The proximity between locations i and j given by w_{ij} is often set to 1 when locations i and j are neighbors, and 0 when they are not, with $w_{ij} = 0$ when $i = j$ (with the weights subsequently row-standardized in this analysis). In this analysis, the Moran's I coefficient was computed using adjacency (queen's case contiguity—using zones sharing edges and vertices) for each of the six balances. ArcGIS 9.3 software (ESRI®, Redlands, CA) was used for this purpose. Table 6 shows Moran's I for each balance.

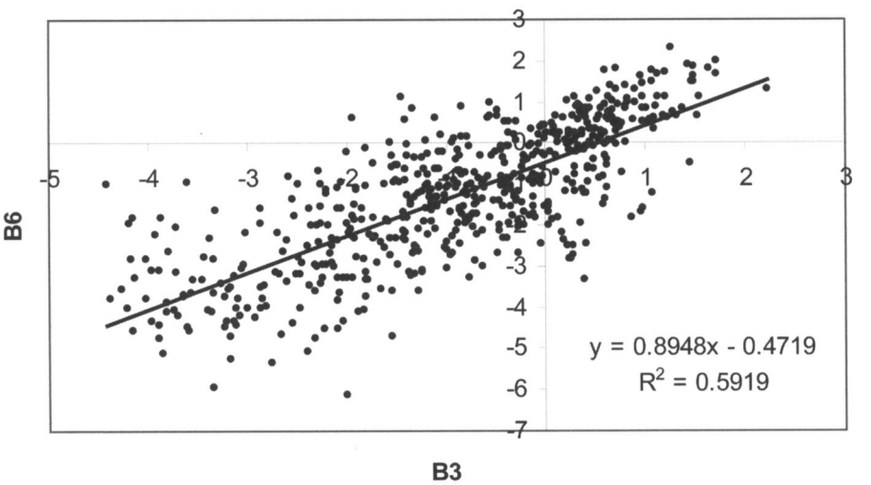

Figure 5. Scatterplot of B3 against B6.

Table 6. Moran's I for each balance: Binary weights based on contiguity with row standardization

Balance	1	2	3	4	5	6
I	0.565	0.731	0.660	0.518	0.397	0.641

All of the balances are positively spatially autocorrelated. Balance 2 has the largest value of I (0.731) and balance 3 is second with a value of 0.660. That the latter has a fairly large value is to be expected given the marked geography of the Catholic as against the Protestant population of Northern Ireland (see Shuttleworth and Lloyd 2009; Lloyd 2010b). The large figure for balance 2 (OR vs. NN) has a less obvious interpretation and reveals an unexpected characteristic of the population of Northern Ireland.

There is reason to believe that the relationships between the balances will vary spatially. For this reason, the relationship between B3 and B6 was explored further through the use of geographically weighted regression (GWR). GWR employs weights that are a function of distance from each location of interest. In this case, the geographical weights employed were determined by a variant of the Gaussian weighting scheme (see Fotheringham, Brunsdon, and Charlton [2002], and Lloyd [2011] for alternative weighting schemes):

$$w_{ij} = \exp[(-d_{ij}/\tau)^2], \tag{10}$$

where d_{ij} is the distance between locations i and j and τ is the bandwidth that determines the size of the kernel. This weighting function has been used widely in other contexts. Various studies have shown that the bandwidth, rather than the specific form of the kernel, has the greatest impact on results (Fotheringham, Brunsdon, and Charlton 2002). In this article, the bandwidth was selected using the corrected Akaike information criterion (AIC; see Fotheringham, Brunsdon, and Charlton 2002) and, through this process, a bandwidth of 13.441 km was chosen. Figure 6 shows the geographically weighted r^2 for B3 and B6.

