

A DATA ANALYTIC TOOL FOR MEASURING COMPOSITIONAL
VARIABILITY

By

Nedaa M. Timraz

Submitted to the

Faculty of the College of Arts and Sciences

of American University

in Partial Fulfillment of

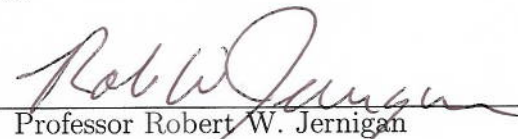
the Requirements for the Degree

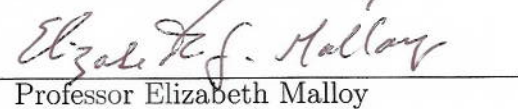
of Doctor of Philosophy

In

Statistics

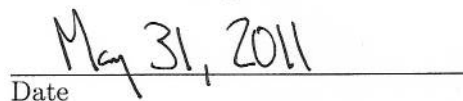
Chair:


Professor Robert W. Jernigan


Professor Elizabeth Malloy


Professor Monica Jackson


Dean of the College


Date

2011
American University
Washington, D.C. 20016

© COPYRIGHT

by

Nedaa M. Timraz

2011

ALL RIGHTS RESERVED

A DATA ANALYTIC TOOL FOR MEASURING COMPOSITIONAL VARIABILITY

by
Nedaa M. Timraz

ABSTRACT

Compositional data are non-negative proportions that sum to one. Under the unit-sum constraint, the standard statistical techniques devised for unconstrained variables can not be applied to analyze compositional data. Aitchison (1986) developed a method based on logratio transformations of compositional data that is widely used. This method is limited by the assumption of strictly positive components or the use of special treatments to accommodate possible zero components. We propose a new data analytic measure of compositional data variability based on the Sum of Coefficients of Variation to address a common objective in compositional data analysis to identify a subset of the variables that retains most of the variability of the full composition. In selecting these subcompositions, this new method resolves the difficulty of zeros in compositional data avoiding any special consideration of zeros. The new technique is investigated analytically and illustrated with real and simulated data sets.

A DATA ANALYTIC TOOL FOR MEASURING COMPOSITIONAL VARIABILITY

by
Nedaa M. Timraz

ABSTRACT

Compositional data are non-negative proportions that sum to one. Under the unit-sum constraint, the standard statistical techniques devised for unconstrained variables can not be applied to analyze compositional data. Aitchison (1986) developed a method based on logratio transformations of compositional data that is widely used. This method is limited by the assumption of strictly positive components or the use of special treatments to accommodate possible zero components. We propose a new data analytic measure of compositional data variability based on the Sum of Coefficients of Variation to address a common objective in compositional data analysis to identify a subset of the variables that retains most of the variability of the full composition. In selecting these subcompositions, this new method resolves the difficulty of zeros in compositional data avoiding any special consideration of zeros. The new technique is investigated analytically and illustrated with real and simulated data sets.

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest sense of gratitude to my supervisor Dr. Robert Jernigan for his encouragement, inspiration, invaluable suggestions, guidance, and support from the initial to the final level. Without his endless help, this work would not be possible. I would also like to extend my deep thanks to Dr. Elizabeth Malloy and Dr. Monica Jackson for their time and efforts in reviewing this work as members of my committee.

I would like to thank all the faculty members and the staff of the Department of Mathematics and Statistics at American University for their tireless efforts and for providing the excellent academic atmosphere throughout my study.

My sincere thank and appreciation go to Dr. Mary Gray who greatly supported and encouraged me throughout my graduate studies.

I take this opportunity to express my profound gratitude to my wonderful husband Wisam for helping me get through the difficult times, and for all moral support, patience, love, and understanding. I truly benefited from his experience and knowledge throughout this study.

I would like to thank my best friend Nawar, my beautiful daughters Yara, Reema and Nadeen, my sister, and my brothers for all the emotional support and the caring they provided.

Lastly, I dedicate this thesis to the souls of my mother and father

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	xii
CHAPTER	
1. INTRODUCTION	1
2. COMPOSITIONAL DATA AND THEIR ANALYSIS	6
Compositional Data : The Sample Space	6
Graphical Representation of Compositional Data	7
Compositional Covariance Structure	8
Reducing the Dimensionality of Compositional Data Sets	11
Subcompositional Analysis	12
Logcontrast Principal Component Analysis	14
Compositional Data Analysis and Zeros	15
Amalgamation	16
Zero Replacement	16
Ranking Methods for zeros in Compositional Data	17
3. MEASURING TOTAL VARIABILITY OF COMPOSITIONAL DATA SETS USING SUM OF COEFFICIENTS OF VARIATION	19

Sum of Coefficients of Variation and Subcompositional Analysis	20
Sum of Coefficients of Variation and Total Variability	21
Estimation of Sum of Coefficients of Variation	21
Estimation of Compositional Total Variability	24
Relationship between Total Variability and Sum of Coefficients of Variation	27
Relationship between Total Variability and Sum of Coefficients of Variation for smaller values of α	36
Relationship between Total Variability and Sum of Coefficients of Variation for different α s	41
Relationship between Total Variability and Sum of Coefficients of Variation for Correlated Variables	48
Illustrative Example using Real Compositional Data Set	56
Garbage Project	56
4. ZEROS IN COMPOSITIONAL DATA: A COMPARISON BETWEEN SUM OF COEFFICIENTS OF VARIATION AND COMPOSITIONAL TOTAL VARIABILITY	70
Replacing 10% of the observations in the variable Food with zeros	71
Replacing 10% of the observations in the variable Paper with zeros	81
Replacing 10% of the observations in the variable Yard with zeros	85
Replacing 20%, 30%, or 40% of the observations with zeros	89
Replacing 20%, 30%, or 40% of the observations in the variable Food with zeros	89
Replacing 20%, 30%, or 40% of the observations in the variable Paper with zeros	97

Replacing 20%, 30%, or 40% of the observations in the variable Yard with zeros	97
5. REAL COMPOSITIONAL DATA WITH ZERO OBSERVATIONS . .	99
Glacial Data Set	99
Archaeological Glass	109
6. CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH . .	115
Conclusion	115
Future Research	116

APPENDIX

A. GARBAGE COMPOSITIONAL DATA	118
B. S-PLUS PROGRAMS	120
Main Functions	120
Estimates of Sum of Coefficients of Variation and Total Variability	122
Relationship between Total Variability and Sum of Coefficients of Variation	124
Relationship between Total Variability and Sum of Coefficients of Variation for different α s	126
Relationship between Total Variability and Sum of Coefficients of Variation for correlated Variables	128
Correlation between Sum of Coefficients of Variation and Sub- compositional Total Variability using Garbage compositional data	130

LIST OF TABLES

Table	Page
1. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$	30
2. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 5-part simulated compositional dataset of size $n=100$	30
3. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 7-part simulated compositional dataset of size $n=100$	31
4. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=50$	32
5. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=30$	33
6. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 5$	37
7. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 2$	38
8. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 1$	38

9.	Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 0.5$	38
10.	Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 0.3$	38
11.	Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 0.1$	39
12.	Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha_1 = 10, \alpha_2 = 20$ and $\alpha_3 = 30$	42
13.	Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha_1 = 10, \alpha_2 = 50$ and $\alpha_3 = 100$	44
14.	Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha_1 = 1, \alpha_2 = 50$ and $\alpha_3 = 100$	45
15.	Summary Statistics of Garbage Compositional Data	56
16.	Garbage Compositional Data: All 3-part subcompositions and the corresponding SCV, Total Variability and R^2	65
17.	Garbage Compositional Data: All 4-part subcompositions and the corresponding SCV, Total Variability and R^2	67
18.	Garbage Compositional Data: All 5-part subcompositions and the corresponding SCV, Total Variability and R^2	69
19.	Garbage Compositional Data: All 3-part subcompositions and the corresponding Sum of Coefficients of Variation and Total Variability after replacing 10% of the observations in the variable Food with zeros.	76
20.	Garbage Compositional Data: All 3-part subcompositions and the corresponding Sum of Coefficients of Variation and total variability after replacing 10% of the observations in the variable Paper with zeros.	84
21.	Garbage Compositional Data: All 3-part subcompositions and the corresponding Sum of Coefficients of Variation and total variability after replacing 10% of the observations in the variable Yard with zeros.	88

22.	Number of 3-part subcompositions of the top five with largest Sum of Coefficients of Variation and largest Total Variability that include the variable Food at different percentages of zeros in the variable Food. .	98
23.	Number of 3-part subcompositions of the top five with largest Sum of Coefficients of Variation and largest Total Variability that include the variable Paper at different percentages of zeros in the variable Paper.	98
24.	Number of 3-part subcompositions of the top five with largest Sum of Coefficients of Variation and largest Total Variability that include the variable Yard at different percentages of zeros in the variable Yard. .	98
25.	Summary Statistics of the Glacial Compositional Data	100
26.	Glacial Compositional Data: 2-part subcompositions and the corresponding Total Variability	101
27.	Glacial Compositional Data: 3-part subcompositions and the corresponding Total Variability	101
28.	Glacial Compositional Data: 2-part subcompositions and the corresponding Sum of Coefficients of Variation obtained using the original and the replaced data sets	102
29.	Glacial Compositional Data: 3-part subcompositions and the corresponding Sum of Coefficients of Variation obtained using the original and the replaced data sets	103
30.	Summary Statistics of the Archaeological Glass Compositional Data .	110
31.	Top 20 4-part subcompositions with largest Total Variability computed after employing Aitchison Additive zero replacement strategy (AA) with two different values $r_1 = 0.0000076$ and $r_2 = 0.0001$	111

LIST OF FIGURES

Figure	Page
1. Graphical representation of a 3-part composition (x_1, x_2, x_3) in the reference triangle 123.	8
2. Graphical representation of a 3-part subcomposition for 25 hongite specimens	9
3. Sum of Coefficients of Variation computed using the standard formula for 3-part compositional data (D=3) simulated from Gamma and Sum of Coefficients of variation using equation (3.5)	23
4. Sum of Coefficients of Variation computed using the standard formula for 3-part compositional data (D=3) simulated from Gamma and Sum of Coefficients of variation using equation (3.5) against the corresponding values of α	24
5. Total Variability computed using Aitchison logratio transformation for 3-part compositional data (D=3) simulated from Gamma and Total Variability using Trigamma Function	28
6. Total Variability computed using Aitchison logratio transformation for 3-part compositional data (D=3) simulated from Gamma and Total Variability using Trigamma Function against the corresponding values of α	29
7. Triangle plot of 3-part compositional data set simulated from Gamma distribution with $\alpha = 10$	31
8. Aitchison's Total Variability and derived Total Variability using Trigamma function for 3-part Simulated Data	32
9. Derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 3-part Simulated Data	33

10. Total Variability using Trigamma function with Aitchison's Total Variability and Sum of Coefficients of Variation for 5-part and 7-part Simulated Data respectively	34
11. Total Variability using Trigamma function with Aitchison's Total Variability and Sum of Coefficients of Variation for 3-part Simulated Data sets with $n=50$ and $n=30$ respectively	35
12. Triangle plots of 3-part compositional data sets simulated from Gamma distributions with $\alpha = 5$, $\alpha = 2$, $\alpha = 1$, $\alpha = 0.5$, $\alpha = 0.3$, and $\alpha = 0.1$	37
13. Aitchison's Total Variability, derived Total Variability using Trigamma function, and Sum of Coefficients of Variation for 3-part Simulated Data sets for $\alpha = 5$, $\alpha = 2$ and $\alpha = 1$	39
14. Aitchison's Total Variability, Total Variability using Trigamma as a function of SCV and Sum of Coefficients of Variation for 3-part Simulated Data sets with $\alpha=0.5$, 0.3 and 0.1	40
15. Triangle plot of 3-part compositional data set simulated from gamma distribution with $\alpha_1 = 10$, $\alpha_2 = 20$ and $\alpha_3 = 30$	43
16. Aitchison's Total Variability and derived Total Variability using Trigamma function for 3-part Simulated Data with $\alpha_1 = 10$, $\alpha_2 = 20$ and $\alpha_3 = 30$	44
17. Derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 3-part Simulated Data with $\alpha_1 = 10$, $\alpha_2 = 20$ and $\alpha_3 = 30$	45
18. Plots of 3-part compositional data set simulated from gamma distribution with $\alpha_1 = 10$, $\alpha_2 = 50$ and $\alpha_3 = 100$	46
19. Plots of 3-part compositional data set simulated from gamma distribution with $\alpha_1 = 1$, $\alpha_2 = 50$ and $\alpha_3 = 100$	47
20. Triangle plot for 3-part compositional data set simulated using correlated Gammas with $\alpha = 5$	49
21. Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 3-part compositional data sets simulated using correlated Gammas with $\alpha = 5$	50
22. Triangle plot for three variables from 5-part compositional data set simulated using correlated Gammas with $\alpha = 10$	52

23.	Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 5-part compositional data sets simulated using correlated Gammas with $\alpha = 10$	53
24.	Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 4-part compositional data sets simulated using correlated Gammas with $\alpha = 10$	54
25.	Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 5-part compositional data sets simulated using Additive Logistic Normal	55
26.	SCV and compositional Total Variability of 3-part subcompositions of the Garbage data. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	58
27.	Distribution of the SCV and Total Variability of all 3-part subcompositions that contain each Garbage component	59
28.	SCV and compositional Total Variability of 4-part subcompositions of the Garbage data. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	60
29.	Distribution of the SCV and Total Variability of all 4-part subcompositions that contain each Garbage component	61
30.	SCV and compositional Total Variability of 5-part subcompositions of the Garbage data. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	62
31.	Distribution of the SCV and Total Variability of all 5-part subcompositions that contain each Garbage component	63
32.	Changes in Total Variability after replacing 10% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	77
33.	Changes in Total Variability after replacing 10% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	78
34.	Distribution of the SCV and Total Variability of all 3-part subcompositions that contain each Garbage component after replacing 10% of the observations in the variable Food with zeros.	79

35.	Changes in the Sum of Coefficients of Variation after replacing 10% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	80
36.	Number of 3-part subcompositions remain in the top five after replacing 10%, 20%, 30%, and 40% of the observations in the variable Food with zeros	90
37.	Changes in Total Variability after replacing 20% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	91
38.	Changes in Total Variability after replacing 20% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	92
39.	Changes in Total Variability after replacing 30% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	93
40.	Changes in Total Variability after replacing 30% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	94
41.	Changes in the Sum of Coefficients of Variation after replacing 20% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	95
42.	Changes in the Sum of Coefficients of Variation after replacing 30% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other	96
43.	Plot of Total Variability for all 2-part subcompositions obtained after employing Aitchison Additive Replacement Strategy with $\delta_1 = 0.001$ and $\delta_2 = 0.0005$	102
44.	Plot of Total Variability for all 3-part subcompositions obtained after employing Aitchison Additive Replacement Strategy with $\delta_1 = 0.001$ and $\delta_2 = 0.0005$	103
45.	Plot of Total Variability for all 2-part subcompositions obtained after employing Multiplicative Replacement Strategy with $r_1 = 0.001$ and $r_2 = 0.0005$	104
46.	Plot of Total Variability for all 3-part subcompositions obtained after employing Multiplicative Replacement Strategy with $r_1 = 0.001$ and $r_2 = 0.0005$	105

47.	Plot of Sum of Coefficients of Variation for all 2-part subcompositions obtained using the original and the replaced data sets using Aitchison Additive Replacement Strategy with $\delta = 0.001$	106
48.	Plot of Sum of Coefficients of Variation for all 2-part subcompositions obtained using the original data and the replaced data sets using Multiplicative Replacement Strategy with $r = 0.001$	107
49.	Plot of the Sum of Coefficients of Variation for all 2-part subcompositions obtained using the replaced data sets using Aitchison Additive Replacement Strategy with $\delta_1 = 0.001$ and $\delta_2 = 0.0005$	108
50.	Plot of the top 20 4-part subcompositions with largest Total Variability obtained after employing Aitchison Additive Replacement Strategy with $r_1 = 0.0000076$ and the Total Variability for the same subcompositions after using $r_2 = 0.0001$	112
51.	Plot of the top 20 4-part subcompositions with largest Total Variability obtained after employing Multiplicative Replacement Strategy with $r_1 = 0.000055$ and Total Variability for the same subcompositions after using $r_2 = 0.0001$	113
52.	Plot of the top 20 4-part subcompositions with largest Sum of Coefficients of Variation obtained using the original data and Sum of Coefficients of Variation for the same subcompositions in the replaced data sets using Aitchison Additive Replacement Strategy with $r_1 = 0.0000076$ and $r_2 = 0.0001$ and Multiplicative Replacement Strategy with $r_1 = 0.000055$ and $r_2 = 0.0001$	114

CHAPTER 1

INTRODUCTION

Many multivariate data sets of interest are compositional, consisting essentially of relative proportions summing to one. For instant, in finance an important aspect of the study of consumer demand is the analysis of household budgets in which attention often focuses on the expenditures of a sample of households on a number of mutually exclusive and exhaustive commodity groups (housing, foodstuffs, other goods, and services) and their relations to total expenditure, income, type of housing, household composition, and so on (Aitchison 1986). In geology the composition of rocks is often studied by classifying each rock according to the relative percentage by weight of chemical oxides (Aitchison 1986). In geochemistry, compositions can be expressed as molar concentrations of each component (Valls 2008). In ecology, ecologists often choose to use proportional or percentage type data to study the relative representation of a species in a particular ecosystem (Jackson 1997).

Definition 1. *Compositional data are vectors of proportions describing the relative contributions of each of D categories to the whole. Mathematically, compositional data with n observations of a D -part composition are of the form*

$$x_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, D,$$

where

$$0 \leq x_{ij} \leq 1 \quad \forall i, j$$

and

$$\sum_{j=1}^D x_{ij} = 1 \quad \forall i$$

are the constraints induced by being a composition.

Compositional data have particular and important numerical properties that have major consequences for any statistical analysis. These have been discussed by a number of authors since Karl Pearson (1897) first highlighted problems in the analysis of compositions (Sarmanov and Vistelius 1959 ; Chayes 1971 ; Butler 1979; Aitchison 1986 ; Davis 1986 ; and Rock 1988). The properties peculiar to compositional data arise from the fact that they represent parts of some whole; therefore, they convey only relative information. They are always positive and usually constrained to a constant sum. The constant sum constraint in definition (1) forces at least one covariance to be negative. Hence correlations are not free to range over the usual interval $(-1, +1)$. Thus, spurious correlations are induced by the fact that the data sum to a constant and there is bias towards negative correlation. Consider the trivial case of a two-part composition summing to a constant: the correlation between the two elements in this composition must be -1.

The essential consequence of these properties is that standard statistical techniques devised for unconstrained random variables, cannot be used to analyze compositional data. The summation constraint and bounded support require special techniques for compositional data. Aitchison (1986) introduced a range of statistical techniques to handle the special problems and questions of inference in analyzing compositional data. These techniques are based on logratio transformed data, recognizing that it is the relative magnitudes and variations of components, rather

than their absolute values, that provide the key to analyzing compositional data. The inference tools developed for multivariate normal data are often applied to the transformed compositions.

Aitchison considered the additional restrictions that $x_{ij} > 0, \forall i, j$, then used the logratio transformation to remove the constraints. This involves choosing one component as a divisor and looking at

$$y_i = \log(x_i/x_D), \quad i = 1, \dots, D - 1.$$

Clearly, the constraint is removed and y can take any real value. Aitchison (1986) suggested that many problems can be analyzed under the assumption that

$$\mathbf{y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

or, equivalently, the D -part composition \mathbf{x} follows an Additive Logistic Normal Distribution.

Although Aitchison's method of logratio transformation of compositional data is widely used in various settings, it suffers from the following limitations:

1. Interpretation of parameter estimates on the multivariate log-odds scale is difficult, specifically, the location parameters $\mu_i = E(\log(x_i/x_D))$ and elements of the covariance matrix, $\sigma_{ij} = \text{cov}(\log(x_i/x_D), \log(x_j/x_D))$. It is often challenging to interpret these parameters (or their estimates) in terms of a motivating scientific problem (Billheimer, Guttorp, and Fagan (1998)).

2. Logratio analysis and normality do not always model data adequately, thus leading to a need for alternative transformations, such as the Box-Cox family (Carles Barceló, Vera Pawlowsky, and Eric Grunsky (1996)).

3. Some common forms of logratio analysis will sometimes produce results with no substantive meaning, even when substantively interpretable structure exists. Specifically, the high relative variation of some variables emphasized in a logratio analysis may derive, in part, from their low absolute levels, and the variation may be of limited practical interest (Baxter, Beardah, Cool, and Jackson (2005), Baxter, Cool, and Jackson (2005), and Hijazi and Jernigan (2009)). For example, Beardah and others (2003) suggested that for compositional data for glass production, bivariate analysis and the crude principal component analysis of standardized data often produced more interpretable results than principal component analysis of logratio transformed data. The reason appears to be that log-ratio analysis emphasizes those variables with a high relative variation and in glass compositional data sets such variables often have a low absolute presence and variation. Baxter and others (2005) and Baxter and Freestone (2006) described using both simulated and real data where crude principal component analysis produces archaeologically interpretable results much more readily than logcontrast principal component analysis

4. Logratio analysis does not produce good predictions for edge cases when some proportions are close to zero. As x_i approach zero, logratios approach negative or positive infinity.