Figure 6 confirms the results of the global correlation analysis presented earlier. It is clear that B3 and B6 are strongly linearly related in the mideast of

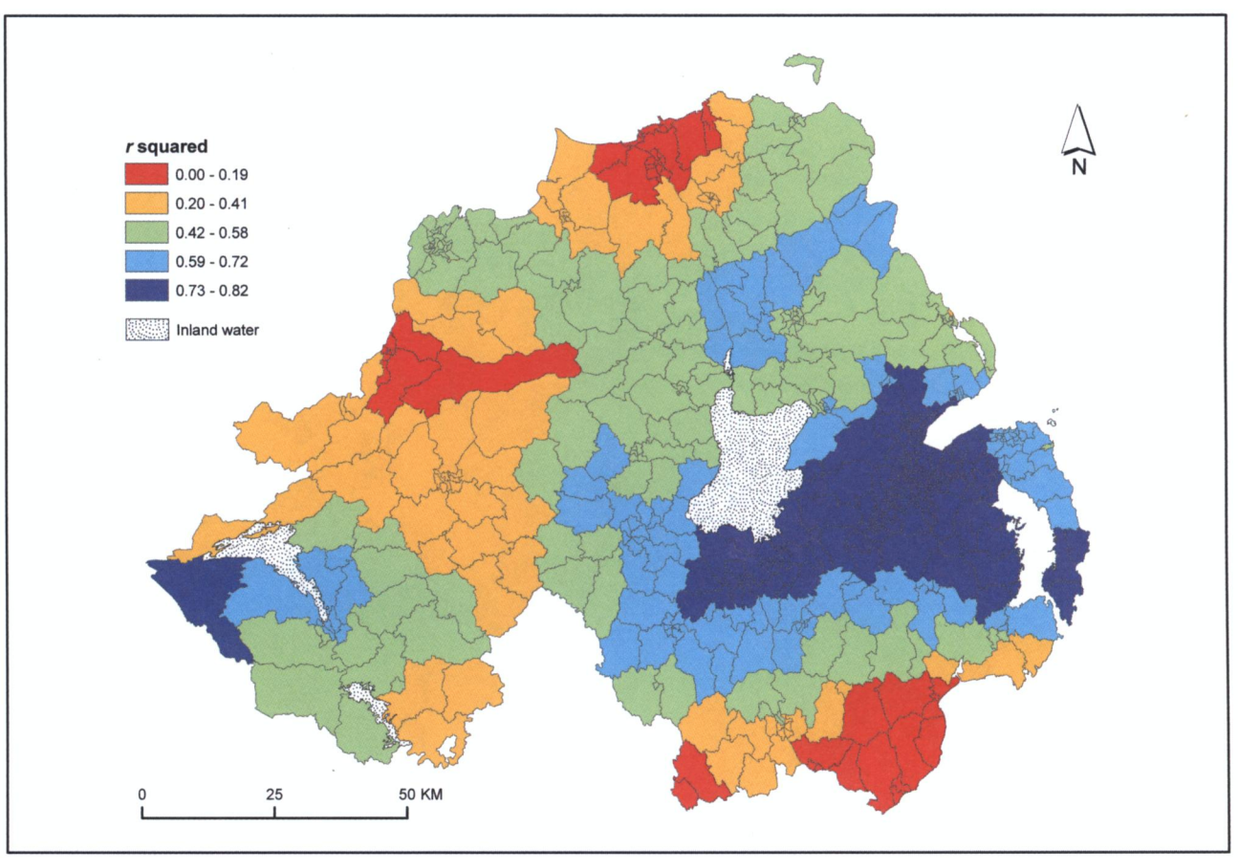

Figure 6. Geographically weighted r^2 for B3 and B6. (Color figure available online.)

Northern Ireland—the area around Belfast—and in a small area of the southwest. In words, in these areas the share of Cath relative to COI, Meth, Pres, and OC and the share of COI relative to Meth are quite strongly linearly related. This reveals additional features of the religious geographies of Northern Ireland that have not been directly alluded to before. In addition to being statistically unsound, an analysis based on raw percentages would obscure the relationships between different subsets of the population as much information is contained in the ratios between the parts of the composition. Further approaches for extracting information from the composition are explored later.

For the final stage of this analysis, agglomerative hierarchical clustering (in S-Plus 2000 software [TIBCO Software Inc., Palo, Alto, CA]) was used to generate clusters based on different combinations of balances. Three clusters were selected and the resulting map of cluster memberships for balance 3 (Catholics/all other Christians) alone is given in Figure 7, whereas clusters based on all six balances are shown in Figure 8. Comparison of Figure 7 with Figure 2 (% Catholics) indicates that wards that are members of cluster 2 are predominantly Catholic in composition. The main trends in the two maps differ in some localities, and this indicates that there is additional information contained in the cluster map. The map of clusters based on all balances (Figure 8) distinguishes some areas in common with the map in Figure 7; wards in cluster 1 include those with the largest proportion of Catholics. Wards in cluster 3 include predominantly Protestant areas, but the map reflects the differential distribution of members of different Protestant denominations (e.g., Presbyterian and Church of Ireland). This analysis supports the earlier analyses in suggesting that the ratio of Catholics to others represents most of the variation and the results reflect the religious geography of Northern Ireland, while indicating that there are subtleties in these geographies that are not apparent from examination of raw percentages. The results of cluster analysis excluding balance 3 (e.g., balances 2 and 4) are generally difficult to interpret. Cluster analysis based on balances offers a powerful (and statistically sound) means of identifying major trends in compositional data and, in this case, could be used to deconstruct the interrelations between the different religious groups of Northern Ireland.