5. Logratio transformations of compositional data are limited by the assumption of strictly positive components and require special treatments for zero components. Aitchison (1986) devised several special treatments to handle zero components of compositional data, and lamented that the problem of zeros is unlikely ever to be satisfactory and generally resolved.

6. The major remaining disadvantage of the logratio models is the complexity of their structure.

We propose a new data analytic tool for measuring compositional data variability that does not involve the use of logratio transformations introduced by Aitchison. The approach is based on the use of the Sum of Coefficients of Variation (SCV) of the components of the compositional data set. Coefficients of Variation are calculated for each component and the sum of these coefficients is computed. The approach is simple, based on a well-known measure of variation, non-parametric, and identifies a set of subcompositions that retains as much of the variability in the full composition as possible. Further, the approach doesn't require any special treatment of zeros and allows much more useful information to be extracted from compositions, in situations where zeros may contain potentially important information.

This study is organized as follows. Chapter 2 introduces compositional data and their analysis. Chapter 3 reviews the two approaches to measuring compositional data variability, Aitchison's approach based on the logratio transformations and the Sum of Coefficients of Variation approach. For special cases we examine distributional properties of the two measures and the theoretical relationship between them as well as applications using simulated and real data sets. In Chapter 4 we compare the Sum of Coefficients of Variation approach and Aitchison's approach in the presence of zeros using a real data set. Aitchison's approach based on the logratio transformations is considered after employing different existing zero treatment techniques. We examined the changes in the two approaches with different percentages of zeros and using different variables in the data. In Chapter 5 we evaluate the performance of the new method based on the Sum of Coefficients of Variation and Aitchison method using two real compositional data sets with zero observations. Finally, conclusions and further research are presented in Chapter 6. We discuss possible ways for improving the Sum of Coefficients of Variation technique. Research directions and possible extensions are given.

CHAPTER 2

COMPOSITIONAL DATA AND THEIR ANALYSIS

Compositional Data : The Sample Space

Compositional data are non-negative proportions summing to one. As such compositional data occupy a restricted space where variables can vary only from 0 to 1. A composition of D proportions is completely specified by the components of a d -part subvector (x_1, \dots, x_d) , where $d = D - 1$, and the remaining component has the value

$$x_D = 1 - x_1 - \dots - x_d$$

This means that a D -part composition is essentially a d -dimensional vector and so can be represented in some convenient d -dimensional set. This restricted space is known as a simplex.

Definition 2. *The d -dimensional simplex is the set defined by $S^d = \{(x_1, \dots, x_d) : x_1 > 0, \dots, x_d > 0; x_1 + \dots + x_d < 1\}$.*

Definition 3. *The d -dimensional simplex embedded in D -dimensional real space is the set defined by $S^d = \{(x_1, \dots, x_D) : x_1 > 0, \dots, x_D > 0; x_1 + \dots + x_D = 1\}$.*

Graphical Representation of Compositional Data

A convenient way of displaying the variability of 3-part compositions is with what is variously termed a *ternary diagram*, a *reference triangle*, or *barycentric coordinate space*. The diagram relies on the geometry of an equilateral triangle to plot each component's proportion of the total composition. Plotting several compositions on the same diagram makes it simple to compare them to each other (Aitchison 1986).

The triangle of Figure (1) with vertices 1, 2, and 3 is equilateral and has unit altitude. For any point P in triangle 123 the perpendiculars x_1, x_2, x_3 from P to the sides opposite 1, 2, 3 satisfy

$$x_i \geq 0 \quad (i = 1, 2, 3), \quad x_1 + x_2 + x_3 = 1. \quad (2.1)$$

Moreover, corresponding to any vector (x_1, x_2, x_3) satisfying (2.1), there is a unique point in triangle 123 with perpendicular lengths x_1, x_2 , and x_3 . There is therefore a one-to-one correspondence between 3-part compositions and points in triangle 123, and so we have a simple means of representing 3-part compositions. In such a representation we may note that the three inequalities in (2.1) are strict if and only if the representative point lies in the interior of triangle 123. The percentage of component 1 could range from 0, if the point is on the base of the triangle, to 100 if the point is at vertex 1.

Although such 3-dimensional displays have a useful expository role in describing the structure of a problem, they have a limited part to play in data analysis, since it is much more difficult to extend this form to a composition that has more than four parts.

Example 1. *The triangle of Figure (2) with vertices A, B, and C is for a 3-part subcomposition of a mineral composition of 25 rock specimens of the type hongite.*

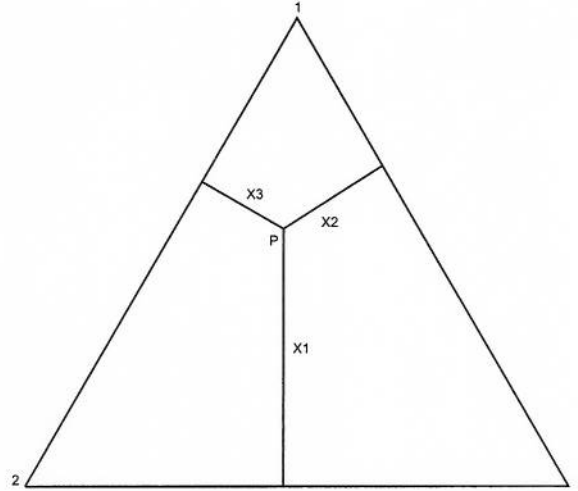


Figure 1. Graphical representation of a 3-part composition (x_1, x_2, x_3) in the reference triangle 123.

The full data set is given in Aitchison (1986) . Each composition consists of the percentages by weight of five minerals, Albite, Blandite, Cornite, Daubite, and Endite. As we can see from the graph there is an extensive and widely scattered variation in the ratio of B to C compare to the other two ratios. This nonlinear "curvature" pattern is common in compositional data.

Compositional Covariance Structure

Aitchison (1986) argues that a composition \mathbf{x} can be completely determined by d ratios such as $x_i/x_D (i = 1, \dots, d)$. This realization that the study of compositions

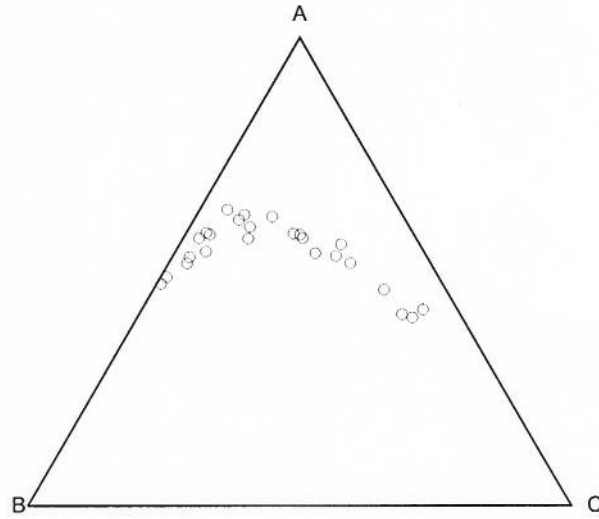


Figure 2. Graphical representation of a 3-part subcomposition for 25 hongite specimens

is essentially concerned with the relative values of the compositional parts and not the absolute values of the individual parts leads to logratios analysis. Aitchison (1986) adopts a new concept of correlation based on covariances of the form

$$\sigma_{ij.kl} = \text{cov}(\log(x_i/x_k), \log(x_j/x_l)).$$

and logratio means

$$\xi_{ij} = E(\log(x_i/x_j)) \quad (i = 1, \dots, d; j = i + 1, \dots, D).$$

Definition 4. *The covariance structure of a D -part composition \mathbf{x} is the set of all*

$$\sigma_{ij.kl} = \text{cov}(\log(x_i/x_k), \log(x_j/x_l))$$

as i, j, k, l run through the values $1, \dots, D$.

Using this definition, there are $\frac{1}{2}dD$ covariances to be specified for which all the others can then be determined. Aitchison (1986) introduced three ways in which this general compositional covariance structure can be specified. These ways are summarized in the following definitions:

Definition 5. *For a D -part composition \mathbf{x} the $D \times D$ matrix*

$$\mathbf{T} = [\tau_{ij}] = [\text{var}\{\log(x_i/x_j)\} : i, j = 1, \dots, D]$$

is termed the variation matrix and determines the covariance structure by the relationships

$$\sigma_{ij.kl} = \frac{1}{2}(\tau_{il} + \tau_{jk} - \tau_{ij} - \tau_{kl}).$$

τ_{ij} in definition (5) satisfies:

$$\tau_{ii} = 0 \quad (i = 1, \dots, D),$$

$$\tau_{ij} = \tau_{ji} \quad (i = 1, \dots, d; j = i + 1, \dots, D),$$

and so are determined by the $\frac{1}{2}dD$ values $\tau_{ij} \quad (i = 1, \dots, d; j = i + 1, \dots, D)$.

Definition 6. *For a D -part composition \mathbf{x} the $d \times d$ matrix*

$$\mathbf{\Sigma} = [\sigma_{ij}] = [\text{cov}\{\log(x_i/x_D), \log(x_j/x_D)\} : i, j = 1, \dots, d]$$

is termed the logratio covariance matrix and determines the covariance structure by the relationships

$$\sigma_{ij.kl} = \sigma_{ij} + \sigma_{kl} - \sigma_{il} - \sigma_{jk}.$$

Definition 7. For a D -part composition \mathbf{x} the $D \times D$ covariance matrix

$$\Gamma = [\gamma_{ij}] = [\text{cov}[\log(x_i/g(x)), \log(x_j/g(x))] : i, j = 1, \dots, D]$$

where $g(x)$ is the geometric mean $(x_1 \dots x_D)^{\frac{1}{D}}$, is termed the centered logratio covariance matrix and determines the covariance structure by the relationships

$$\sigma_{ij.kl} = \gamma_{ij} + \gamma_{kl} - \gamma_{il} - \gamma_{jk}.$$

However, Aitchison (1986) describes that the analysis of any particular problem may be simpler in terms of one of the specifications than the others and that each specification has some apparent disadvantage relative to the others: \mathbf{T} is not a covariance matrix, Σ is not symmetric in the compositional parts, Γ is singular; there is no specification which is free from all of these three disadvantageous features.

Finally, Aitchison (1986) introduces a measure of Total Variability of a composition \mathbf{x} as :

$$Totvar(\mathbf{x}) = tr(\Gamma) = \frac{1}{D} \sum_{i < j} \tau_{ij} = \frac{1}{D} \sum_{i < j} var(\log \frac{x_i}{x_j}). \quad (2.2)$$

Reducing the Dimensionality of Compositional Data Sets

Aitchison (1984) considers two dimension-reducing procedures for compositional data, subcompositional analysis and Logcontrast Principal Components.

Subcompositional Analysis

Data reduction is most easily accomplished by considering smaller decompositions. For example, in a household expenditure enquiry we may start with a budget-share composition (x_1, \dots, x_9) of proportions of total expenditure spent on nine commodity groups: foodstuffs, housing, fuel and light, tobacco and alcohol, clothing and footwear, durable goods, miscellaneous goods, transport and vehicles, and services. We may wish to identify subcompositions that retain as much of the total variability in the entire composition as possible. For example, we may ask the extent to which the subcompositions based on the components (foodstuffs, housing, and fuel and light) and (foodstuffs, clothing and footwear, and services) retain the variability of the complete nine compositions. We can form a subcomposition simply by rescaling the original proportions of these selected groups so that the scaled proportions sum to 1.

The purpose of the analysis is to identify subcompositions that retain as much of the total variability in the entire composition as possible.

If (x_1, \dots, x_D) is a complete D -part composition with the constraint

$$x_1 + \dots + x_D = 1$$

then any subvector with its elements rescaled so that their sum is 1 is a subcomposition. For example, the subvector (x_1, \dots, x_C) gives a subcomposition

$$\mathbf{X} = (X_1, \dots, X_C) = x_i / (x_1 + \dots + x_C) \quad (i = 1, \dots, C). \quad (2.3)$$

Such a subcomposition is technically a composition with dimension $c = C - 1$, smaller than the dimension d of the original composition. It can be completely represented by the logratio vector $\mathbf{Y} = \log(\mathbf{X}_{-C}/X_C)$. In any subcomposition, the ratio of any

pair of components is identical to their ratio in the full composition. It follows that the logratio means ξ_{ij} and variances τ_{ij} are the same for the subcomposition as for the composition. Moreover, we can obtain the variation array for any subcomposition by selection of the appropriate entries from the full array (Aitchison (1986)). With such a covariance structure any relevant information which we have about a subcomposition provides direct and exact information about the full composition. Additionally, Aitchison (1986) explains that the covariance structure of a composition is completely determined by knowledge of the covariance structure of the logratio variances of all of its 2-part subcompositions.

Definition 8. *A selection matrix \mathbf{S} is any matrix of order $C \times D$ ($C < D$), with C elements equal to 1, one in each row and at most one in each column, and the remaining $C(D - 1)$ elements 0.*

The measure of variability of the subcomposition \mathbf{X} is simply the Total Variability of \mathbf{X} regarded as C -part composition or of \mathbf{Y} regarded as a c -dimensional logratio vector. For a particular subcomposition this is

$$tr(\mathbf{\Gamma}_S) = tr(\mathbf{G}_C \mathbf{S} \mathbf{\Gamma} \mathbf{S}' \mathbf{G}_C) \quad (2.4)$$

where \mathbf{S} is $C \times D$ selection matrix, $\mathbf{\Gamma}$ is the centered logratio covariance matrix of the full composition, and $\mathbf{G}_C = \mathbf{I}_C - \mathbf{C}^{-1} \mathbf{J}_C$ with \mathbf{J}_C the $C \times C$ matrix of units and \mathbf{I}_C identity matrix. Hence the proportion of the Total Variability of a composition retained by a subcomposition with selection matrix \mathbf{S} is

$$tr(\mathbf{\Gamma}_S)/tr(\mathbf{\Gamma}). \quad (2.5)$$

and Aitchison finds a much simpler computation form which involves only simple

summation of subsets of the variation matrix \mathbf{T} and avoids the complicated construction of $\Gamma_{\mathbf{S}}$.

$$tr(\Gamma_{\mathbf{S}}) = \frac{1}{C} \sum_{i < j} \tau_{ij} \quad i, j = 1, \dots, C$$

$$tr(\Gamma) = \frac{1}{D} \sum_{i < j} \tau_{ij} \quad i, j = 1, \dots, D$$

$$tr(\Gamma_{\mathbf{S}})/tr(\Gamma) = (\mathbf{D}\mathbf{j}'_{\mathbf{C}}\mathbf{S}\mathbf{T}\mathbf{S}'\mathbf{j}_{\mathbf{C}})/(\mathbf{C}\mathbf{j}'_{\mathbf{D}}\mathbf{T}\mathbf{j}_{\mathbf{D}}). \quad (2.6)$$

where \mathbf{j} is a column vector of units. This is simply the average of all entries in the variation matrix \mathbf{T} that contribute to the C -part subcomposition divided by the average of all the entries in \mathbf{T} , this is

$$\frac{D \sum_{i=1}^C \tau_{ij}}{C \sum_{i=1}^D \tau_{ij}}$$

Logcontrast Principal Component Analysis

A common technique for reducing the dimensionality in multivariate studies is principal component analysis. Aitchison (1983) describes that a principal component analysis which regards a data set in S^d as embedded in R^d or R^{d+1} , and finds lines and hyperplanes of closest fit using Euclidean distance and orthogonality may seem geometrically attractive but suffers from the curvature problem resulting from nonlinear patterns and variation about curved lines in the compositional data set, see Figure (2) for an example (where the crude principal component analysis would result in a linear reduction technique with straight line principal axes). Like so many other statistical procedures for compositional data, principal component analysis is subject to all the difficulties of interpretation associated with the use of crude covariance structures. Moreover, such a procedure is mathematically linear in nature and

so cannot hope to capture patterns of curved variability which are commonly present in many compositional data sets. Aitchison (1983) has reviewed the unsatisfactory nature of the versions of principal component analysis and proposed a different approach which has the capability of capturing the structure of the variability of the data set. His method, referred to as logcontrast principal component, is to study the dependence structure of compositions through the covariance matrix of logarithms of the ratios of components. For a composition \mathbf{x} in S^d , the counterpart of a linear combination in R^d is a linear combination of the logratio vector \mathbf{y} or equivalently a logcontrast $\mathbf{a}'\log\mathbf{x}$ where $\mathbf{a}'\mathbf{j} = 0$.

Although the loglinear-contrast approach to principal-component and sub-compositional analysis has undoubtedly attractive features and appears successful in application, it is far from being a complete answer to dimension-reducing problems in compositional data analysis Aitchison (1984) . One draw back is that we cannot take logarithms of zero proportions. Moreover, the logcontrast approach is not a general solution for all curved data sets. While it may succeed in straightening out curved sets where the linear approach fails, it can prove just as inadequate.

Compositional Data Analysis and Zeros

As we mentioned earlier, one of the troublesome problems in compositional data analysis using the logratio models is the presence of zeros in the data. In the statistical literature, two explanations for the occurrence of zero observations are proposed. These are rounding (or trace elements) and essential (or true) zeros. The first explanation rationalizes that zero observation is an artifact of the measurement process. Basically, if we had a more accurate measurement instrument we would record a non-zero observation. Thus the observed zero is a proxy for a very small number. The second rationalization argues that the observation should be zero as the true generating process leads to the occurrence of zeros. The proposed modifications

to deal with zero observations can then be derived by considering the causes of the zero (Fry and others 1996). The following possible strategies have been investigated.

Amalgamation

Amalgamation is the reduction of the number of components in the composition by the grouping together of certain components. This is a simple approach and it avoids the potential problems of the other options. In particular, spurious clusters associated with any replaced zeros may occur (Fry and others 1996). However, this is not an appropriate technique to deal with the zero observations when the omitted variables are important for the analysis.

Zero Replacement

The zero replacement techniques assumes that a composition has C zero and $D - C$ non-zero components. It is recommended that the zeros to be replaced by small values. The possible strategies for the replacements are:

1. The additive replacement strategy of Aitchison (1986) (AA) suggested a procedure which replaces any composition with C zero and $D - C$ non-zero components by another composition in which the zeros become $r = \delta(C + 1)(D - C)/D^2$ and the positive components are reduced by $\delta C(C + 1)/D^2$, where δ is the maximum rounding-off error.
2. The alternative zero replacement (AZR) procedure modified Aitchison of Fry and others (1996) suggested that we replace the zeros by $r = \delta(C + 1)(D - C)/D^2$, but to reduce each non-zero by $\omega_i \times \delta C(C + 1)/D^2$, where ω_i is the share ratio of the component i . This both retains the share ratios for the non-zero components and makes an appropriate zero replacement. Fry and others (1996) suggested that we can get a sensible minimum value of r_i by dividing the minimum possible value

any observation can be by the maximum value in the data.

3. Replace zero values with some value r , less than 0.01%, and recalculate the variable "Other" by differencing (RZRO). Example of this approach is presented in Beardah, Baxter, Cool, and Jackson (2003) .

4. Replace zeros with some small value, r , and other elements by $(x_{ij} - rx_{ij}/100)$. This is the multiplicative replacement strategy (MR), proposed by Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2003). It is a particular case of what they call the simple replacement strategy, in which zeros are replaced with a small constant and then all elements rescaled so that the sum is 1.

The difficulty in general with zero replacement approach is to decide how much to add while retaining as much of the original structure in the data as possible. Moreover, the constant-sum-constraint of compositional data forces modification of the zero and the non-zero values and the imputed value depends not only on δ but also on the dimension D and the number C of zeros. Note also that a different δ_i could be considered for every component x_i leading to a slightly more complicated expression. Finally, Tauber (1999) illustrated that Aitchison's distance between two replaced observations defined by:

$$\Delta(X, x) = \sqrt{\sum_{i=1}^D \left(\log \frac{x_i}{g(x)} - \log \frac{X_i}{g(X)} \right)^2},$$

where $g(\cdot)$ is the geometric mean of the composition, is extremely sensitive to the change in δ .

Ranking Methods for zeros in Compositional Data

When ranking is applied to multivariate data, it is usually applied separately to each variable, across cases. This avoids the problem that the variables may be

of very different scales. It would, however, be possible to rank each case across variables, or even to rank across all variables and cases. The type of ranking used will depend on the analysis to be undertaken. Bacon-Shone (1991) introduced ranking methods of handling zeros in compositional data analysis. In his method, Bacon-Shone suggested that ranking across both cases and components retains most of the useful information in a robust way that does not require optimizing over a parameter as in the zero replacement methods and thus seems a potentially useful alternative for handling zeros in compositional data. He also suggested that the greater the number of variables and cases, the more effective ranking will be in retaining the structure of the data. Furthermore, a weakness of ranking is that the transform will depend on the set of components chosen. This suggests ranking data at the highest level of disaggregation, and doing any aggregation after the rank transformation.