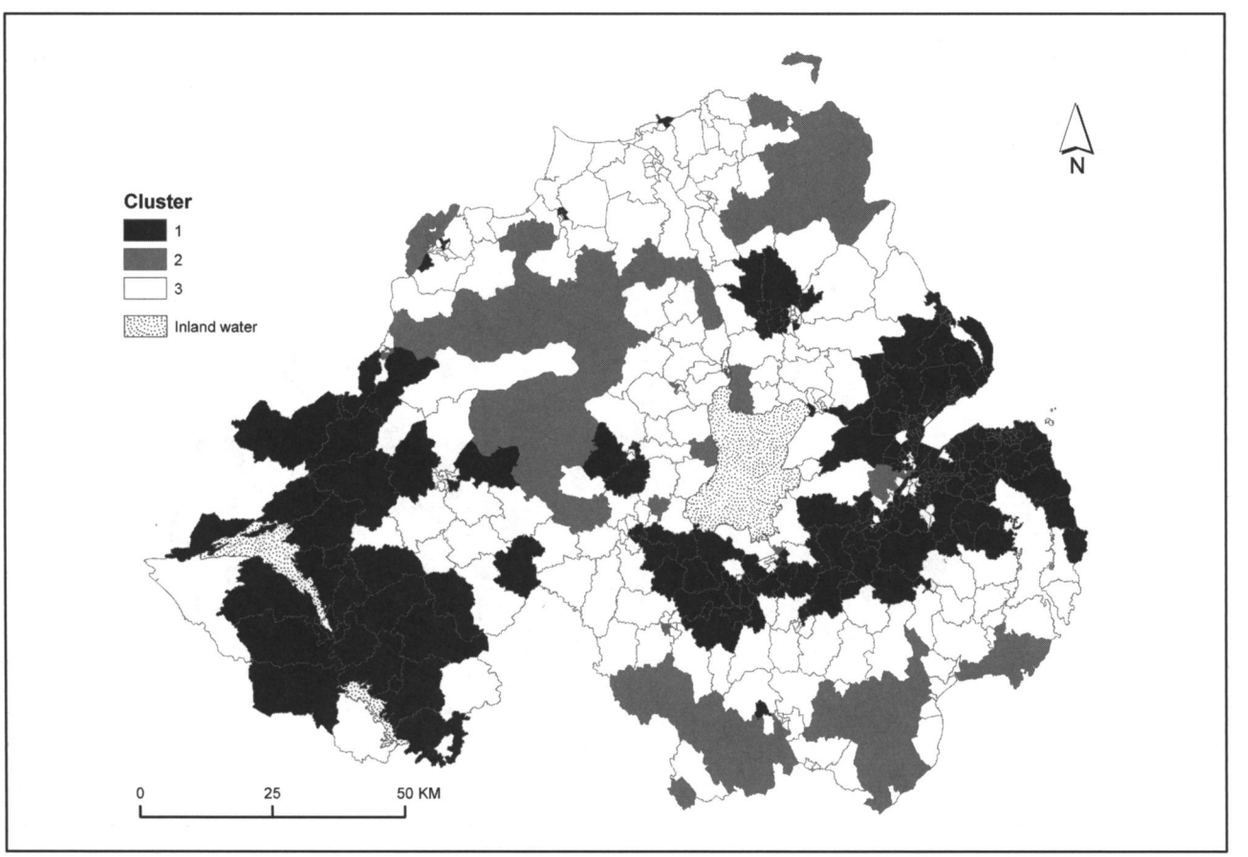

Figure 7. Clusters based on balance 3.

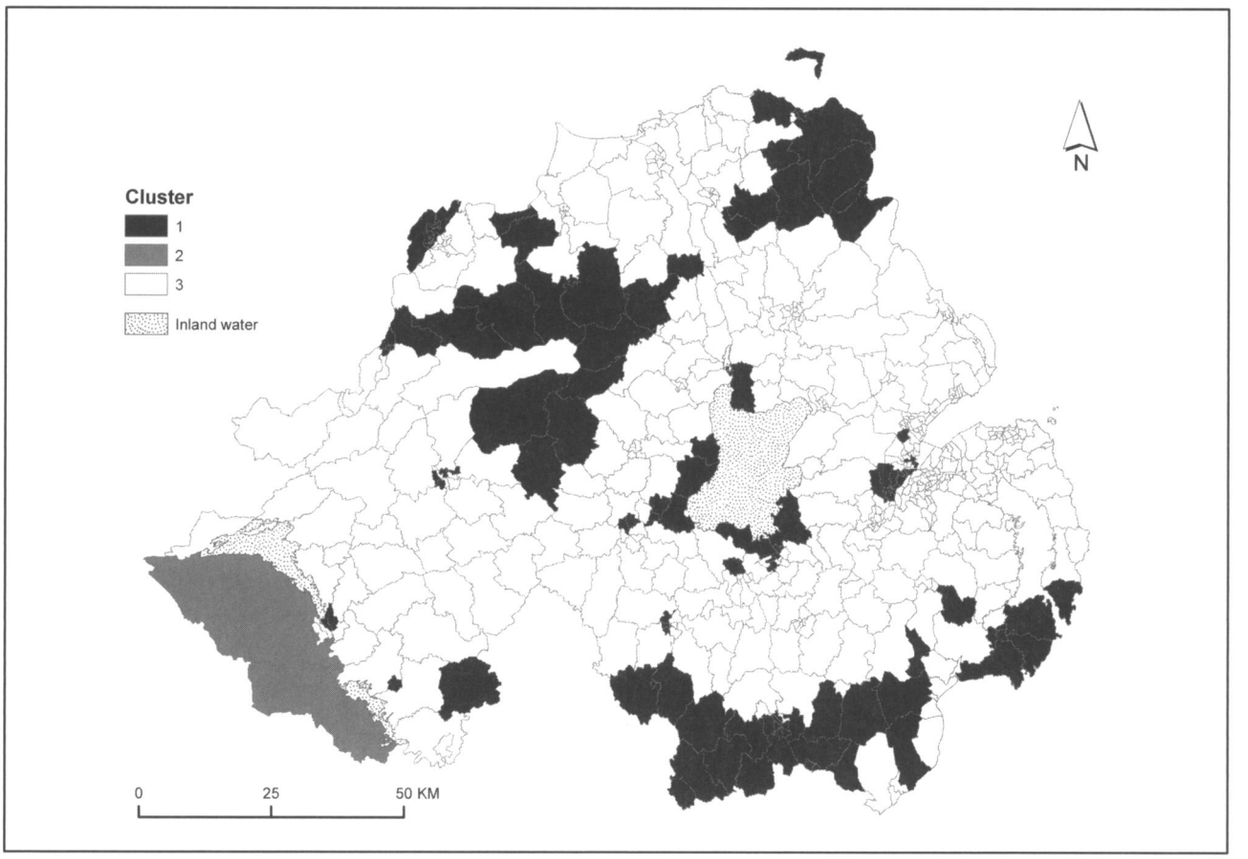

Figure 8. Clusters based on all six balances.