CHAPTER 3

MEASURING TOTAL VARIABILITY OF COMPOSITIONAL DATA SETS USING SUM OF COEFFICIENTS OF VARIATION

Coefficient of Variation (CV) is a normalized measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation to the mean:

$$CV = \text{Standard Deviation} / \text{Mean}$$

This is only defined for non-zero mean, and is most useful for variables that are always non-negative and for data measured on a ratio scale. Coefficient of variation is useful because the standard deviation of the data must always be understood in the context of the mean of the data. It is a dimensionless number and when comparing data sets with different units or wildly different means, it can be more informative than the standard deviation. *CV* has been found useful in many scientific areas. Reed et al. (2002) developed a simple procedure to determine the probability that an assay will accurately discern whether two samples have the same analytic concentration or not based on a knowledge of the assay variability as measured by the Coefficient of Variation. The Coefficient of Variation has been used by organizational researchers to index and compare the internal variability of top management teams, task groups, boards of directors, departments, and other social aggregates on numerous dimensions (Bedeian and Mossholder 2000). The *CV* has also seen some

applications in compositional data. In laboratory medicine the *CV* is widely used measure in External Quality Assessment to assess and compare the reproducibility of techniques and equipments. Zhang and others (2010) proposed a multivariate *CV* for comparing the performance of the electrophoretic techniques in External Quality Assessment based on the logratio transformed compositional electrophoretic data. Graf (2006) introduced a global *CV* to assess precision of compositional data in a stratified two-stage sample using the Swiss Earnings Structure Survey. The global *CV* he introduced is the square root of the average squared *CV* for all possible ratios of components.

Finally, archaeologists who were aware of the statistical literature on compositional data analysis expressed concern about the unsatisfactory experience in applying the logratio transformation to pottery and glass compositional data. Baxter and others (2005) and Baxter and Freestone (2006) described using both simulated and real data where crude principal component analysis and absolute differences in composition can convey archaeological interpretable results much more readily than logcontrast principal component analysis. Baxter and Freestone (2006) noted that "One problem for logratio analysis is that results can be overly influenced by minor oxides, present at low absolute levels and not structure-carrying, that dominate the logratio analysis to no good effect because of their high relative variance."

Sum of Coefficients of Variation and Subcompositional Analysis

We discussed in Chapter 2 measures of Total Variability and dimension-reduction techniques introduced by Aitchison. His approach can prove inadequate for some data producing uninterpretable parameters and limited by the assumption of strictly positive components as well as the requirement of special treatments in practice of the zero components.

We introduce an alternative technique of measuring compositional data variability and variables selection based on the Sum of Coefficients of Variation (SCV) of the components. We simply sum up the Coefficients of Variation of the components of a C -part subcomposition. A high SCV is associated with sets of components that retain most of the variability of the full composition.

For example, if \mathbf{X}_C is the C -part subcomposition formed from the leading subvector (x_1, \dots, x_C) of the full composition (x_1, \dots, x_D) , then the Sum of Coefficients of Variation of the C -part subcomposition would be

$$SCV = CV(x_1) + \dots + CV(x_C) \quad (3.1)$$

Sum of Coefficients of Variation and Total Variability

In this section we investigate the relationship between Sum of Coefficients of Variation and Aitchison Total Variability based on logratio transformations. First we study the distributional properties for each measure and from the derived distributions we examine the relationship between them. We consider D -part compositional data generated from D independent random variables from a Gamma distribution.

Estimation of Sum of Coefficients of Variation

Let W_1, \dots, W_D be independent random variables from $Gamma(\alpha_i, \beta)$. Recall that the probability density function (pdf) of a random variable W following a Gamma distribution is defined as follows:

$$f(w; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} \exp -\frac{w}{\beta},$$

with $0 \leq w < \infty, \alpha, \beta > 0$. The mean and variance are defined by:

$$E(W) = \alpha\beta$$

$$Var(W) = \alpha\beta^2$$

The theoretical coefficient of variation under the Gamma distribution is thus given by

$$CV(W) = \frac{\sqrt{Var(W)}}{E(W)} = \frac{\sqrt{\alpha\beta^2}}{\alpha\beta} = \frac{1}{\sqrt{\alpha}} \quad (3.2)$$

Let $x_i = \frac{w_i}{\sum_{i=1}^D w_i}$, then $X = (x_1, \dots, x_D)$ has a Dirichlet distribution $D^{D-1}(\alpha_1, \dots, \alpha_D)$ with $\alpha_1, \dots, \alpha_D > 0$. The probability density function of X is given by

$$f(x_1, \dots, x_D; \alpha_1, \dots, \alpha_D) = \frac{1}{B(\alpha)} \prod_{i=1}^D x_i^{\alpha_i-1},$$

for all $x_1, \dots, x_{D-1} > 0$ satisfying $x_1 + \dots + x_{D-1} < 1$, where x_D is $1 - x_1 - \dots - x_{D-1}$. The density is zero outside the open $(D-1)$ -dimensional simplex. The normalizing constant is the multinomial Beta function, which can be expressed in terms of the Gamma function: $B(\alpha) = \frac{\prod_{i=1}^D \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^D \alpha_i)}$, $\alpha = (\alpha_1, \dots, \alpha_D)$.

Define $\alpha_+ = \sum_{i=1}^D \alpha_i$, then

$$E(x_i) = \frac{\alpha_i}{\alpha_+},$$

$$Var(x_i) = \frac{\alpha_i(\alpha_+ - \alpha_i)}{\alpha_+^2(\alpha_+ + 1)},$$

and the coefficient of variation is given by

$$CV(x_i) = \sqrt{\frac{\alpha_+ - \alpha_i}{\alpha_i(\alpha_+ + 1)}} \quad (3.3)$$

and if $\alpha_i = \alpha$

$$CV(x) = \sqrt{\frac{D-1}{D\alpha+1}} \quad (3.4)$$

and the Sum of the Coefficients of Variation

$$SCV = D\sqrt{\frac{D-1}{D\alpha+1}}. \quad (3.5)$$

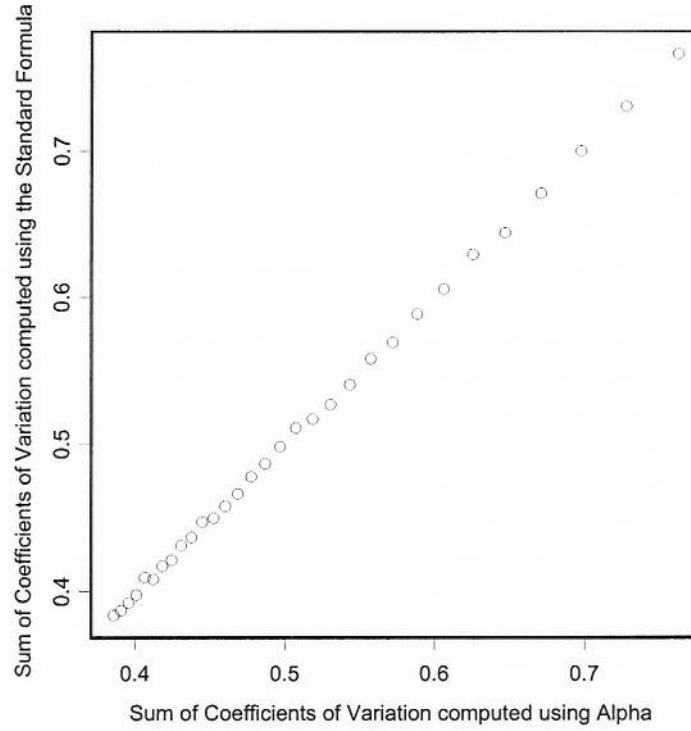


Figure 3. Sum of Coefficients of Variation computed using the standard formula for 3-part compositional data ($D=3$) simulated from Gamma and Sum of Coefficients of variation using equation (3.5)

Figure (3) shows a scatter plot of the Sum of Coefficients of Variation computed using Equation (3.5) for different values of α ($\alpha = 10, \dots, 40$) and the Sum of Coefficients of Variation computed using the standard formula $SCV = \sum_{i=1}^D CV(x_i)$, where $CV(X) = \frac{SD}{mean}$, for 3-part compositional data sets generated from Gamma distributions with 100 observations and different values of α ($\alpha = 10, \dots, 40$). Figure (4) shows the two sums against α . It can be seen clearly from the graph that the two measures are nearly the same and decreasing with α .

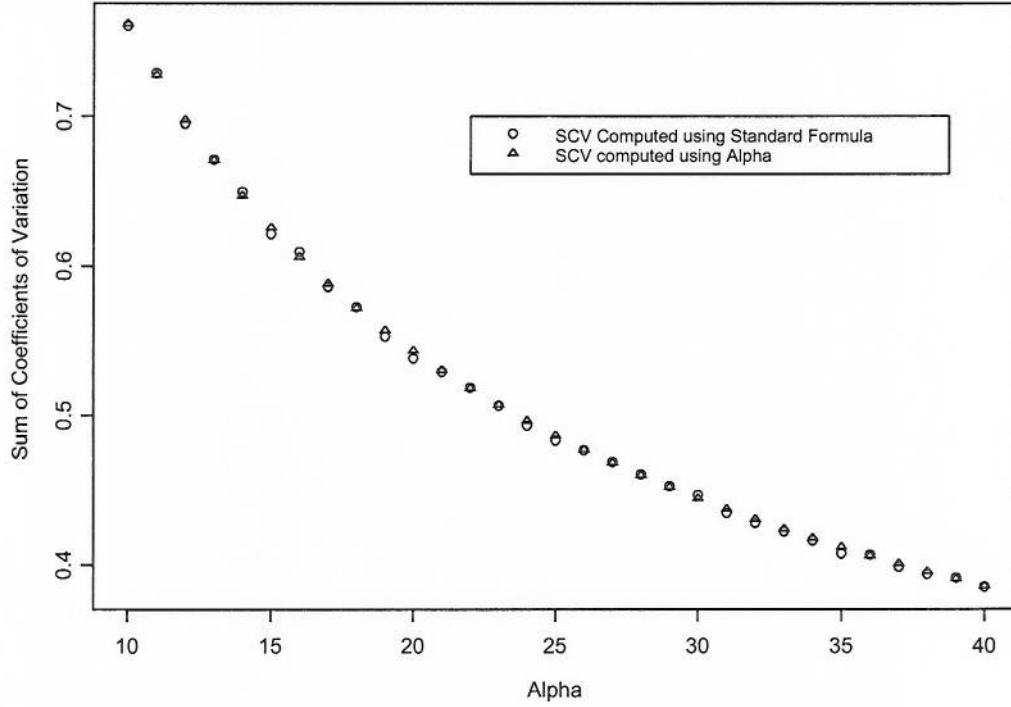


Figure 4. Sum of Coefficients of Variation computed using the standard formula for 3-part compositional data ($D=3$) simulated from Gamma and Sum of Coefficients of variation using equation (3.5) against the corresponding values of α

Estimation of Compositional Total Variability

Recall that the Total Variation of a random composition \mathbf{x} is given by

$$Totvar = \frac{1}{D} \sum_{i < j} \tau_{ij} = \frac{1}{D} \sum_{i < j} Var(\log \frac{x_i}{x_j}). \quad (3.6)$$

From the closure operation, $\mathbf{x} = (x_1, \dots, x_D) = \frac{(w_1, \dots, w_D)}{\sum_{i=1}^D w_i}$ and $W = (w_1, \dots, w_D)$ are compositionally equivalent and that

$$\tau_{ij} = Var(\log \frac{x_i}{x_j}) = Var(\log \frac{\frac{w_i}{\sum_{i=1}^D w_i}}{\frac{w_j}{\sum_{i=1}^D w_i}}) = Var(\log \frac{w_i}{w_j})$$

First we will derive the distribution of $(\frac{w_i}{w_j})$. Let $v = \frac{w_1}{w_2}$ and $u = w_1 + w_2$, then the jacobian is $|J| = \frac{u}{(1+v)^2}$,

$$\begin{aligned} g(u, v) &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \exp(-u) \left(\frac{uv}{1+v}\right)^{\alpha_1-1} \left(\frac{u}{1+v}\right)^{\alpha_2-1} \frac{u}{(1+v)^2} \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \exp(-u) u^{\alpha_1+\alpha_2-1} v^{\alpha_1-1} (1+v)^{-\alpha_1-\alpha_2} \end{aligned}$$

The ratio $\frac{w_1}{w_2}$ therefore has the distribution

$$h(v) = \int_0^\infty g(u, v) du = \frac{1}{\beta(\alpha_1, \alpha_2)} v^{\alpha_1-1} (1+v)^{-\alpha_1-\alpha_2}, v > 0, \alpha_1, \alpha_2 > 0 \quad (3.7)$$

which is a Beta Prime Distribution with parameters (α_1, α_2) (Johnson, Kotz and Balakrishnan 1995).

Now let $Z = -\log(v) = -\log(\frac{w_1}{w_2})$, then

$$v = \exp(-z) \text{ and } \left| \frac{d}{dz} \exp(-z) \right| = \exp(-z)$$

hence,

$$f(z) = \frac{1}{\beta(\alpha_1, \alpha_2)} \exp(-z(\alpha_1 - 1)) (1 + \exp(-z))^{-\alpha_1-\alpha_2} \exp(-z)$$

$$f(z) = \frac{1}{\beta(\alpha_1, \alpha_2)} \frac{\exp(-\alpha_1 z)}{(1 + \exp(-z))^{\alpha_1+\alpha_2}}, -\infty < z < \infty, \alpha_1, \alpha_2 > 0. \quad (3.8)$$

which is a Generalized Logistic Distribution (*GLD*) with moment generating

function (Balakrishnan 1992) :

$$M(t) = \frac{\Gamma(\alpha_1 - t)\Gamma(\alpha_2 + t)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}, -\alpha_2 < t < \alpha_1. \quad (3.9)$$

When $\alpha_1 = \alpha_2 = \alpha$, then (3.8) equals

$$f(z) = \frac{1}{\beta(\alpha, \alpha)} \frac{\exp(-\alpha z)}{(1 + \exp(-z))^{2\alpha}}, -\infty < z < \infty, \alpha > 0. \quad (3.10)$$

which is a Generalized Logistic Distribution with moment generating function given by

$$M(t) = \frac{\Gamma(\alpha - t)\Gamma(\alpha + t)}{(\Gamma(\alpha))^2}, -\alpha < t < \alpha. \quad (3.11)$$

and for the special case when $\alpha = 1$

$$f(z) = \frac{\exp(-z)}{(1 + \exp(-z))^2}, -\infty < z < \infty, \quad (3.12)$$

which is Logistic distribution with $E(Z) = 0$ and $Var(Z) = \frac{\pi^2}{3}$.

Using the moment generating function in (3.9), the mean and variance of Z can be written as (Balakrishnan 1992 and Wu et al. 2000):

$$E(Z) = \psi(\alpha_2) - \psi(\alpha_1)$$

and

$$Var(Z) = \psi'(\alpha_1) + \psi'(\alpha_2)$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the first and the second derivatives of the logarithm of the Gamma function, digamma and trigamma functions, respectively

(Balakrishnan 1992). Hence,

$$\tau_{ij} = \text{Var}(\log \frac{x_i}{x_j}) = \text{Var}(Z) = \psi'(\alpha_i) + \psi'(\alpha_j)$$

and the Total Variability defined in (3.6)

$$\text{Totvar}(\mathbf{x}) = \frac{D-1}{D} \sum_{i=1}^D \psi'(\alpha_i). \quad (3.13)$$

When $\alpha_i = \alpha_j = \alpha$

$$\tau_{ij} = 2\psi'(\alpha)$$

and

$$\text{Totvar}(\mathbf{x}) = (D-1)\psi'(\alpha). \quad (3.14)$$

Figure (5) is a scatter plot of the Total Variability computed using Aitchison logratio transformation for 3-part compositional data generated from gamma distributions with 100 observations and different values of α ($\alpha = 10, \dots, 40$) and the derived Total Variability in Equation (3.14) for each value of α . Figure (6) shows the two measures of Total Variability against α . It is clear that the two measures are nearly the same and decreasing with α .

Relationship between Total Variability and Sum of Coefficients of Variation

For a Dirichlet distribution with $\alpha_i = \alpha$, from equation (3.5), the *SCV* is given by

$$\text{SCV} = D \sqrt{\frac{D-1}{D\alpha+1}}$$

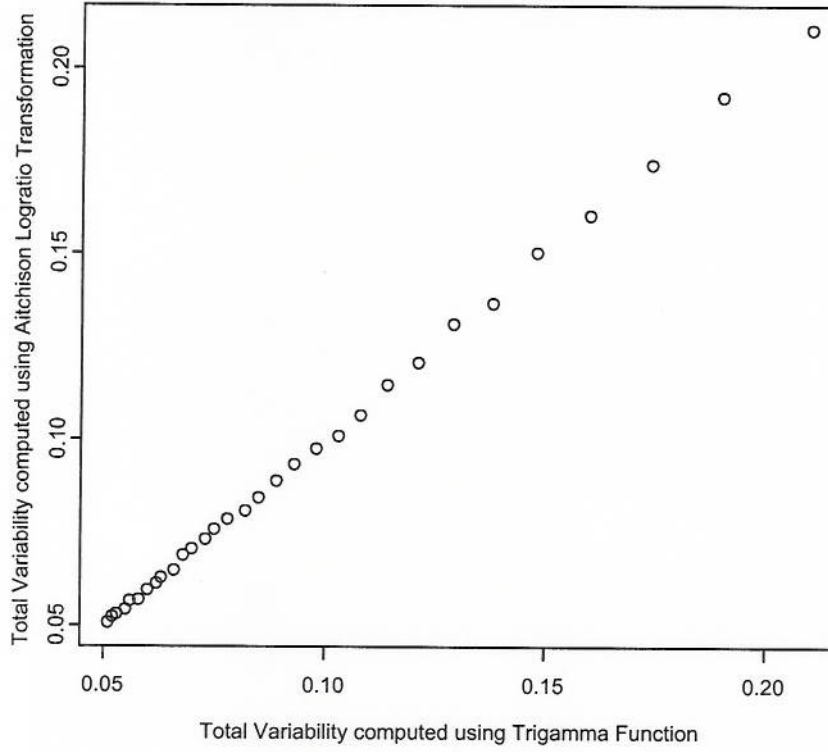


Figure 5. Total Variability computed using Aitchison logratio transformation for 3-part compositional data ($D=3$) simulated from Gamma and Total Variability using Trigamma Function

Solving for α we have

$$\hat{\alpha} = \frac{D - 1 - (\frac{S\hat{C}V}{D})^2}{D(\frac{S\hat{C}V}{D})^2}$$

Then an estimate for the Total Variability is

$$Totvar(\mathbf{x}) = (D - 1)\psi'(\hat{\alpha}) = (D - 1)\psi'\left(\frac{D - 1 - (\frac{S\hat{C}V}{D})^2}{D(\frac{S\hat{C}V}{D})^2}\right). \quad (3.15)$$

To illustrate the above findings we simulate 1000 3-part compositional data

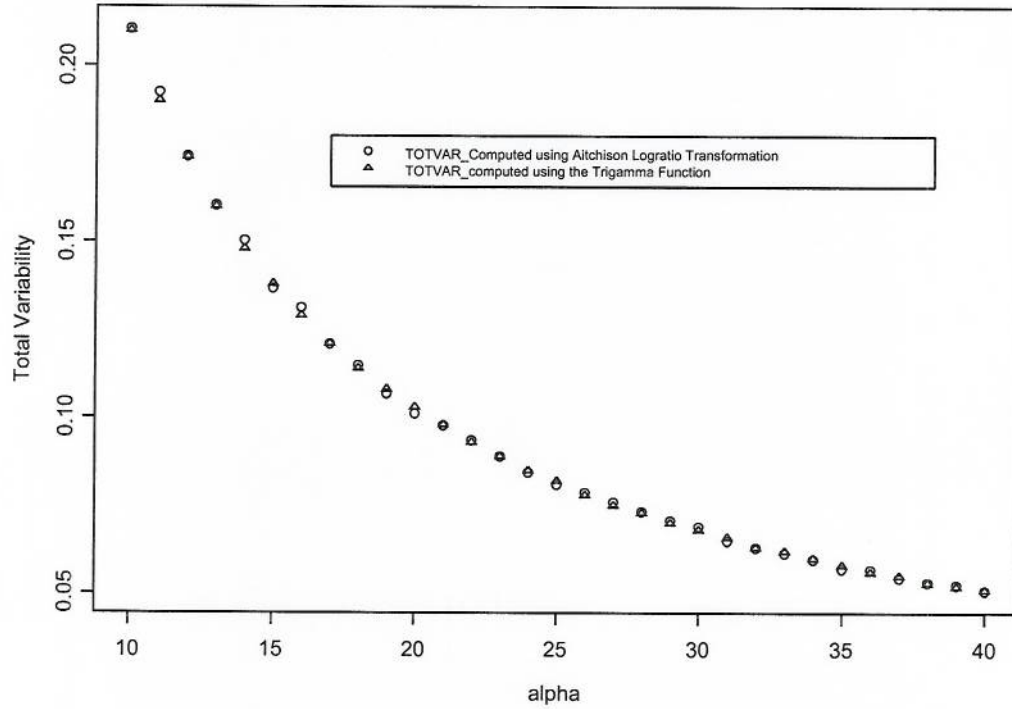


Figure 6. Total Variability computed using Aitchison logratio transformation for 3-part compositional data ($D=3$) simulated from Gamma and Total Variability using Trigamma Function against the corresponding values of α

sets from three independent random Gammas with 100 observations and $\alpha = 10$, then apply the closure operation which divides each component by the sum of the components, thus scaling the data to the constant sum 1. Figure (7) shows Triangle plot for one such data set. Table (1) presents summary statistics of Coefficients of Variation, Sum of Coefficients of Variation and Total Variability using both Aitchison logratio and Total Variability defined in equation (3.15) for the simulated data. The estimates of the means of Aitchison Total Variability and Total Variability using Trigamma as a function of the Sum of Coefficients of Variation are the same to three

	$CV(X_1)$	$CV(X_2)$	$CV(X_3)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.254	0.254	0.253	0.761	0.211	0.211

Table 1. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.279	1.40	0.420	0.419

Table 2. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 5-part simulated compositional dataset of size $n=100$

decimal places.