Even a relatively simple analysis such as that presented in this section cannot be conducted with raw percentages and this approach, making use of log-ratios, offers a way forward.

Conclusions

This article has identified some problems with the analysis of compositional data using standard statistical methods. Problems such as prediction of values outside the sample space and subcompositional incoherence have been defined and illustrated. Some approaches based on log-ratios are outlined and their application demonstrated using case studies. This article is only a brief review of compositional data analysis and necessarily skims over some important issues. Some topics likely to be of interest to population geographers are not mentioned at all because of space limitations. Analysis of correlations between different compositions (e.g., persons by age ranges and households by number of persons), as opposed to single parts of compositions, can be conducted with canonical correlation analysis. The application of canonical correlation analysis to compositional data was considered by Aitchison (1986). For many census outputs, cross-tabulations are available (e.g., economic activity by religion of persons), and such data offer further opportunities for exploring population characteristics. Aitchison (1986) termed such data *multiway compositions* and presents methods for dealing with them.

Compositional data are at the heart of quantitative population studies. A large majority of publications making use of such data treat these data as though they have an unrestricted sample space. The use of log-ratios for the analysis of compositional data has not met with universal acceptance for a variety of reasons. Many authors have seen no particular problem with the application of standard methods, and others have pointed to difficulties in interpreting results based on log-ratios (see the postscript to the 2003 printing of Aitchison [1986] for a summary). Certainly, the latter criticism holds true in many applications, but this does not validate the use of inappropriate methods for the analysis of compositional data. As noted by Lloyd (2010b),

log-ratio transforms are underused in population studies. It is hoped that this work will persuade readers of the importance of such approaches in the analysis of population data. The use of standard statistical approaches in the analysis of compositional data is not appropriate, and the methods detailed in this article and the associated references offer a way forward for the analyses of such data.

References

Aitchison, J. 1986. *The statistical analysis of compositional data*. London: Chapman and Hall.

———. 2008. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. *CODAWORK'08*. Girona, Spain: Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona. http://hdl.handle.net/10256/706 (last accessed 27 January 2012).

Aitchison, J., C. Barcelo-Vidal, and V. Pawlowsky-Glahn. 2002. Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of log-ratio analysis. *Archaeometry* 44:295–304.

Aitchison, J., and J. J. Egozcue. 2005. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology* 37:829–50.

Aitchison, J., and M. Greenacre. 2002. Biplots of compositional data. *Applied Statistics* 51:375–92.

Aitchison, J., J. W. Kay, and I. J. Lauder. 2004. *Statistical concepts and applications in clinical medicine*. Boca Raton, FL: Chapman and Hall.

Bacon-Shone, J. 2003. Modelling structural zeros in compositional data. *CODAWORK'03*. Girona, Spain: Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona. http://hdl.handle.net/10256/661 (last accessed 27 January 2012).

Chayes, F. 1971. *Ratio correlation: A manual for students of petrology and geochemistry*. Chicago: University of Chicago Press.

Cliff, A. D., and J. K. Ord. 1973. *Spatial autocorrelation*. London: Pion.

Egozcue, J. J. 2009. Reply to "On the Harker variation diagrams; ..." by J. A. Cortés. *Mathematical Geosciences* 41:829–34.

Egozcue, J. J., and V. Pawlowsky-Glahn. 2005. Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37:773–93.