Figure (8) is a scatter plot of Aitchison's Total Variability and the derived Total Variability using the Trigamma as a function of the Sum of Coefficients of Variation. It is clear that the two measures of Total Variability are very similar. Figure (9) shows the scatter plot of the derived Total Variability using the Trigamma function and the Sum of Coefficients of Variation. The plot indicates a strong positive correlation between the two measures.

Tables (2) and (3) and Figure (10) show similar findings for 5-part and 7-part compositional data sets generated from a Gamma distribution with $\alpha = 10$ and $n = 100$. Furthermore, tables (4) and (5) and Figure (11) show similar findings for 3-part compositional data sets generated from a Gamma distribution with smaller sizes, $n = 50$ and $n = 30$.

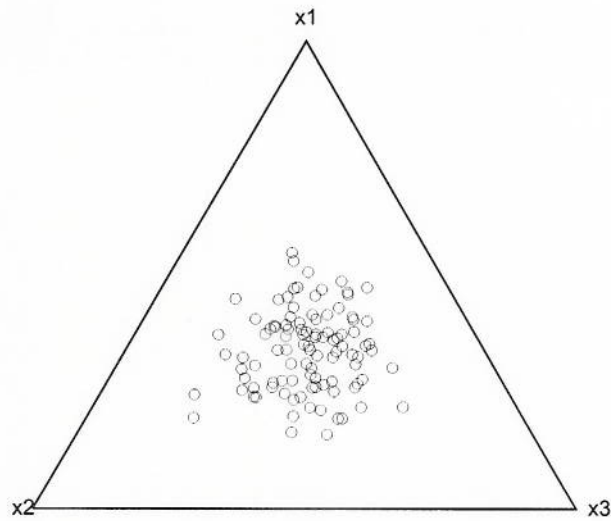


Figure 7. Triangle plot of 3-part compositional data set simulated from Gamma distribution with $\alpha = 10$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.290	2.030	0.631	0.629

Table 3. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 7-part simulated compositional dataset of size $n=100$

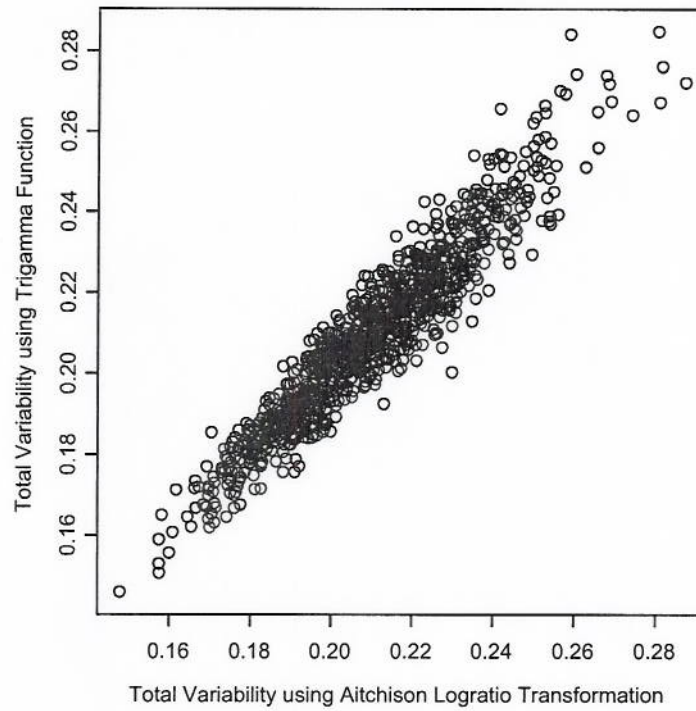


Figure 8. Aitchison's Total Variability and derived Total Variability using Trigamma function for 3-part Simulated Data

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.253	0.760	0.211	0.211

Table 4. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=50$

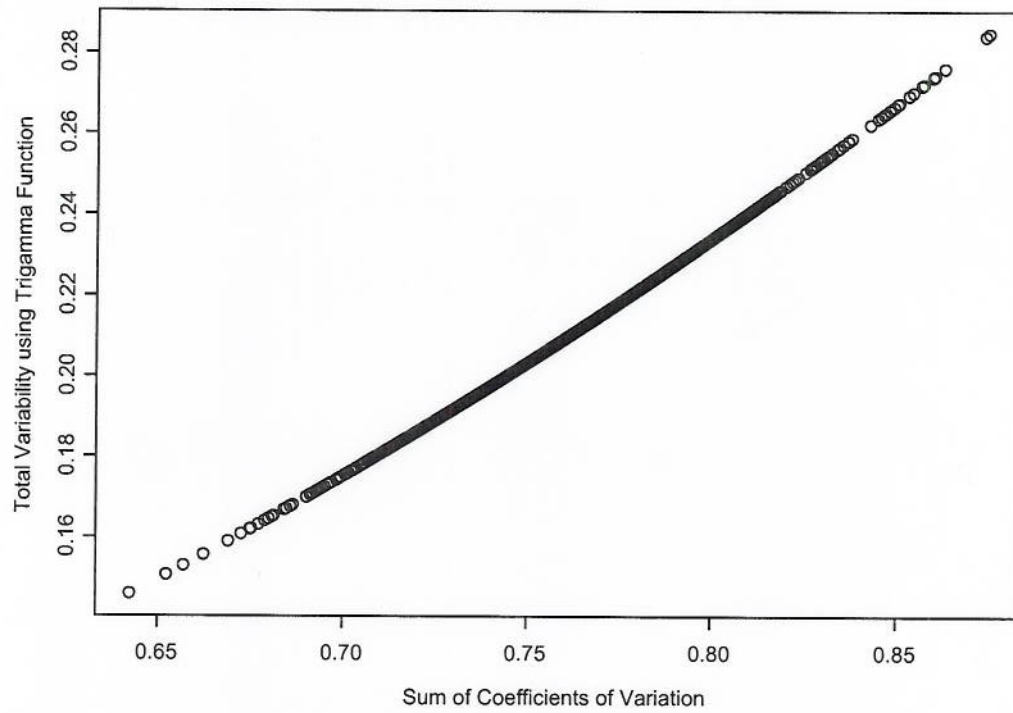


Figure 9. Derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 3-part Simulated Data

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.253	0.758	0.211	0.211

Table 5. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=30$

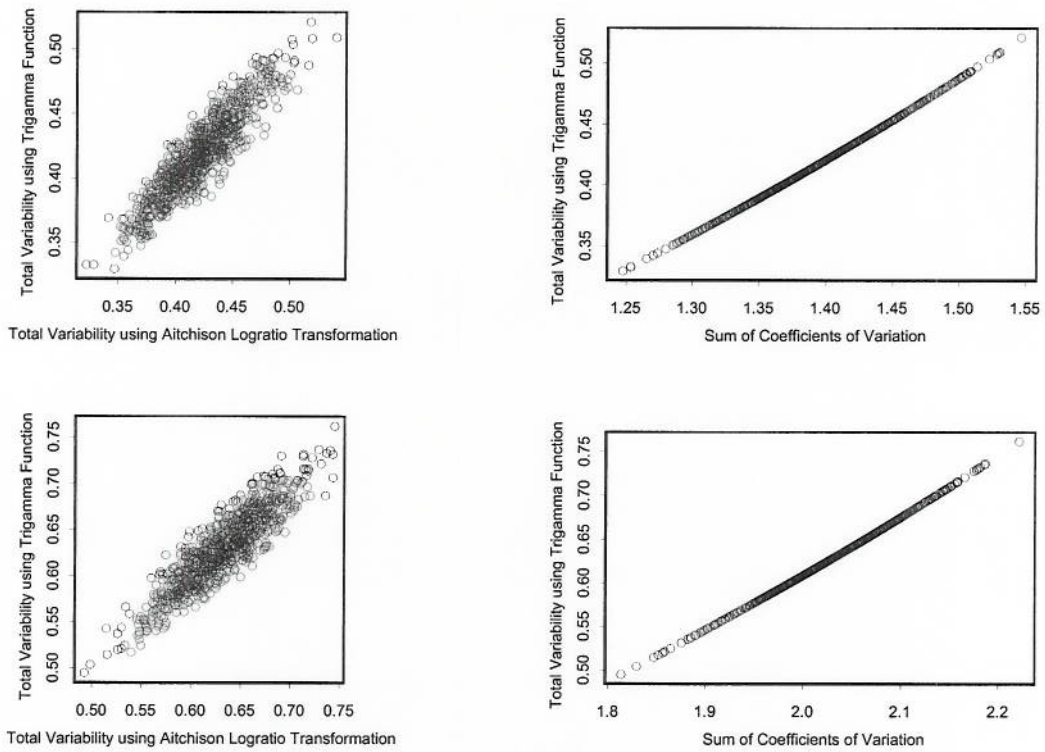


Figure 10. Total Variability using Trigamma function with Aitchison's Total Variability and Sum of Coefficients of Variation for 5-part and 7-part Simulated Data respectively

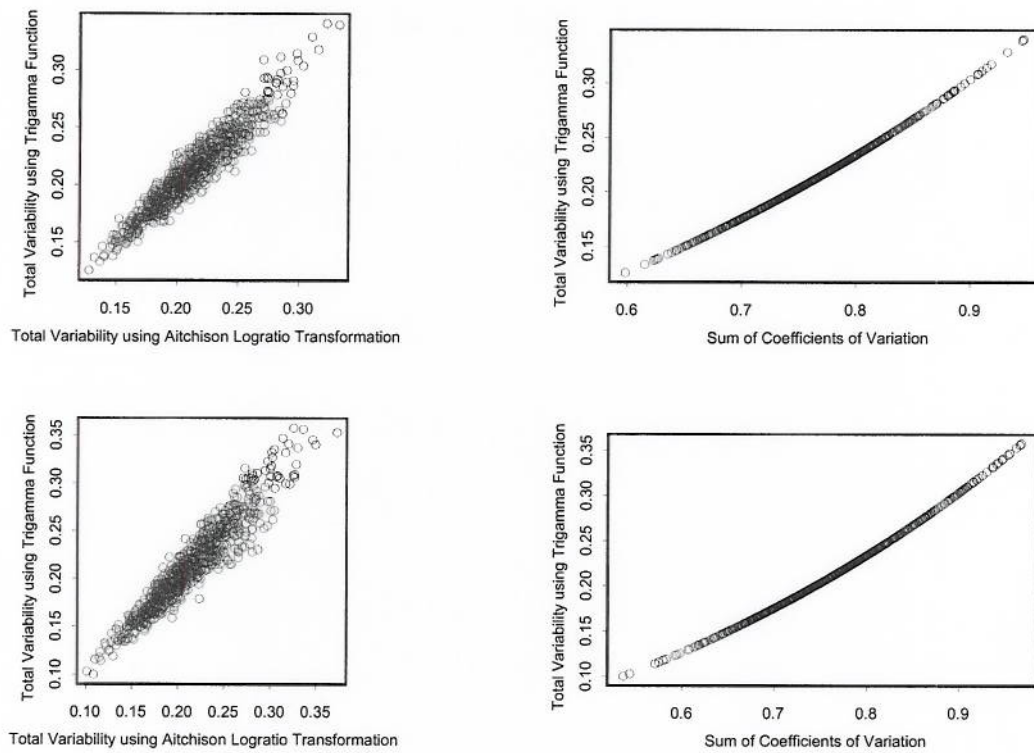


Figure 11. Total Variability using Trigamma function with Aitchison's Total Variability and Sum of Coefficients of Variation for 3-part Simulated Data sets with $n=50$ and $n=30$ respectively

Relationship between Total Variability and Sum of Coefficients of Variation for smaller values of α

In this section we investigate the relationship between Total Variability and Sum of Coefficients of Variation for smaller values of α . Figure (12) below shows triangle plots of 3-part compositional data sets generated from Gamma distribution with $\alpha = 5, 2, 1, 0.5, 0.3$, and 0.1 . The findings for $\alpha > 1$ are similar to what we found previously. Tables (6-8) present summary statistics of Sum of Coefficients of Variation, Aitchison Total Variability and Total Variability using the derived formula of Trigamma as a function of Sum of Coefficients of Variation for 3-part compositional data sets generated from Gamma distributions with $\alpha = 5$, $\alpha = 2$ and $\alpha = 1$. Figure (13) shows plots of the derived Total Variability against Aitchison Total Variability and Sum of Coefficients of Variation computed for these data sets. The plots reveal strong correlations between the derived Total Variability and Aitchison Total Variability as well as between the derived Total Variability and Sum of Coefficients of Variation. However, the relationship between Aitchison Total Variability and Sum of Coefficients of Variation as well as relationship between Aitchison Total Variability and the derived Total Variability using the Trigamma function change when $\alpha < 1$. Tables (9-11) present summary statistics for these measures. Clearly when $\alpha < 1$, the Trigamma as a function of the Sum of Coefficients of Variation is not a good estimate of Aitchison Total Variability and the relationship between Aitchison Total Variability and Sum of Coefficients of Variation is not strong as it appears from Figure (14). Furthermore, the relationship between the derived Total Variability and Sum of Coefficients of Variation is not linear like what we saw before when $\alpha > 1$.

One explanation of these results is that Logratio analysis and Aitchison Total Variability do not produce good predictions for edge cases when some proportions are close to zero. As x_i approach zero, logratios approach negative or positive infinity and this is the case when $\alpha < 1$.

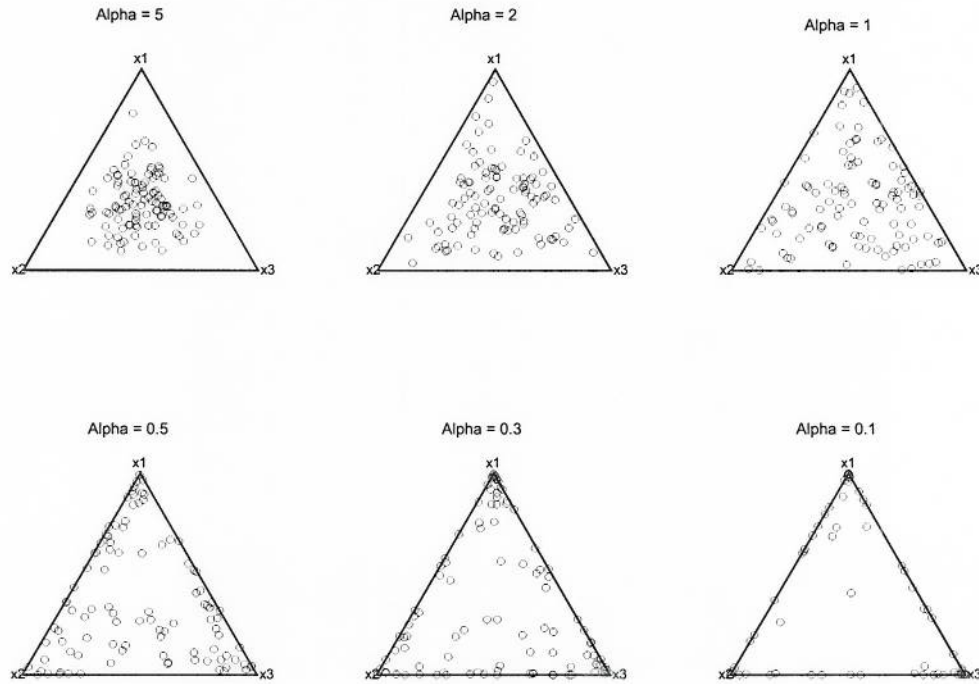


Figure 12. Triangle plots of 3-part compositional data sets simulated from Gamma distributions with $\alpha = 5$, $\alpha = 2$, $\alpha = 1$, $\alpha = 0.5$, $\alpha = 0.3$, and $\alpha = 0.1$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.353	1.058	0.441	0.442

Table 6. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 5$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.534	1.601	1.290	1.300

Table 7. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 2$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.708	2.123	3.280	3.343

Table 8. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 1$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.897	2.693	9.816	10.285

Table 9. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 0.5$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	1.033	3.095	24.661	26.367

Table 10. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 0.3$

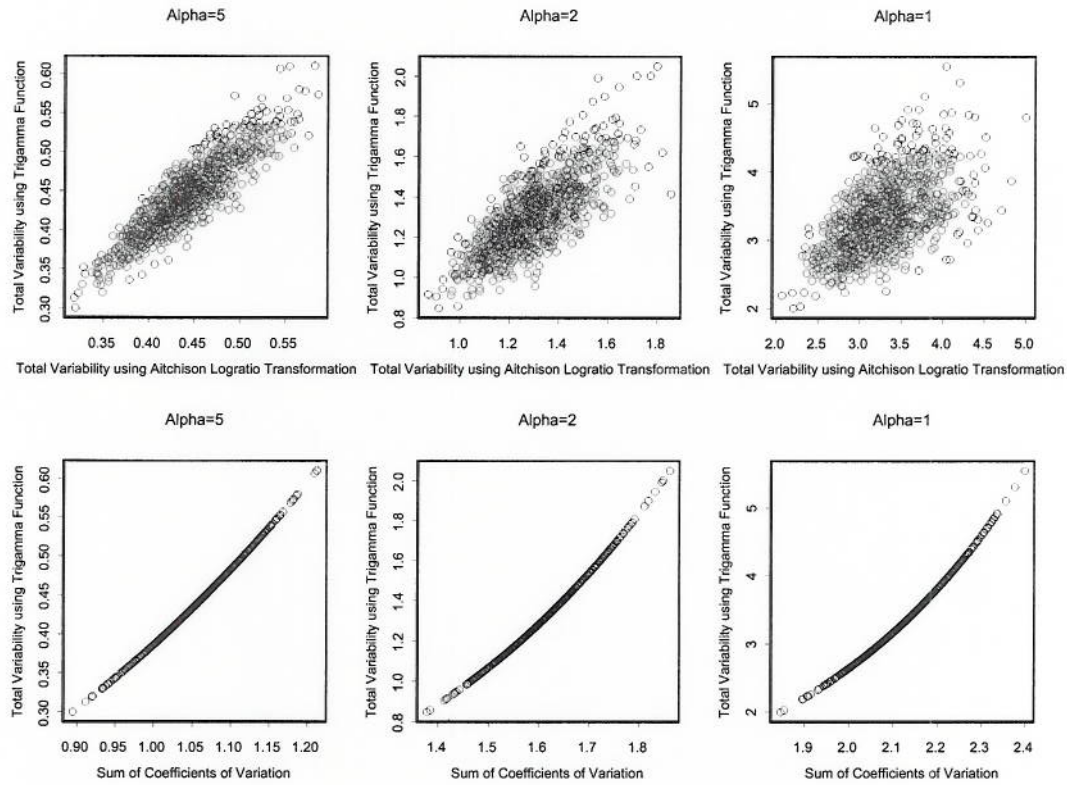


Figure 13. Aitchison's Total Variability, derived Total Variability using Trigamma function, and Sum of Coefficients of Variation for 3-part Simulated Data sets for $\alpha = 5$, $\alpha = 2$ and $\alpha = 1$

	$CV(X_i)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	1.253	3.759	202.550	270.019