———. 2006. Simplicial geometry for compositional data. In *Compositional data analysis in the geosciences: From theory to practice*, ed. A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, 145–60. London: Geological Society.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. 2003. Isometric log-ratio transformations for compositional data analysis. *Mathematical Geology* 35:279–300.

Evans, I. S., and K. Jones. 1981. Ratios and closed number systems. In *Quantitative geography: A British view*, ed. N. Wrigley and R. J. Bennett, 123–34. London and New York: Routledge and Kegan Paul.

Filzmoser, P., and K. Hron. 2008. Outlier detection for compositional data using robust methods. *Mathematical Geosciences* 40:233–48.

———. 2009. Correlation analysis for compositional data. *Mathematical Geosciences* 41:905–19.

Filzmoser, P., K. Hron, and C. Reimann. 2009. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment* 407:6100–6108.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester, UK: Wiley.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58:453–67.

Lloyd, C. D. 2010a. Analysing population characteristics using geographically weighted principal components analysis: A case study of Northern Ireland in 2001. *Computers, Environment and Urban Systems* 34:389–99.

———. 2010b. Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: An application of geographically weighted spatial statistics. *International Journal of Geographical Information Science* 24:1193–1221.

———. 2011. *Local models for spatial analysis*. 2nd ed. Boca Raton, FL: CRC Press.

———. 2012. Analysing the spatial scale of population concentrations by religion in Northern Ireland using global and local variograms. *International Journal of Geographical Information Science* 26:57–73.

Mateu-Figueras, G., and J. Daunis-i-Estadella. 2008. Compositional amalgamations and balances: A critical approach. *CODAWORK'08*. Girona, Spain: Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona. http://hdl.handle.net/10256/738 (last accessed 27 January 2012).

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37:17–23.

Office for National Statistics. 2001. *Census 2001: A guide to the One Number Census*. London: Office for National Statistics, General Register Office Scotland, Northern Ireland Statistical and Research Agency. http://www.nisra.gov.uk/archive/census/oncguide.pdf (last accessed 27 January 2012).

Pawlowsky-Glahn, V., and J. J. Egozcue. 2001. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* 15:384–98.

———. 2006. Compositional data analysis: An introduction. In *Compositional data analysis in the geosciences: From theory to practice*, ed. A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, 1–10. London: Geological Society.

Pawlowsky-Glahn, V., and R. A. Olea. 2004. *Geostatistical analysis of compositional data*. New York: Oxford University Press.

Pearson, K. 1897. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* 60:489–98.

Rock, N. M. S. 1988. *Numerical geology. Lecture Notes in Earth Sciences 18*. Berlin: Springer-Verlag.
Shuttleworth, I. G., and C. D. Lloyd. 2009. Are Northern Ireland's communities dividing? Evidence from geographically consistent Census of Population data, 1971–2001. *Environment and Planning A* 41:213–29.
Tangri, D., and R. V. S. Wright. 1993. Multivariate analysis of compositional data: Applied comparisons favour standard principal components analysis over Aitchison's loglinear contrast method. *Archaeometry* 35:103–15.
Templ, M., K. Hron, and P. Filzmoser. 2010. robCompositions: Robust estimation for compositional data. Manual and package, version 1.4.1. http://cran.r-project.org/web/packages/robCompositions/index.html (last accessed 27 January 2012).
Thió-Henestrosa, S., J. J. Egozcue, V. Pawlowsky-Glahn, L. Ó. Kovács, and G. P. Kovács. 2008. Balance-dendrogram: A new routine of CoDaPack. *Computers and Geosciences* 34:1682–96.
Thió-Henestrosa, S., and J. A. Martín-Fernández. 2005. Dealing with compositional data: The freeware CoDaPack. *Mathematical Geology* 37:773–93.
Udina, F. 2005. Interactive biplot construction. *Journal of Statistical Software* 13:1–16.
van den Boogaart, K. G., and R. Tolosano-Delgado. 2008. "compositions": A unified R package to analyze compositional data. *Computers and Geosciences* 34:320–38.
Williamson, P. 2007. The impact of cell adjustment on the analysis of aggregate census data. *Environment and Planning A* 39:1058–78.
Wrigley, N. 1973. The use of percentages in geographical research. *Area* 5:183–86.