Table 11. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha = 0.1$

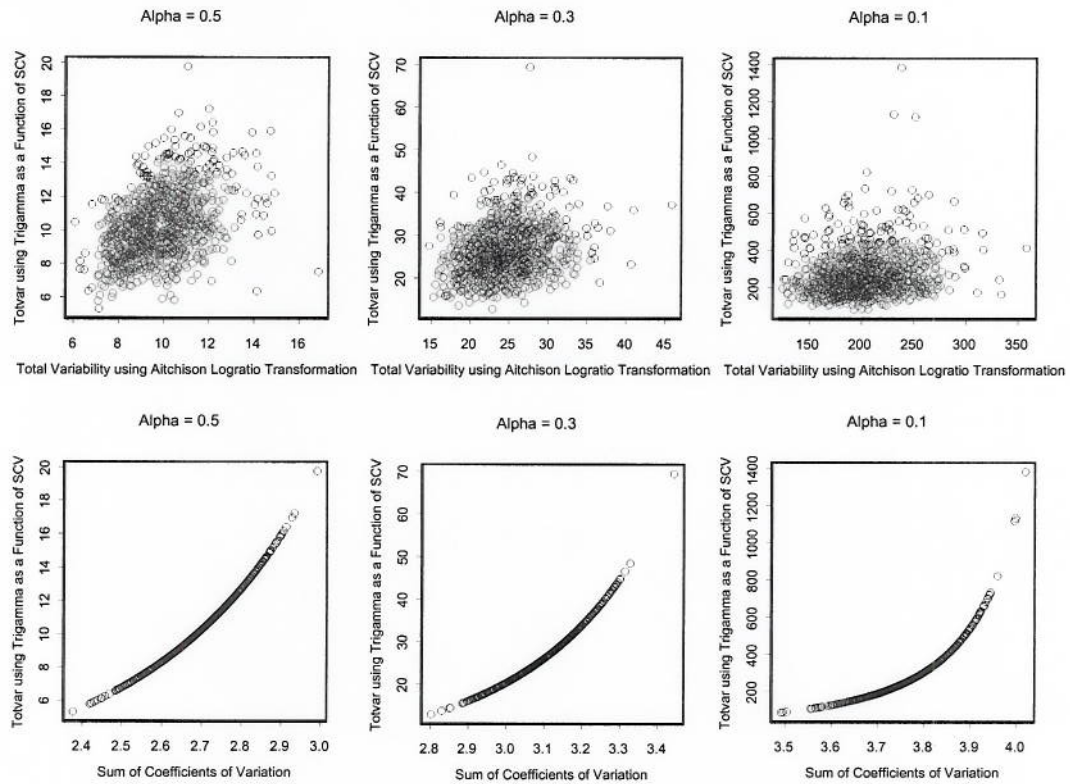


Figure 14. Aitchison's Total Variability, Total Variability using Trigamma as a function of SCV and Sum of Coefficients of Variation for 3-part Simulated Data sets with $\alpha=0.5, 0.3$ and 0.1

**Relationship between Total Variability and Sum of
Coefficients of Variation for different α s**

Consider W_1, \dots, W_D be independent distributed random variables from $Gamma(i\alpha, \beta)$ with $i = 1, \dots, D$. Recall from equation (3.3),

$$CV(x_i) = \sqrt{\frac{\alpha_+ - \alpha_i}{\alpha_i(\alpha_+ + 1)}}$$

when $\alpha_i = i\alpha$,

$$\alpha_+ = \sum_{i=1}^D \alpha_i = \alpha \sum_{i=1}^D i = \alpha \frac{D(D+1)}{2}$$

and

$$CV(x_i) = \sqrt{\frac{\alpha \frac{D(D+1)}{2} - i\alpha}{i\alpha(\alpha \frac{D(D+1)}{2} + 1)}} = \sqrt{\frac{\frac{D(D+1)}{2} - i}{i(\alpha \frac{D(D+1)}{2} + 1)}},$$

for $i = 1, \dots, D$. Hence,

$$SCV = \sum_{i=1}^D \sqrt{\frac{\frac{D(D+1)}{2} - i}{i(\alpha \frac{D(D+1)}{2} + 1)}}$$

and

$$SCV^2 = \frac{1}{\alpha \frac{D(D+1)}{2} + 1} \left(\sum_{i=1}^D \sqrt{\frac{D(D+1)}{2i} - 1} \right)^2$$

hence,

$$\hat{\alpha} = \frac{(\sum_{i=1}^D \sqrt{\frac{D(D+1)}{2i} - 1})^2 - SCV^2}{\frac{D(D+1)}{2} SCV^2}$$

Recall from equation (3.15)

$$Tot\hat{var}(\mathbf{x}) = \frac{D-1}{D} \sum_{i=1}^D \psi'(\alpha_i) = \frac{D-1}{D} \sum_{i=1}^D \psi'(i\hat{\alpha})$$

	$CV(X_1)$	$CV(X_2)$	$CV(X_3)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.286	0.181	0.128	0.595	0.127	0.127

Table 12. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha_1 = 10, \alpha_2 = 20$ and $\alpha_3 = 30$

then

$$Tot\hat{v}ar(\mathbf{x}) = \frac{D-1}{D} \sum_{i=1}^D \psi'(i) \frac{(\sum_{i=1}^D \sqrt{\frac{D(D+1)}{2i}} - 1)^2 - S\hat{C}V^2}{\frac{D(D+1)}{2} S\hat{C}V^2}. \quad (3.16)$$

To illustrate the above findings numerically we simulate 1000 3-part compositional data sets from three independent random Gammas with 100 observations and with parameters $\alpha_1 = 10, \alpha_2 = 20$ and $\alpha_3 = 30$, then apply the closure operation so the sum of components add to 1. Figure (15) is a Triangle plot for one data set. Table (12) presents the summary statistics of Coefficients of Variation, Sum of Coefficients of Variation, Aitchison Total Variability, and the derived Total Variability in equation (3.16) for the simulated data.

Figure (16) shows a scatter plot of Aitchison's Total Variability and the derived Total Variability using the Trigamma as a function of the Sum of Coefficients of Variation. It is clear that the estimated Total Variability using the Trigamma function is a good estimate of Aitchison Total Variability. Figure (17) shows a scatter plot of the derived Total Variability using Trigamma function and the Sum of Coefficients of Variation. The plot reveals a strong correlation between derived Total Variability and Sum of Coefficients of Variation.

We repeated similar analysis in the cases where W_1, \dots, W_D are independent distributed random variables from Gamma with parameters $\alpha, 5\alpha$ and 10α and again

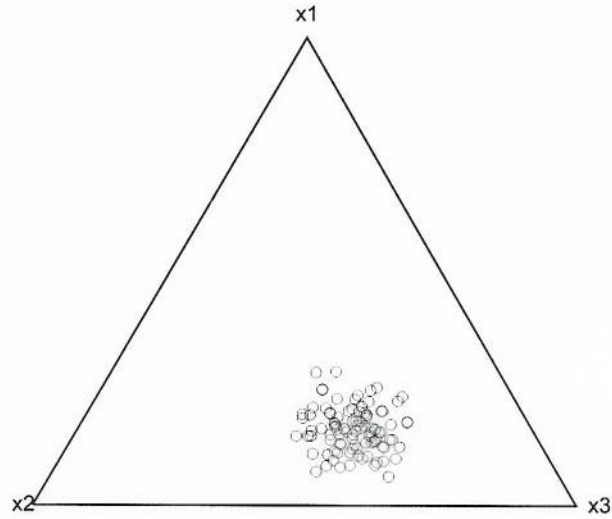


Figure 15. Triangle plot of 3-part compositional data set simulated from gamma distribution with $\alpha_1 = 10$, $\alpha_2 = 20$ and $\alpha_3 = 30$

with parameters α , 50α and 100α . Tables (13) and (14) as well as Figures (18) and (19) show similar findings for data sets generated from Gamma distributions with $\alpha_1 = 10$, $\alpha_2 = 50$ and $\alpha_3 = 100$ and Gamma distributions with $\alpha_1 = 1$, $\alpha_2 = 50$ and $\alpha_3 = 100$

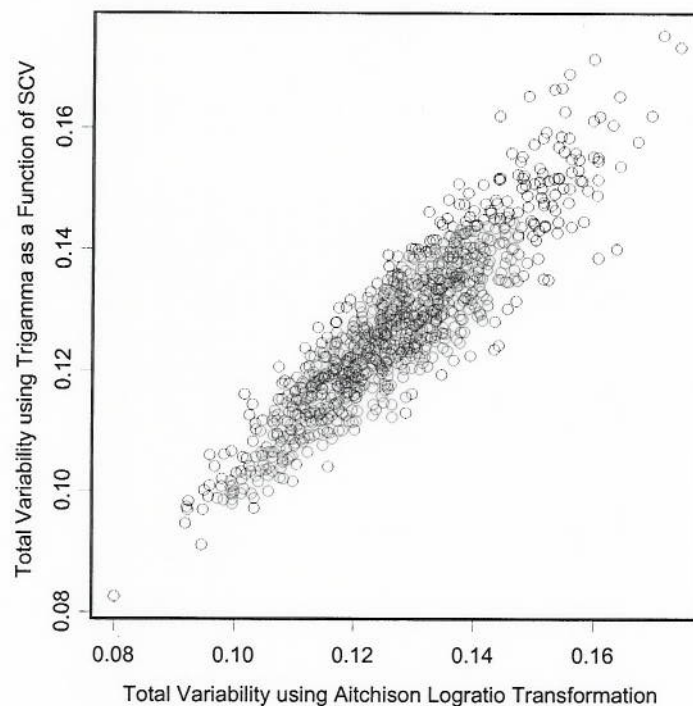


Figure 16. Aitchison's Total Variability and derived Total Variability using Trigamma function for 3-part Simulated Data with $\alpha_1 = 10$, $\alpha_2 = 20$ and $\alpha_3 = 30$

	$CV(X_1)$	$CV(X_2)$	$CV(X_3)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.305	0.116	0.061	0.482	0.090	0.090

Table 13. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha_1 = 10$, $\alpha_2 = 50$ and $\alpha_3 = 100$

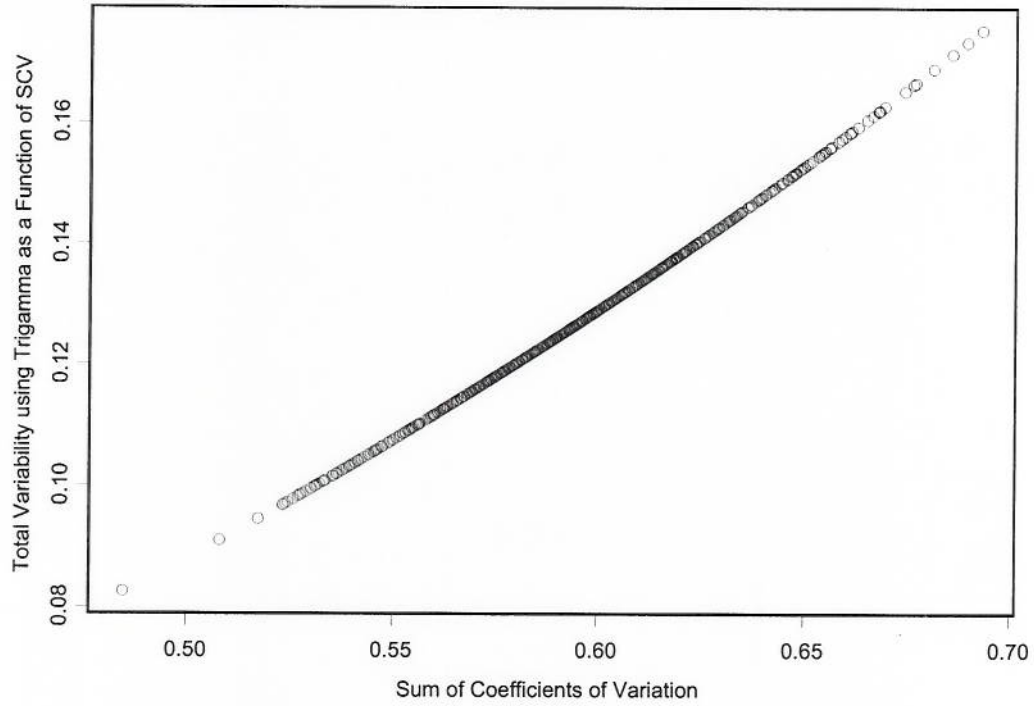


Figure 17. Derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 3-part Simulated Data with $\alpha_1 = 10$, $\alpha_2 = 20$ and $\alpha_3 = 30$

	$CV(X_1)$	$CV(X_2)$	$CV(X_3)$	Sum of Coefficients of Variation	Aitchison Total Variability	Total Variability using Trigamma Function and SCV
Mean	0.985	0.115	0.058	1.158	1.117	1.117

Table 14. Summary Statistics of 1000 simulated Coefficients of Variation, Sum of Coefficients of Variation and Total Variability for 3-part simulated compositional dataset of size $n=100$ and $\alpha_1 = 1$, $\alpha_2 = 50$ and $\alpha_3 = 100$

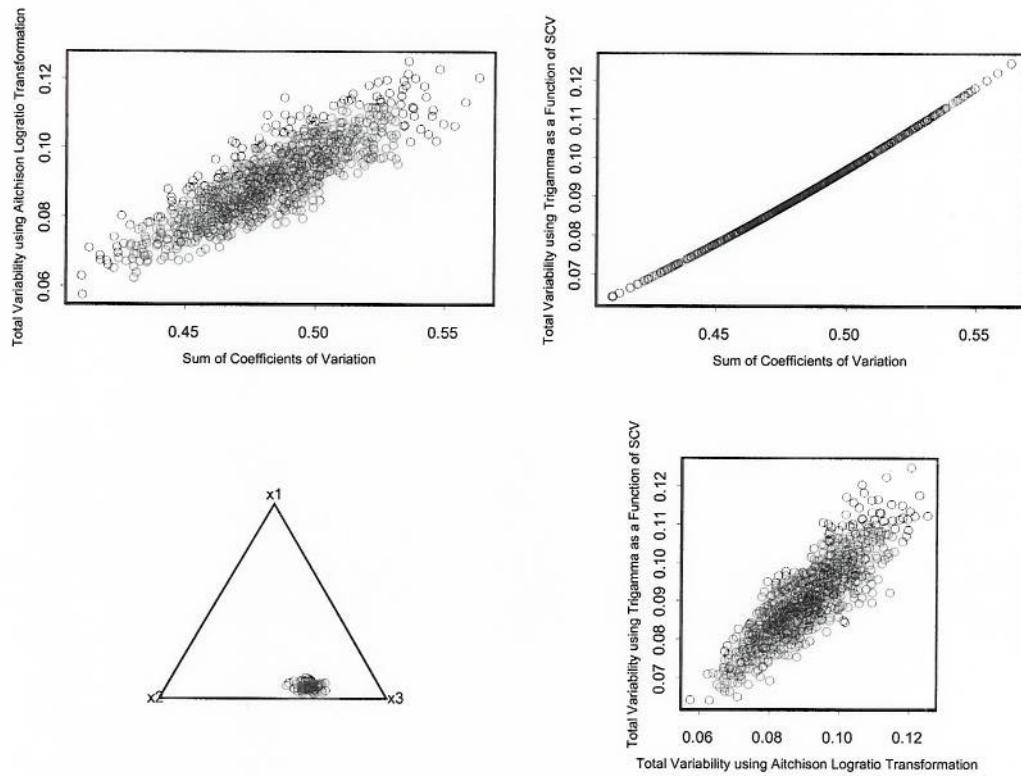


Figure 18. Plots of 3-part compositional data set simulated from gamma distribution with $\alpha_1 = 10$, $\alpha_2 = 50$ and $\alpha_3 = 100$

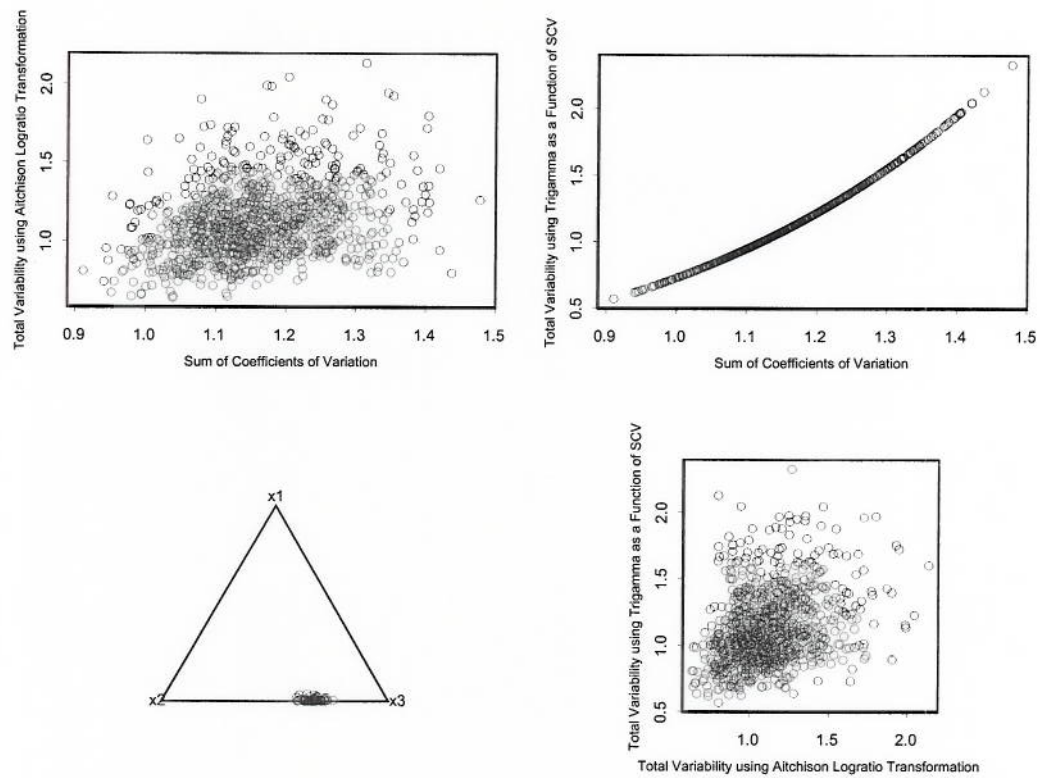


Figure 19. Plots of 3-part compositional data set simulated from gamma distribution with $\alpha_1 = 1$, $\alpha_2 = 50$ and $\alpha_3 = 100$

Relationship between Total Variability and Sum of Coefficients of Variation for Correlated Variables

In this section, we examine the relationship between the two measures for correlated variables. To generate correlated Gamma variables, we first generate correlated Multivariate Normal sample with known correlation matrix. This imposes a similar rank correlation on the Normal sample. We then impose this same rank correlation on randomly generated independent Gammas. We investigate the relationship using three different data sets. For the first dataset, we simulate a Multivariate Normal sample with three variables using the following correlation matrix:

$$\begin{bmatrix} 1.00 & 0.70 & 0.70 \\ 0.70 & 1.00 & 0.70 \\ 0.70 & 0.70 & 1.00 \end{bmatrix}$$

From the simulated Multivariate Normal sample we impose its same rank correlation on three independent random variables simulated from Gamma distribution. The estimated Gamma correlation matrix for one dataset is:

$$\begin{bmatrix} 1.00 & 0.71 & 0.71 \\ 0.71 & 1.00 & 0.72 \\ 0.71 & 0.72 & 1.00 \end{bmatrix}$$

Figure (20) is a triangle plot for one data set and Figure (21) displays a scatter plot of Aitchison's Total Variability and the derived Total Variability based on the Trigamma function and Sum of Coefficients of Variation. The results are based on 1000 simulated 3-part compositional data sets from correlated Gamma variables with 100 observations and $\alpha = 5$. Again a strong positive correlation is clear from the plot.

The second dataset is a 5-part compositional data generated from five corre-

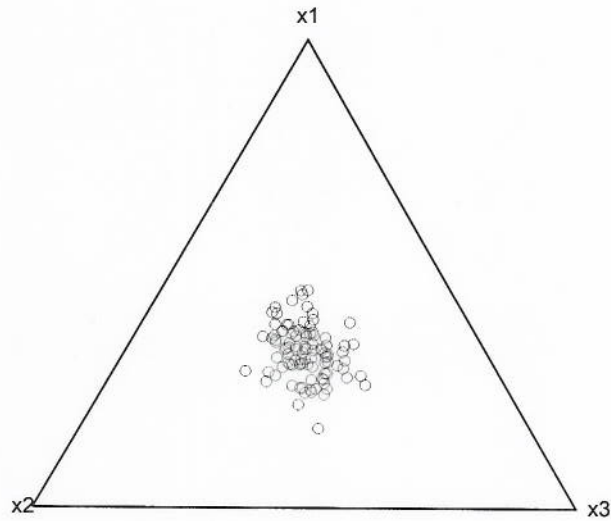


Figure 20. Triangle plot for 3-part compositional data set simulated using correlated Gammas with $\alpha = 5$

lated Gamma variables. The correlation matrix used for this simulation derived from Aitchison and Greenacre (2002). They investigated a 6-part colour compositions in 22 paintings. In each painting the artist used black, white, blue, red, yellow, and other. The correlation matrix of the logratios is:

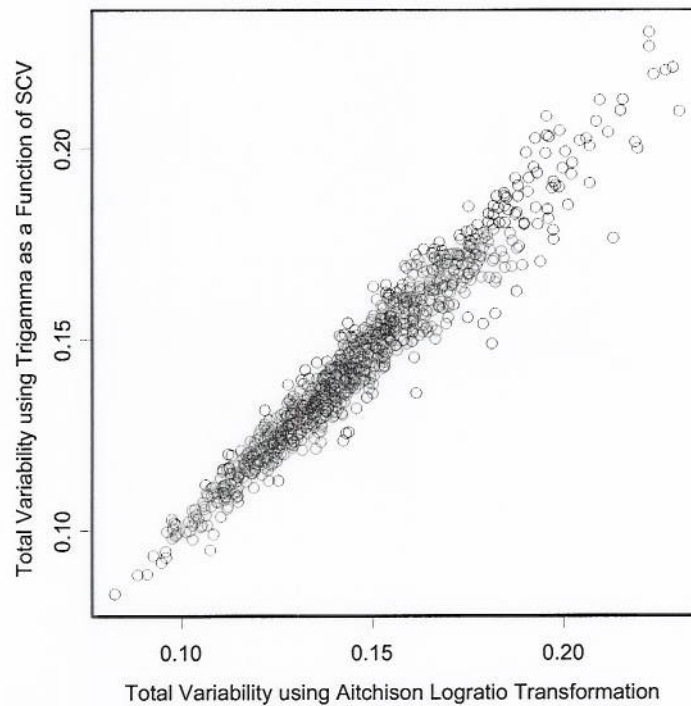


Figure 21. Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 3-part compositional data sets simulated using correlated Gammas with $\alpha = 5$

$$\begin{bmatrix} 1.000 & -0.069 & 0.005 & 0.188 & 0.213 \\ -0.069 & 1.000 & 0.352 & 0.147 & 0.606 \\ 0.005 & 0.352 & 1.000 & -0.845 & 0.883 \\ 0.188 & 0.147 & -0.845 & 1.000 & -0.561 \\ 0.213 & 0.606 & 0.883 & -0.561 & 1.000 \end{bmatrix}$$

As an illustration, we use this correlation matrix to determine the multivariate rank correlation of a Normal sample. We impose the rank correlations on independent

Gamma random variables. The estimated Gamma correlation matrix for one such sample is:

$$\begin{bmatrix} 1.000 & -0.019 & 0.079 & 0.170 & 0.231 \\ -0.019 & 1.000 & 0.264 & 0.207 & 0.563 \\ 0.079 & 0.264 & 1.000 & -0.831 & 0.851 \\ 0.170 & 0.207 & -0.831 & 1.000 & -0.538 \\ 0.231 & 0.563 & 0.851 & -0.538 & 1.000 \end{bmatrix}$$

Figure (22) is a triangle plot for three variables from one data set and Figure (23) displays a scatter plot of Aitchison's Total Variability and the derived Total Variability based on the Trigamma function and Sum of Coefficients of Variation. The results were based on 1000 simulated 5-part compositional data sets from correlated Gamma variables with 100 observations and $\alpha = 10$. There is a strong positive correlation between Aitchison Total Variability and Total Variability from Sum of Coefficients of Variation consistent with findings from earlier.

The last dataset is 4-part compositional data where we used the correlation matrix of the logratios of the hongite data presented in Chapter 2. The correlation matrix of the logratios is:

$$\begin{bmatrix} 1.00 & 0.88 & -0.62 & 0.75 \\ 0.88 & 1.00 & -0.89 & 0.41 \\ -0.62 & -0.89 & 1.00 & 0.03 \\ 0.75 & 0.41 & 0.03 & 1.00 \end{bmatrix}$$

The estimated Gamma correlation matrix for one simulated dataset is:

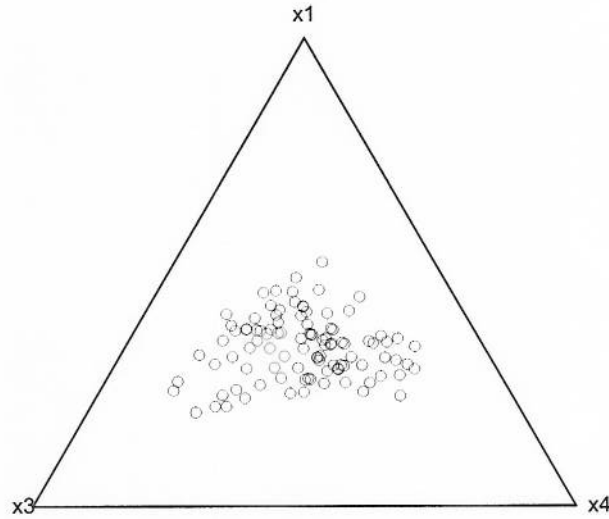


Figure 22. Triangle plot for three variables from 5-part compositional data set simulated using correlated Gammas with $\alpha = 10$

$$\begin{bmatrix} 1.00 & 0.88 & -0.64 & 0.70 \\ 0.88 & 1.00 & -0.87 & 0.36 \\ -0.64 & -0.87 & 1.00 & 0.03 \\ 0.70 & 0.36 & 0.03 & 1.00 \end{bmatrix}$$

Figure (24) displays a scatter plot of Aitchison's Total Variability and the derived Total Variability based on the Trigamma function and Sum of Coefficients of Variation. The results were based on 1000 simulated 4-part compositional data sets from correlated Gamma variables with 100 observations and $\alpha = 10$. Again

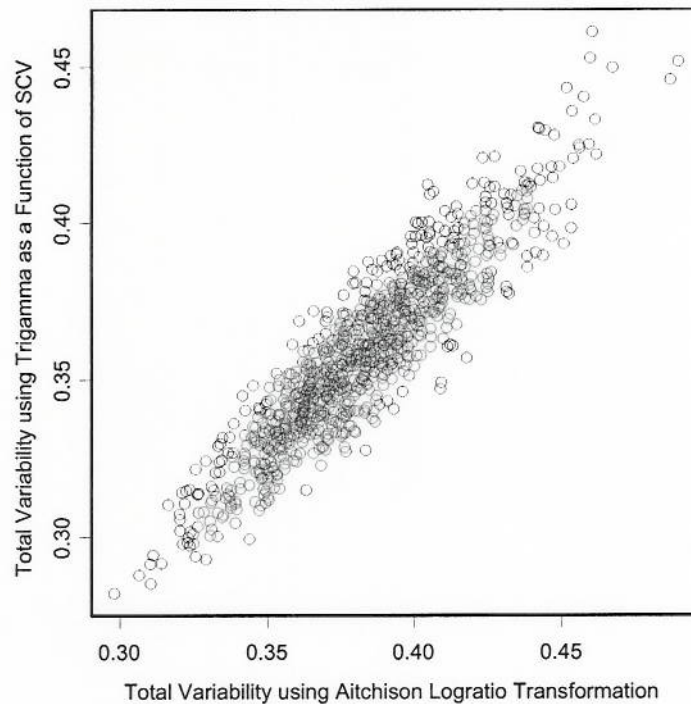


Figure 23. Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 5-part compositional data sets simulated using correlated Gammas with $\alpha = 10$

there is a strong positive correlation between Aitchison Total Variability and Total Variability from Sum of Coefficients of Variation.

Finally, we close this section by generating 5-part compositional datasets from Additive Logistic Normal Distribution simulated using correlated Multivariate Normal samples. For this simulation we used the covariance matrix of the logratios of the hongite data. Figure (25) displays a scatter plot of Aitchison's Total Variability and the derived Total Variability based on the Trigamma function and Sum of Coefficients of Variation. The results were based on 1000 simulated 5-part com-

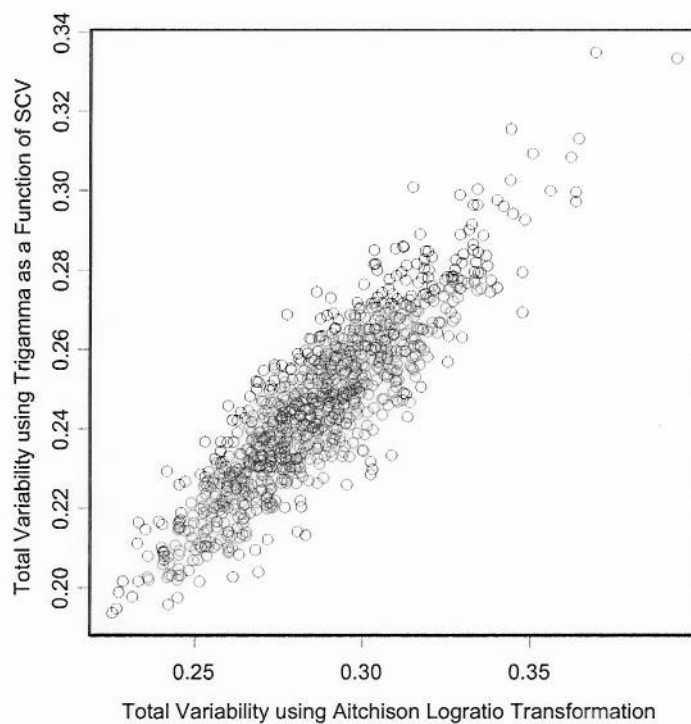


Figure 24. Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 4-part compositional data sets simulated using correlated Gammas with $\alpha = 10$

positional datasets from Additive Logistic Normal variables with 100 observations. Again there is a strong positive correlation between Aitchison Total Variability and Total Variability from Sum of Coefficients of Variation.

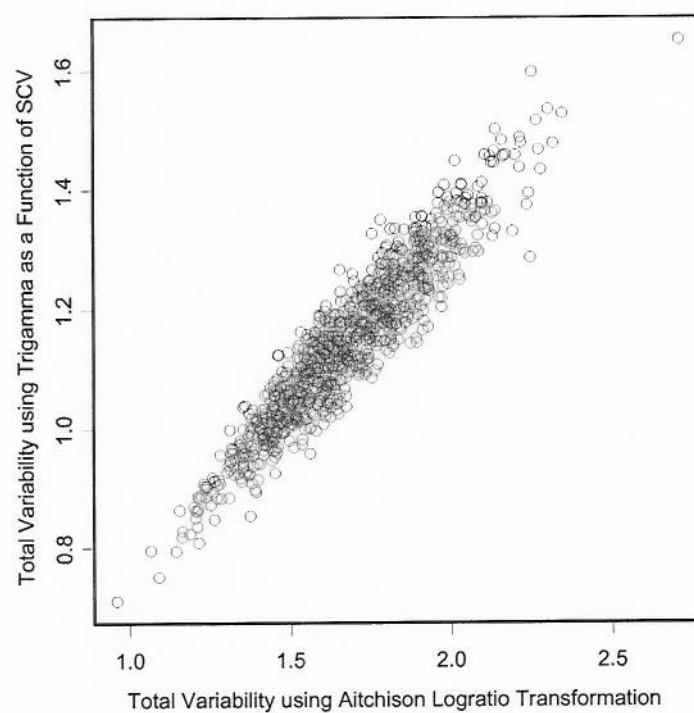


Figure 25. Aitchison's Total Variability and derived Total Variability using Trigamma function and Sum of Coefficients of Variation for 5-part compositional data sets simulated using Additive Logistic Normal

Components	Mean	Standard Deviation	Coefficient of Variation
Paper	0.363	0.115	0.317
Food	0.177	0.089	0.504
Glass	0.141	0.094	0.666
Other	0.092	0.112	1.222
Metal	0.085	0.034	0.401
Plastic	0.070	0.024	0.347
Yard	0.052	0.080	1.549
Text	0.022	0.024	1.096

Table 15. Summary Statistics of Garbage Compositional Data

Illustrative Example using Real Compositional Data Set

Garbage Project

Between 1987 and 1995, a group of archaeologists directed by W.L.Rathje from the University of Arizona's Garbage Project systematically excavated, hand sorted, measured, and recorded thirty tons of contents from fifteen landfills located across North America (W.L.Rathje 2005). In contrast to all of the public concern about fast food packaging and disposable diapers, the results demonstrated that both items together accounted for less than two percent of landfill volume. In contrast, paper, which received little public attention, was the largest proportion of landfill volume.

Proportions of weights of discarded garbage for one week for a sample of 62 households from the Garbage Project, University of Arizona are shown in Appendix A. Table (15) presents summary statistics of the components and Tables (16-18) show all 3-part, 4-part, and 5-part subcompositions formed from this data set and the corresponding SCV, Aitchison's Total Variability and the proportion of Total Variability retained by the subcomposition (R^2). The data in these tables have been ordered in descending order by Total Variability.

Figure (26) shows scatter plots of the Total Variability of 3-part subcompositions and the corresponding Sum of Coefficients of Variation. The plot indicates

a strong positive correlation between SCV and Aitchison's Total Variability. Correlation coefficient between the two measures is 0.95. It was this graph that first suggested the usefulness of SCV. The top five 3-part subcompositions with largest Total Variability sorted in descending order are: (Yard, Text, Other), (Glass, Yard, Other), (Food, Yard, Other), (Paper, Yard, Other), and (Metal, Yard, Other) which match four of the top five subcompositions with largest Sum of Coefficients of Variation. Moreover, the components in this graph fall into groups. The first group in the left bottom of the graph consists of 3-part subcompositions that contain the variables Metal, Plastic, or Paper. The second group consists of subcompositions that contain variable Text and the third group consists of subcompositions that contain variable Other. The last group at the right top of the graph consists of subcompositions contain variables Yard and Other. The two boxplots in Figure (27) displays Sum of Coefficients of Variation and Total Variability for all 3-part subcompositions that contain each component. The graph shows a large agreement between the ordering of the components for the two measures.

Figure (28) shows scatter plots of the Total Variability of all 4-part subcompositions and the corresponding Sum of Coefficients of Variation. The plot indicates a strong correlation between the two measures with a correlation coefficient of 0.95. The top five 4-part subcompositions with largest Total Variability sorted in descending order are: (Food, Yard, Text, Other), (Glass, Yard, Text, Other), (Paper, Yard, Text, Other), (Metal, Yard, Text, Other), and (Plastic, Yard, Text, Other) which match all the five top 4-part subcompositions with largest Sum of Coefficients of Variation but with different order. The two boxplots in Figure (29) displays the Sum of Coefficients of Variation and Total Variability for all 4-part subcompositions that contain each component. The graph shows a large agreement between the ordering of the components for the two measures.

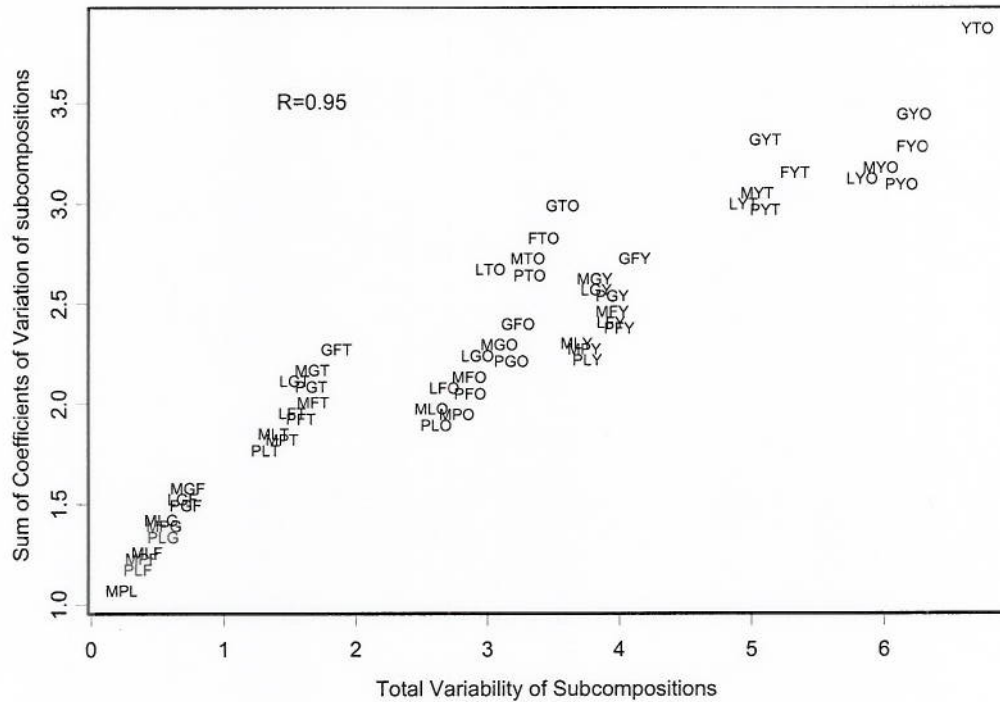


Figure 26. SCV and compositional Total Variability of 3-part subcompositions of the Garbage data. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

Figure (30) shows scatter plots of the Total Variability of 5-part subcompositions and the corresponding Sum of Coefficients of Variation. The plot indicates a strong correlation between SCV and Aitchison's total variability. Correlation coefficient between the two measures is 0.95. The top five 5-part subcompositions with largest Total Variability sorted in descending order are: (Glass, Food, Yard, Text, Other), (Paper, Glass, Yard, Text, Other), (Paper, Food, Yard, Text, Other), (Metal, Plastic, Yard, Text, Other), and (Metal, Food, Yard, Text, Other) which match four of the top five subcompositions with largest Sum of Coefficients of Vari-

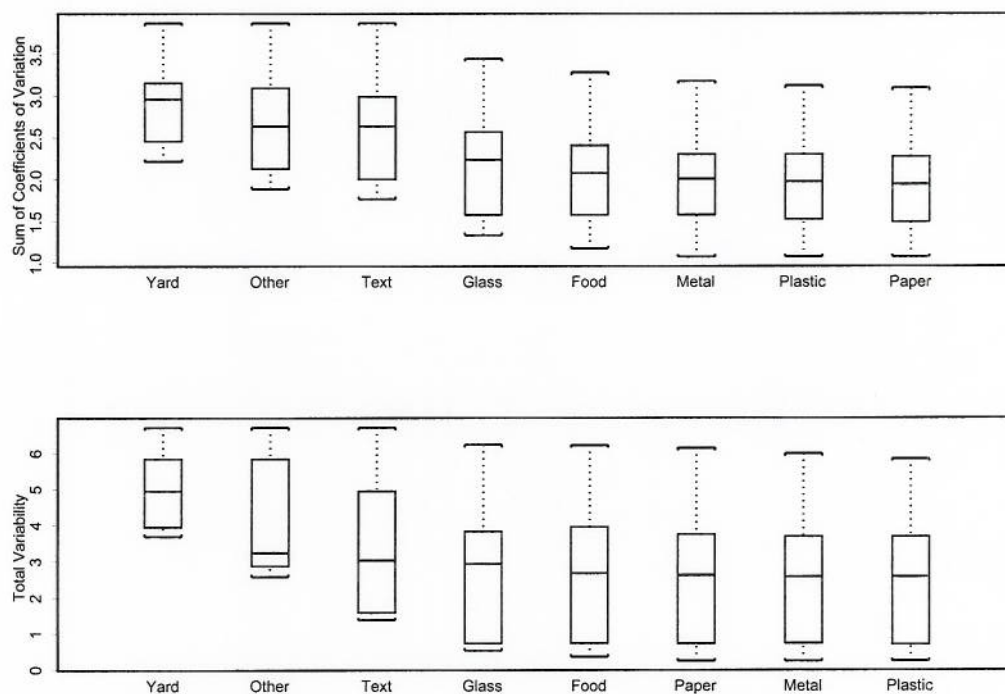
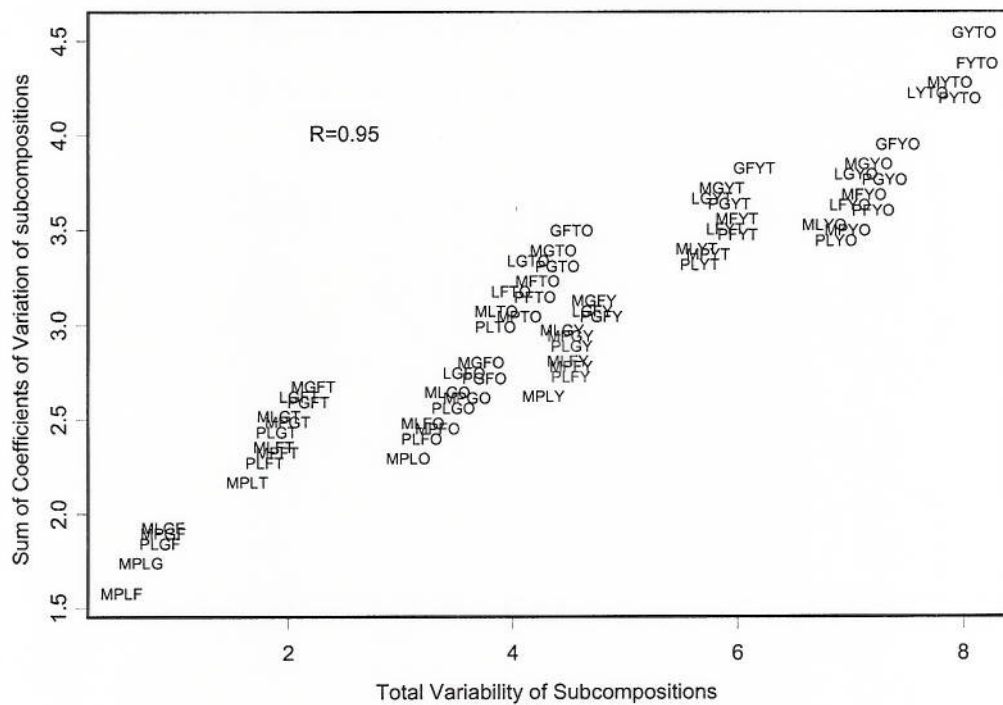


Figure 27. Distribution of the SCV and Total Variability of all 3-part subcompositions that contain each Garbage component

ations. The two boxplots in Figure (31) display the Sum of Coefficients of Variation and Total Variability for all 5-part subcompositions that contain each component. The graph shows a large agreement between the ordering of the components for the two measures.