Appendix: Worked Examples of Log-Ratio Transforms and Back-Transforms

Each of the log-ratio transforms is illustrated with reference to the three-part composition in Table A.1. (so, $D = 3$).

Table A.1. Three-part composition

Religion	Catholics	Protestants and other Christians	Others and no religion	Total
Variable no.	1	2	3	
Number	1,065	2,528	1,631	5,234
Percentage	20.35	48.49	31.16	100.00

The alr transform is given by:

$$\ln\frac{20.35}{31.16} = -0.4262,\ \ln\frac{48.49}{31.16} = 0.4422.$$

The inverse of the alr transform, the agl, is given by:

$$\frac{\exp(-0.4262)}{1+\exp(-0.4262)+\exp(0.4422)} = 0.2035,$$
$$\frac{\exp(-0.4422)}{1+\exp(-0.4262)+\exp(0.4422)} = 0.4849,$$
$$\frac{1}{1+\exp(-0.4262)+\exp(0.4422)} = 0.3116.$$

Each expression can be multiplied by 100 to obtain the original values in percentages.

The geometric mean is obtained with $(20.35 \times 48.49 \times 31.16)^{1/3} = 31.33$ and the clr transform is given by:

$$\ln\frac{20.35}{31.33} = -0.4315,\ \ln\frac{48.49}{31.33} = 0.4369,$$
$$\ln\frac{31.16}{31.33} = -0.0053.$$

The inverse transformation is given by:

$$\frac{\exp(-0.4315)}{\exp(-0.4315)+\exp(0.4369)+\exp(-0.0053)} = 0.2035,$$
$$\frac{\exp(-0.4369)}{\exp(-0.4315)+\exp(0.4369)+\exp(-0.0053)} = 0.4849,$$
$$\frac{\exp(-0.0053)}{\exp(-0.4315)+\exp(0.4369)+\exp(-0.0053)} = 0.3116.$$

The ilr transform is given by:

$$\frac{1}{\sqrt{2}}\ln\frac{20.35}{48.49} = -0.6141,$$
$$\frac{1}{\sqrt{6}}\ln\frac{20.35 \times 48.49}{31.16^2} = 0.0065.$$

The inverse transformation can be conducted as follows:

$$\sum_{j=1}^{D}\frac{y_j}{\sqrt{j(j+1)}} = -0.4315,$$
$$\sum_{j=2}^{D}\frac{y_j}{\sqrt{j(j+1)}} = 0.0027,$$
$$\sum_{j=3}^{D}\frac{y_j}{\sqrt{j(j+1)}} = 0,$$

using these values as follows:

$$-0.4315 - \sqrt{\frac{1-1}{1}} \times 0 = -0.4315,$$

$$0.0027 - \sqrt{\frac{2-1}{2}} \times -0.6141 = 0.4369,$$

$$0 - \sqrt{\frac{3-1}{3}} \times 0.0065 = -0.0053,$$

and these values are used as follows (the same as for the inverse of the clr transformation):

$$\frac{\exp(-0.4315)}{\exp(-0.4315) + \exp(0.4369) + \exp(-0.0053)} = 0.2035,$$

$$\frac{\exp(-0.4369)}{\exp(-0.4315) + \exp(0.4369) + \exp(-0.0053)} = 0.4849,$$

$$\frac{\exp(-0.0053)}{\exp(-0.4315) + \exp(0.4369) + \exp(-0.0053)} = 0.3116.$$

Correspondence: School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, Belfast, UK, e-mail: c.lloyd@qub.ac.uk (Lloyd); Departament Informàtica i Matemàtica Aplicada, Universitat de Girona, Spain, e-mail: vera.pawlowsky@udg.edu (Pawlowsky-Glahn); Departament Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Spain, e-mail: juan.jose.egozcue@upc.edu (Egozcue).