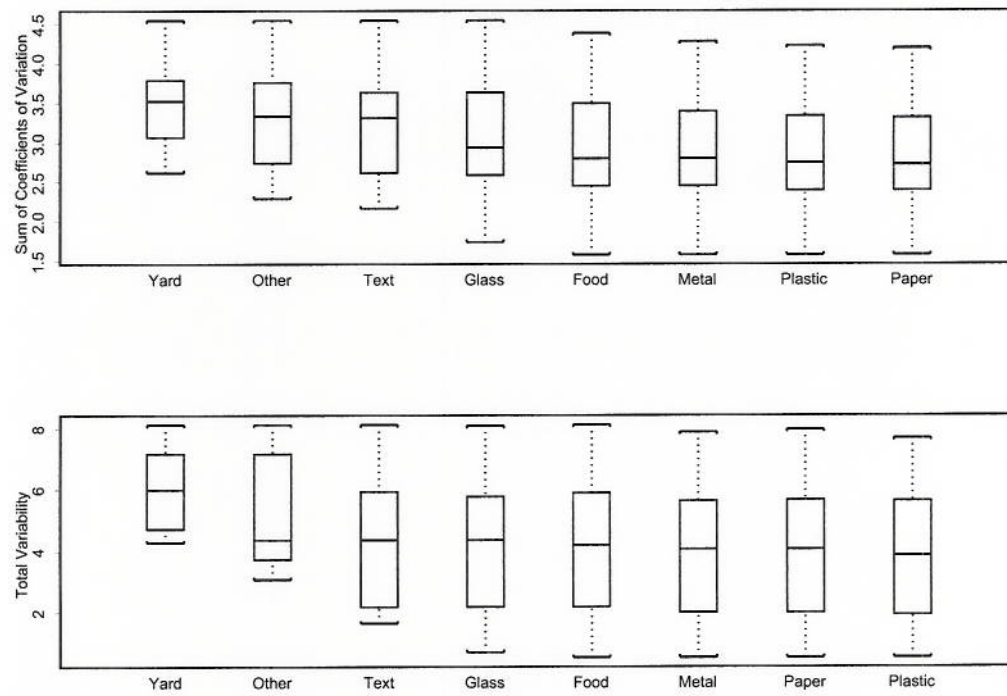


Figure 29. Distribution of the SCV and Total Variability of all 4-part subcompositions that contain each Garbage component

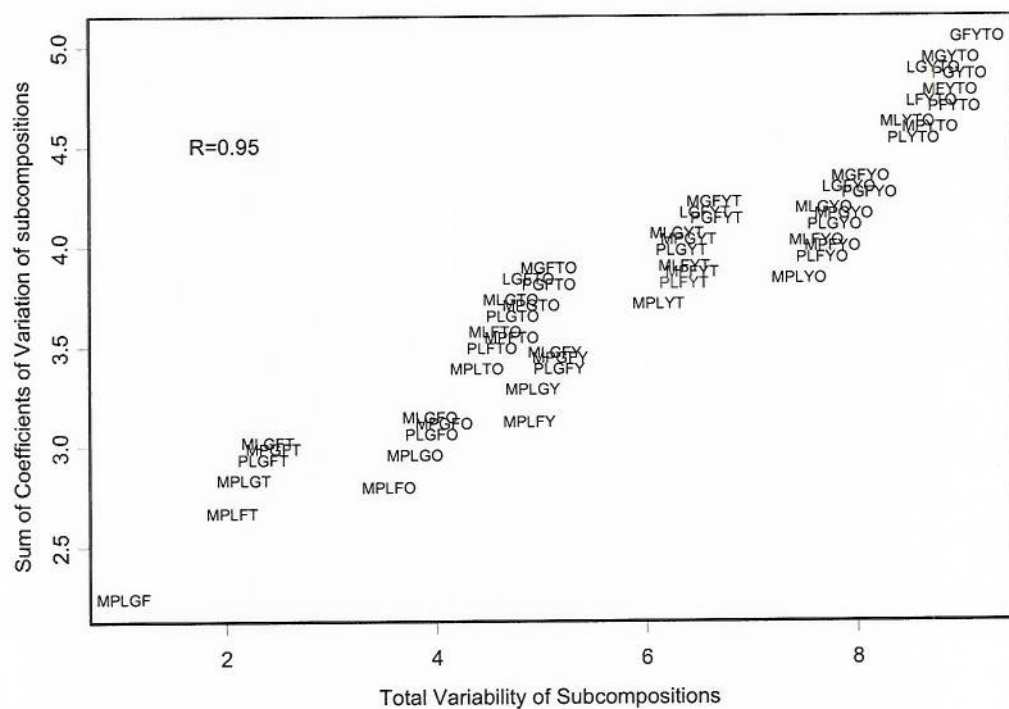


Figure 30. SCV and compositional Total Variability of 5-part subcompositions of the Garbage data. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

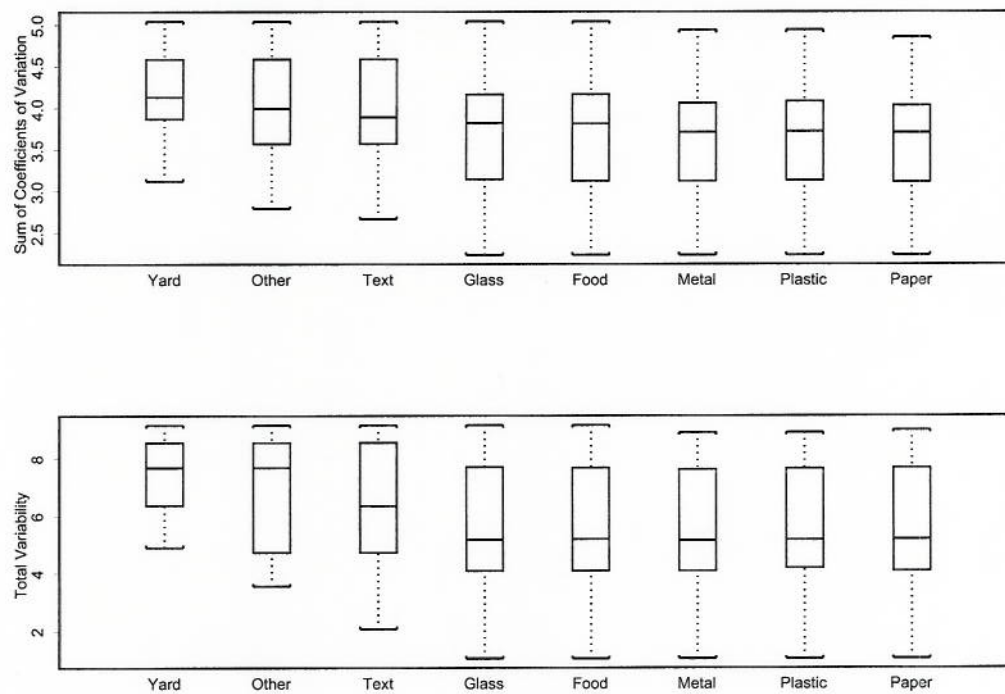


Figure 31. Distribution of the SCV and Total Variability of all 5-part subcompositions that contain each Garbage component

Table 16. All 3-part subcompositions and the corresponding SCV, Total Variability and R^2

	3-part Subcomposition	Sum of Coefficients of Variation	Total Variability	R^2
1	Yard , Text , Other	3.868	6.704	0.640
2	Glass , Yard , Other	3.438	6.224	0.594
3	Food , Yard , Other	3.276	6.214	0.593
4	Paper , Yard , Other	3.089	6.131	0.585
5	Metal , Yard , Other	3.173	5.974	0.570
6	Plastic , Yard , Other	3.118	5.835	0.557
7	Food , Yard , Text	3.149	5.328	0.508
8	Glass , Yard , Text	3.312	5.105	0.487
9	Paper , Yard , Text	2.963	5.105	0.487
10	Metal , Yard , Text	3.047	5.046	0.481
11	Plastic , Yard , Text	2.992	4.940	0.471
12	Glass , Food , Yard	2.720	4.118	0.393
13	Paper , Food , Yard	2.371	4.002	0.382
14	Paper , Glass , Yard	2.533	3.950	0.377
15	Metal , Food , Yard	2.455	3.950	0.377
16	Plastic , Food , Yard	2.400	3.939	0.376
17	Plastic , Glass , Yard	2.562	3.827	0.365
18	Metal , Glass , Yard	2.617	3.815	0.364
19	Paper , Plastic , Yard	2.213	3.761	0.359
20	Metal , Paper , Yard	2.268	3.739	0.357
21	Metal , Plastic , Yard	2.297	3.679	0.351
22	Glass , Text , Other	2.984	3.571	0.341
23	Food , Text , Other	2.822	3.430	0.327
24	Paper , Text , Other	2.636	3.324	0.317
25	Metal , Text , Other	2.719	3.310	0.316
26	Glass , Food , Other	2.393	3.235	0.309
27	Paper , Glass , Other	2.206	3.185	0.304
28	Metal , Glass , Other	2.290	3.095	0.295
29	Plastic , Text , Other	2.665	3.024	0.289
30	Plastic , Glass , Other	2.235	2.927	0.279
31	Paper , Food , Other	2.044	2.873	0.274
32	Metal , Food , Other	2.127	2.865	0.273
33	Metal , Paper , Other	1.941	2.772	0.264
34	Plastic , Food , Other	2.073	2.674	0.255
35	Paper , Plastic , Other	1.886	2.614	0.249
36	Metal , Plastic , Other	1.970	2.577	0.246
37	Glass , Food , Text	2.266	1.857	0.178
38	Metal , Food , Text	2.001	1.677	0.160
39	Metal , Glass , Text	2.163	1.674	0.160
40	Paper , Glass , Text	2.080	1.666	0.159

	3-part Subcomposition	Sum of Coefficients of Variation	Total Variability	R^2
41	Paper , Food , Text	1.917	1.587	0.151
42	Paper , Plastic , Text	1.760	1.539	0.126
43	Plastic , Glass , Text	2.109	1.539	0.147
44	Plastic , Food , Text	1.947	1.520	0.145
45	Metal , Paper , Text	1.814	1.444	0.138
46	Metal , Plastic , Text	1.844	1.380	0.132
47	Metal , Glass , Food	1.571	0.732	0.070
48	Paper , Glass , Food	1.488	0.717	0.068
49	Plastic , Glass , Food	1.517	0.693	0.066
50	Metal , Paper , Glass	1.385	0.554	0.053
51	Paper , Plastic , Glass	1.330	0.547	0.052
52	Metal , Plastic , Glass	1.414	0.532	0.051
53	Metal , Plastic , Food	1.252	0.421	0.040
54	Metal , Paper , Food	1.223	0.386	0.037
55	Paper , Plastic , Food	1.168	0.357	0.034
56	Metal , Paper , Plastic	1.065	0.236	0.022

Table 16. Garbage Compositional Data: All 3-part subcompositions and the corresponding SCV, Total Variability and R^2

Table 17. All 4-part subcompositions and the corresponding SCV, Total Variability and R^2

	4-part Subcomposition	Sum of Coefficients of Variation	Total Variability	R^2
1	Food , Yard , Text , Other	4.372	8.128	0.776
2	Glass , Yard , Text , Other	4.534	8.102	0.773
3	Paper , Yard , Text , Other	4.185	7.974	0.761
4	Metal , Yard , Text , Other	4.269	7.888	0.753
5	Plastic , Yard , Text , Other	4.214	7.688	0.734
6	Glass , Food , Yard , Other	3.942	7.421	0.708
7	Paper , Glass , Yard , Other	3.755	7.309	0.697
8	Paper , Food , Yard , Other	3.593	7.207	0.688
9	Metal , Glass , Yard , Other	3.839	7.166	0.684
10	Metal , Food , Yard , Other	3.677	7.126	0.680
11	Plastic , Glass , Yard , Other	3.785	7.055	0.673
12	Plastic , Food , Yard , Other	3.622	6.998	0.668
13	Metal , Paper , Yard , Other	3.490	6.981	0.666
14	Paper , Plastic , Yard , Other	3.436	6.878	0.656
15	Metal , Plastic , Yard , Other	3.519	6.774	0.646
16	Glass , Food , Yard , Text	3.816	6.153	0.587
17	Paper , Food , Yard , Text	3.467	6.008	0.573
18	Metal , Food , Yard , Text	3.551	6.000	0.572
19	Paper , Glass , Yard , Text	3.629	5.935	0.566
20	Plastic , Food , Yard , Text	3.496	5.898	0.563
21	Metal , Glass , Yard , Text	3.713	5.865	0.560
22	Plastic , Glass , Yard , Text	3.658	5.779	0.551
23	Metal , Paper , Yard , Text	3.364	5.750	0.549
24	Paper , Plastic , Yard , Text	3.309	5.672	0.541
25	Metal , Plastic , Yard , Text	3.393	5.642	0.538
26	Paper , Glass , Food , Yard	3.037	4.795	0.458
27	Metal , Glass , Food , Yard	3.121	4.730	0.451
28	Plastic , Glass , Food , Yard	3.066	4.716	0.450
29	Glass , Food , Text , Other	3.488	4.535	0.433
30	Paper , Plastic , Glass , Yard	2.880	4.532	0.432
31	Metal , Paper , Food , Yard	2.772	4.529	0.432
32	Paper , Plastic , Food , Yard	2.718	4.522	0.431
33	Metal , Paper , Glass , Yard	2.934	4.522	0.431
34	Metal , Plastic , Food , Yard	2.801	4.497	0.429
35	Metal , Plastic , Glass , Yard	2.963	4.445	0.424
36	Paper , Glass , Text , Other	3.302	4.405	0.420
37	Metal , Glass , Text , Other	3.386	4.369	0.417
38	Metal , Paper , Plastic , Yard	2.615	4.281	0.408
39	Metal , Food , Text , Other	3.223	4.230	0.404
40	Paper , Food , Text , Other	3.140	4.205	0.401

	4-part Subcomposition	Sum of Coefficients of Variation	Total Variability	R^2
41	Plastic , Glass , Text , Other	3.331	4.148	0.396
42	Metal , Paper , Text , Other	3.037	4.068	0.388
43	Plastic , Food , Text , Other	3.169	3.993	0.381
44	Metal , Plastic , Text , Other	3.066	3.859	0.368
45	Paper , Plastic , Text , Other	2.982	3.855	0.368
46	Paper , Glass , Food , Other	2.710	3.753	0.358
47	Metal , Glass , Food , Other	2.794	3.722	0.355
48	Metal , Paper , Glass , Other	2.607	3.602	0.344
49	Plastic , Glass , Food , Other	2.739	3.573	0.341
50	Paper , Plastic , Glass , Other	2.552	3.477	0.332
51	Metal , Plastic , Glass , Other	2.636	3.424	0.327
52	Metal , Paper , Food , Other	2.445	3.336	0.318
53	Metal , Plastic , Food , Other	2.474	3.202	0.306
54	Paper , Plastic , Food , Other	2.390	3.194	0.305
55	Metal , Paper , Plastic , Other	2.287	3.074	0.293
56	Metal , Glass , Food , Text	2.667	2.227	0.212
57	Paper , Glass , Food , Text	2.585	2.185	0.208
58	Plastic , Glass , Food , Text	2.613	2.103	0.201
59	Metal , Paper , Glass , Text	2.481	2.001	0.191
60	Metal , Plastic , Glass , Text	2.510	1.922	0.183
61	Metal , Paper , Food , Text	2.319	1.910	0.182
62	Paper , Plastic , Glass , Text	2.426	1.901	0.181
63	Metal , Plastic , Food , Text	2.348	1.875	0.179
64	Paper , Plastic , Food , Text	2.264	1.794	0.171
65	Metal , Paper , Plastic , Text	2.161	1.642	0.157
66	Metal , Paper , Glass , Food	1.889	0.896	0.085
67	Metal , Plastic , Glass , Food	1.918	0.893	0.085
68	Paper , Plastic , Glass , Food	1.834	0.867	0.083
69	Metal , Paper , Plastic , Glass	1.731	0.700	0.067
70	Metal , Paper , Plastic , Food	1.569	0.526	0.050

Table 17. Garbage Compositional Data: All 4-part subcompositions and the corresponding SCV, Total Variability and R^2

Table 18. All 5-part subcompositions and the corresponding SCV, Total Variability and R^2

	5-part Subcomposition	Sum of Coefficients of Variation	Total Variability	R^2
1	Glass,Food,Yard,Text,Other	5.038	9.157	0.874
2	Paper,Glass,Yard,Text,Other	4.851	8.993	0.858
3	Paper,Food,Yard,Text,Other	4.689	8.939	0.853
4	Metal,Plastic,Yard,Text,Other	4.935	8.904	0.850
5	Metal,Food,Yard,Text,Other	4.773	8.899	0.849
6	Plastic,Glass,Yard,Text,Other	4.880	8.739	0.834
7	Plastic,Food,Yard,Text,Other	4.718	8.721	0.832
8	Metal,Paper,Yard,Text,Other	4.586	8.710	0.831
9	Paper,Plastic,Yard,Text,Other	4.532	8.551	0.816
10	Metal,Plastic,Yard,Text,Other	4.615	8.493	0.810
11	Paper,Glass,Food,Yard,Other	4.259	8.130	0.776
12	Metal,Glass,Food,Yard,Other	4.343	8.044	0.768
13	Plastic,Glass,Food,Yard,Other	4.289	7.937	0.757
14	Metal,Paper,Glass,Yard,Other	4.157	7.888	0.753
15	Paper,Plastic,Glass,Yard,Other	4.102	7.800	0.744
16	Metal,Paper,Food,Yard,Other	3.994	7.781	0.742
17	Metal,Plastic,Glass,Yard,Other	4.186	7.697	0.734
18	Paper,Plastic,Food,Yard,Other	3.940	7.680	0.733
19	Metal,Plastic,Food,Yard,Other	4.023	7.626	0.728
20	Metal,Paper,Plastic,Yard,Other	3.837	7.463	0.712
21	Paper,Glass,Food,Yard,Text	4.133	6.687	0.638
22	Metal,Glass,Food,Yard,Text	4.217	6.660	0.635
23	Plastic,Glass,Food,Yard,Text	4.162	6.573	0.627
24	Metal,Paper,Food,Yard,Text	3.868	6.453	0.616
25	Metal,Paper,Glass,Yard,Text	4.030	6.419	0.612
26	Metal,Plastic,Food,Yard,Text	3.897	6.376	0.608
27	Paper,Plastic,Food,Yard,Text	3.813	6.372	0.608
28	Paper,Plastic,Glass,Yard,Text	3.976	6.352	0.606
29	Metal,Plastic,Glass,Yard,Text	4.059	6.307	0.602
30	Metal,Paper,Plastic,Yard,Text	3.710	6.129	0.585
31	Metal,Paper,Glass,Food,Yard	3.438	5.193	0.495
32	Paper,Plastic,Glass,Food,Yard	3.384	5.182	0.494
33	Metal,Plastic,Glass,Food,Yard	3.468	5.142	0.491
34	Metal,Glass,Food,Text,Other	3.890	5.089	0.486
35	Paper,Glass,Food,Text,Other	3.806	5.089	0.486
36	Metal,Paper,Plastic,Glass,Yard	3.281	4.928	0.470
37	Metal,Paper,Glass,Text,Other	3.703	4.919	0.469
38	Metal,Paper,Plastic,Food,Yard	3.119	4.895	0.467
39	Plastic,Glass,Food,Text,Other	3.835	4.894	0.467
40	Paper,Plastic,Glass,Text,Other	3.648	4.743	0.453

	5-part Subcomposition	Sum of Coefficients of Variation	Total Variability	R^2
41	Metal,Paper,Food,Text,Other	3.541	4.733	0.452
42	Metal,Plastic,Glass,Text,Other	3.732	4.726	0.451
43	Metal,Plastic,Food,Text,Other	3.570	4.576	0.437
44	Paper,Plastic,Food,Text,Other	3.486	4.544	0.434
45	Metal,Paper,Plastic,Text,Other	3.383	4.400	0.420
46	Metal,Paper,Glass,Food,Other	3.111	4.082	0.390
47	Paper,Plastic,Glass,Food,Other	3.057	3.964	0.378
48	Metal,Plastic,Glass,Food,Other	3.140	3.950	0.377
49	Metal,Paper,Plastic,Glass,Other	2.954	3.807	0.363
50	Metal,Paper,Plastic,Food,Other	2.791	3.555	0.339
51	Metal,Paper,Glass,Food,Text	2.985	2.458	0.235
52	Metal,Plastic,Glass,Food,Text	3.014	2.405	0.229
53	Paper,Plastic,Glass,Food,Text	2.930	2.360	0.225
54	Metal,Paper,Plastic,Glass,Text	2.827	2.178	0.208
55	Metal,Paper,Plastic,Food,Text	2.665	2.066	0.197
56	Metal,Paper,Plastic,Glass,Food	2.235	1.035	0.099

Table 18. Garbage Compositional Data: All 5-part subcompositions and the corresponding SCV, Total Variability and R^2

CHAPTER 4

ZEROS IN COMPOSITIONAL DATA: A COMPARISON BETWEEN SUM OF COEFFICIENTS OF VARIATION AND COMPOSITIONAL TOTAL VARIABILITY

As we discussed earlier, logratio analysis of compositional data introduced by Aitchison is limited by the assumption of strictly positive components. Therefore, Total Variability and subcompositional analysis techniques based on the logratio transformations require complete data matrices, thus calling for a strategy of imputation of zeros. Different zero treatment strategies in compositional data and their advantages/disadvantages were discussed in Chapter 2. There is no agreement on one best strategy to handle zeros and this problem is unlikely ever to be satisfactory and generally resolved in analysis based on the logratio transformations (Aitchison 1986). However, using the new approach of measuring compositional data variability based on Sum of Coefficients of Variation finding informative subcompositions when zero observations are present does not require any special treatment or imputation. In this chapter we will examine the behavior of the Sum of Coefficients of Variation approach introduced in Chapter 3 and compositional Total Variability introduced by Aitchison (1986) in the existence of zeros.

We investigate the behavior of SCV method and Aitchison Total Variability in the existence of zeros using the following procedure: we replace 10%, 20%, 30%, and

40% of the observations in one variable with zeros at random and then we compare the changes in the Sum of Coefficients of Variation and Aitchison Total Variability of the subcompositions before and after zeros. Aitchison's approach based on logratios is applied after treating zeros with current zero treatment techniques introduced in Chapter 2. We illustrate these analysis with the Garbage data first using the component Food. We choose the variable Food because it has an intermediate variation between the eight components. We repeat the analysis using the two most extremist components in the data (with the smallest and largest variation), Paper and Yard.

Replacing 10% of the observations in the variable
Food with zeros

Table (19) presents Sum of Coefficients of Variation of all 3-part subcompositions and the corresponding Total Variability computed after employing the following zero treatment techniques introduced in Chapter 2:

1. Multiplicative Replacement (MR) with $r = 0.0001256881$
2. Multiplicative Replacement (MR) with $r = 0.001$
3. Aitchison Additive (AA) with $C = 1$, $D = 8$, $\delta = 0.005$, and $r = 0.001$
4. Aitchison Additive (AA) with $C = 1$, $D = 8$, $\delta = 0.00005$, and $r = 0.00001$
5. Alternative zero replacement (AZR) with $r = 0.0004859086$
6. Alternative zero replacement (AZR) with $r = 0.0001256881$
7. Replace zero and recalculate other (RZRO) with $r = 0.0001256881$
8. Rank across cases and variables (Rank)

As a result of the existence of zeros, the correlation coefficients between SCV and total variability of all 3-part subcompositions dropped from 0.95 in the original

data to 0.716, 0.886, 0.887, 0.452, 0.840, 0.716, 0.717, and 0.522 after replacing zeros across the eight zero treatment techniques respectively. In addition, Table (19) shows that order and the amount of Total Variability computed for all 3-part subcompositions change dramatically with 10% of zeros in the component Food. Correlation coefficients between original compositional Total Variability for 3-part subcompositions and the new Total Variability after replacing zeros are 0.707, 0.909, 0.911, 0.410, 0.852, 0.707, 0.708, 0.506 across the eight zero treatment techniques. Figures (32) and (33) show scatter plots of the original Total Variability and the new Total Variability after replacing zeros. There are two separate lines in these plots, the upper line represents all 3-part subcompositions that contain the component Food. This indicates that Aitchison's compositional Total Variability is extremely affected by the existence of zeros in the compositional data. From Table (16) in Chapter 3, the top five 3-part subcompositions with largest Total Variability in the original data with no zeros sorted in descending order are: (Yard, Text, Other), (Glass, Yard, Other), (Food, Yard, Other), (Paper, Yard, Other), and (Metal, Yard, Other). Out of these top five subcompositions, only one of them contains the variable Food. However, Table (19) below shows that in the top five 3-part subcompositions with largest Total Variability in the data after replacing zeros, the number of subcompositions that contain the replacement variable for Food increased to 5, 2, 2, 5, 3, 5, 5, and 5 using Aitchison's method across the eight zero treatment techniques respectively. Finally, boxplots in Figure (34) show Sum of Coefficients of Variation and Total Variability for all 3-part subcompositions that contain each Garbage component after replacing zeros. We selected two zero-replacement techniques (MR) and (AA). The graph shows that subcompositions that include the variable Food have the largest Total Variability which is a dramatic change from what we saw before in Figure (27) from chapter 3 where variable Food came in the fifth place.

In contrast, we don't see the same amount of change in the Sum of Coefficients of Variation before and after zeros. Correlation coefficient between SCV of all 3-part subcompositions of the original data and SCV of the new data after replacing zeros is 0.995 (see Figure (35)). In addition, the top five 3-part subcompositions with largest SCV in the original data with no zeros sorted in descending order are: (Yard, Text, Other), (Glass, Yard, Other), (Glass, Yard, Text), (Food, Yard, Other), and (Metal, Yard, Other) with only one subcomposition that includes the variable Food. The new top five subcompositions with largest SCV after replacing zeros are: (Yard, Text, Other), (Glass, Yard, Other), (Food, Yard, Other), (Glass, Yard, Text), and (Food, Yard, Text) with only two subcompositions that contains the variable Food. Boxplot in Figure (34) shows that even after replacing Food with 10% zeros, it stayed in the fifth place.

Table 19. Sum of Coefficients of Variation and Total Variability after replacing 10% of the observations in the variable Food with zeros

	3-part Subcomposition	SCV	MR	MR	AA	AA	AA	AZR	AZR	RZRO	Rank
			$r =$ $1.3E^{-4}$	$r =$ $1.0E^{-3}$	$r =$ $1.0E^{-3}$	$r =$ $1.0E^{-5}$	$r =$ $4.9E^{-4}$	$r =$ $1.3E^{-4}$	$r =$ $1.3E^{-4}$	$r =$ $1.3E^{-4}$	
1	Metal, Paper, Plastic	1.063	0.236	0.236	0.236	0.236	0.236	0.236	0.236	0.236	0.105
2	Metal, Paper, Glass	1.409	0.554	0.554	0.554	0.554	0.554	0.554	0.554	0.554	0.159
3	Metal, Paper, Food	1.362	3.434	1.926	1.925	5.965	2.393	3.434	3.434	3.434	1.366
4	Metal, Paper, Yard	2.268	3.739	3.739	3.764	3.739	3.739	3.739	3.739	3.739	1.059
5	Metal, Paper, Text	1.817	1.444	1.444	1.450	1.444	1.444	1.444	1.444	1.444	0.315
6	Metal, Paper, Other	1.973	2.772	2.772	2.795	2.772	2.772	2.772	2.772	2.792	0.678
7	Metal, Plastic, Glass	1.426	0.532	0.532	0.532	0.532	0.532	0.532	0.532	0.532	0.203
8	Metal, Plastic, Food	1.379	3.430	1.934	1.933	5.947	2.3960	3.430	3.430	3.430	1.410
9	Metal, Plastic, Yard	2.285	3.679	3.679	3.701	3.679	3.679	3.679	3.679	3.679	1.118
10	Metal, Plastic, Text	1.834	1.380	1.380	1.385	1.380	1.380	1.380	1.380	1.380	0.348
11	Metal, Plastic, Other	1.989	2.577	2.577	2.598	2.577	2.577	2.577	2.577	2.594	0.677
12	Metal, Glass, Food	1.718	3.919	2.366	2.365	6.505	2.848	3.919	3.919	3.919	1.518
13	Metal, Glass, Yard	2.624	3.815	3.815	3.839	3.816	3.815	3.815	3.815	3.815	1.128
14	Metal, Glass, Text	2.173	1.674	1.674	1.68	1.674	1.674	1.674	1.674	1.674	0.397
15	Metal, Glass, Other	2.329	3.095	3.095	3.116	3.095	3.095	3.095	3.095	3.113	0.792
16	Metal, Food, Yard	2.577	7.009	5.485	5.488	9.558	5.957	7.008	7.008	7.009	2.441
17	Metal, Food, Text	2.126	4.638	3.153	3.150	7.141	3.612	4.638	4.638	4.638	1.617
18	Metal, Food, Other	2.282	6.131	4.560	4.568	8.739	5.048	6.131	6.131	6.134	2.076
19	Metal, Yard, Text	3.033	5.046	5.046	5.065	5.046	5.046	5.046	5.046	5.046	1.354
20	Metal, Yard, Other	3.188	5.974	5.974	6.005	5.975	5.974	5.974	5.974	5.987	1.657

	3-part Subcomposition	SCV	MR $r = 1.3E^{-4}$	MR $r = 1.0E^{-3}$	AA $r = 1.0E^{-3}$	AA $r = 1.0E^{-5}$	AZR $r = 4.9E^{-4}$	AZR $r = 1.3E^{-4}$	RZRO $r = 1.3E^{-4}$	Rank
21	Metal, Text, Other	2.737	3.310	3.310	3.329	3.310	3.310	3.310	3.324	0.827
22	Paper, Plastic, Glass	1.344	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.158
23	Paper, Plastic, Food	1.297	3.311	1.831	1.830	5.808	2.288	3.311	3.311	1.327
24	Paper, Plastic, Yard	2.203	3.761	3.761	3.783	3.761	3.761	3.761	3.761	1.069
25	Paper, Plastic, Text	1.752	1.319	1.319	1.324	1.319	1.319	1.319	1.319	0.285
26	Paper, Plastic, Other	1.908	2.614	2.614	2.634	2.614	2.614	2.614	2.631	0.625
27	Paper, Glass, Food	1.636	3.853	2.316	2.315	6.419	2.793	3.853	3.853	1.439
28	Paper, Glass, Yard	2.543	3.950	3.950	3.974	3.95	3.950	3.950	3.950	1.082
29	Paper, Glass, Text	2.091	1.666	1.666	1.672	1.666	1.666	1.666	1.666	0.338
30	Paper, Glass, Other	2.247	3.185	3.185	3.205	3.185	3.185	3.185	3.202	0.744
31	Paper, Food, Yard	2.495	7.009	5.502	5.505	9.539	5.968	7.009	7.009	2.356
32	Paper, Food, Text	2.044	4.496	3.027	3.024	6.980	3.480	4.496	4.496	1.519
33	Paper, Food, Other	2.20	6.087	4.531	4.540	8.676	5.015	6.087	6.090	1.990
34	Paper, Yard, Text	2.951	5.105	5.105	5.124	5.105	5.105	5.105	5.105	1.290
35	Paper, Yard, Other	3.107	6.131	6.131	6.161	6.131	6.131	6.131	6.144	1.604
36	Paper, Text, Other	2.655	3.324	3.324	3.343	3.324	3.324	3.324	3.338	0.761
37	Plastic, Glass, Food	1.653	3.787	2.261	2.260	6.339	2.734	3.786	3.787	1.474
38	Plastic, Glass, Yard	2.559	3.827	3.827	3.849	3.828	3.827	3.827	3.827	1.133
39	Plastic, Glass, Text	2.108	1.539	1.539	1.544	1.539	1.539	1.539	1.539	0.362
40	Plastic, Glass, Other	2.264	2.927	2.927	2.945	2.927	2.927	2.927	2.943	0.734

	3-part Subcomposition	SCV	MR $r = 1.3E^{-4}$	MR $r = 1.0E^{-3}$	AA $r = 1.0E^{-3}$	AA $r = 1.0E^{-5}$	AZR $r = 4.9E^{-4}$	AZR $r = 1.3E^{-4}$	RZRO $r = 1.3E^{-4}$	Rank
41	Plastic, Food, Yard	2.512	6.904	5.409	5.409	9.420	5.871	6.904	6.904	2.408
42	Plastic, Food, Text	2.061	4.388	2.930	2.926	6.857	3.379	4.388	4.388	1.544
43	Plastic, Food, Other	2.217	5.847	4.303	4.309	8.422	4.782	5.847	5.848	1.980
44	Plastic, Yard, Text	2.968	4.940	4.940	4.956	4.940	4.940	4.940	4.940	1.330
45	Plastic, Yard, Other	3.123	5.835	5.835	5.860	5.835	5.835	5.835	5.846	1.610
46	Plastic, Text, Other	2.672	3.024	3.024	3.039	3.024	3.024	3.024	3.036	0.740
47	Glass, Food, Yard	2.851	7.265	5.713	5.715	9.850	6.195	7.265	7.265	2.475
48	Glass, Food, Text	2.40	4.906	3.392	3.388	7.444	3.860	4.906	4.906	1.650
49	Glass, Food, Other	2.556	6.589	4.989	4.995	9.233	5.488	6.589	6.591	2.153
50	Glass, Yard, Text	3.307	5.105	5.105	5.123	5.106	5.105	5.105	5.105	1.338
51	Glass, Yard, Other	3.463	6.224	6.224	6.251	6.225	6.224	6.224	6.235	1.684
52	Glass, Text, Other	3.011	3.571	3.571	3.587	3.571	3.571	3.571	3.583	0.853
53	Food, Yard, Text	3.259	8.249	6.764	6.753	10.750	7.223	8.248	8.249	2.625
54	Food, Yard, Other	3.415	9.439	7.868	7.864	12.046	8.357	9.439	9.436	3.036
55	Food, Text, Other	2.964	6.558	5.025	5.020	9.118	5.500	6.557	6.555	2.112
56	Yard, Text, Other	3.871	6.704	6.704	6.719	6.704	6.704	6.704	6.711	1.770

Table 19. Garbage Compositional Data: All 3-part subcompositions and the corresponding Sum of Coefficients of Variation and Total Variability after replacing 10% of the observations in the variable Food with zeros.

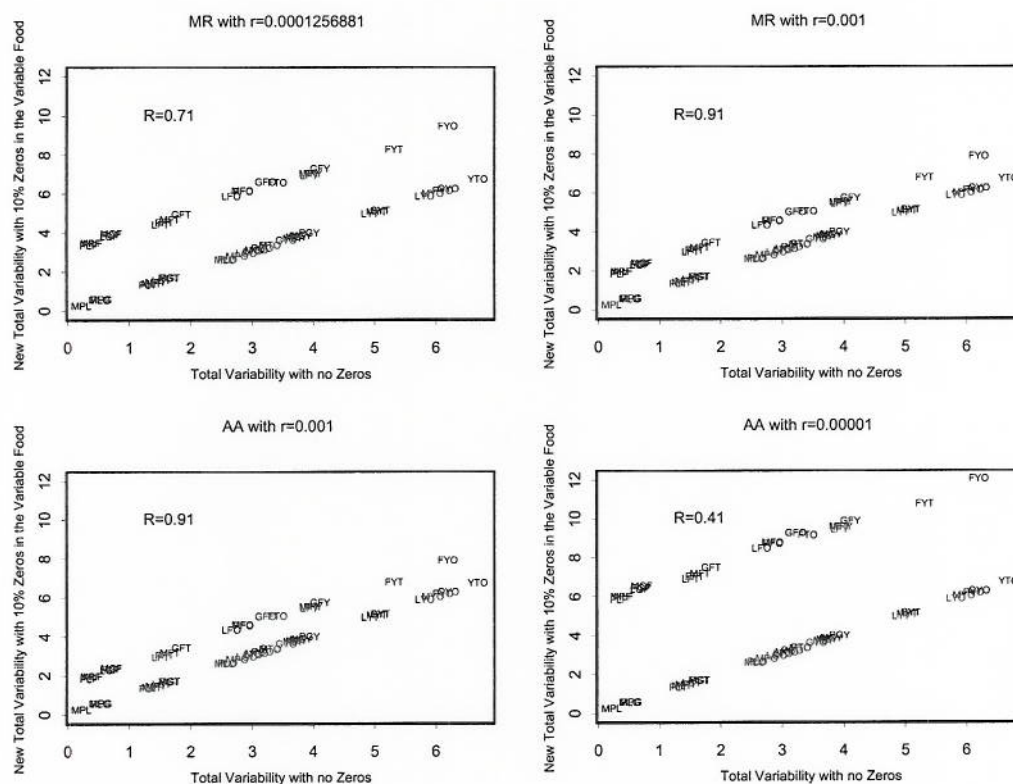


Figure 32. Changes in Total Variability after replacing 10% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

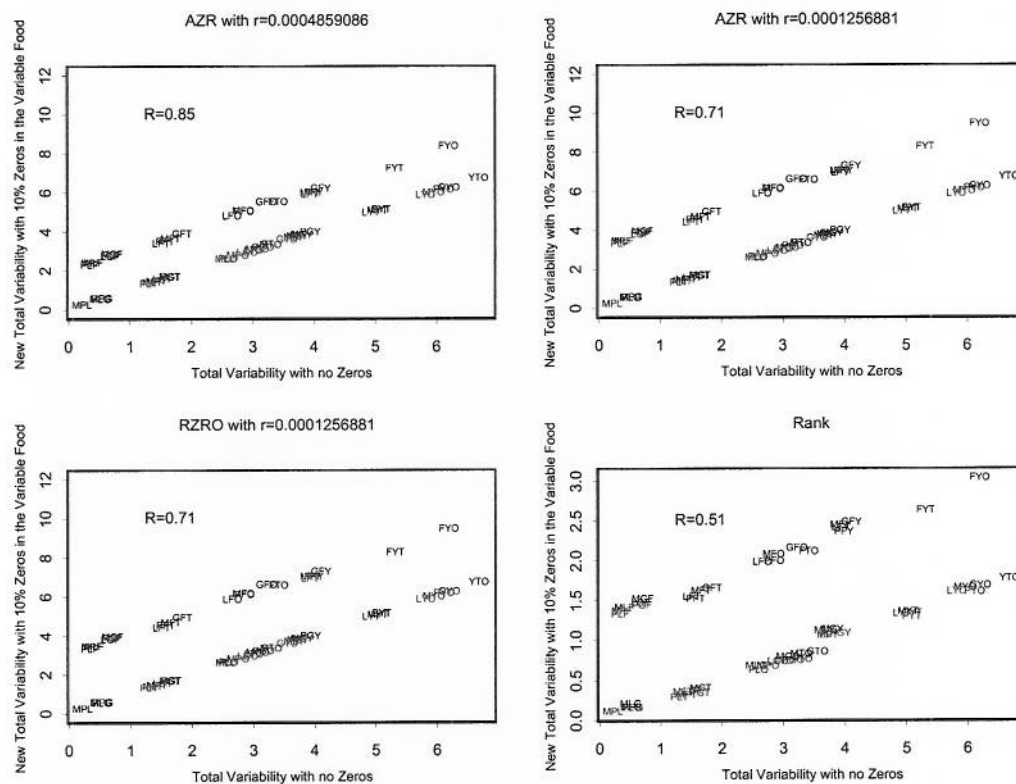


Figure 33. Changes in Total Variability after replacing 10% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

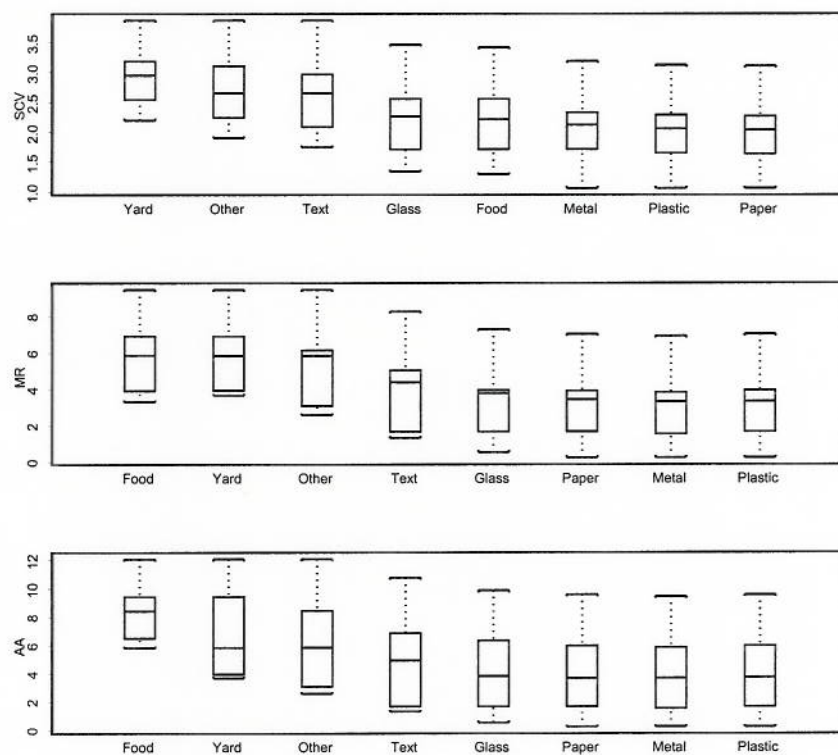


Figure 34. Distribution of the SCV and Total Variability of all 3-part subcompositions that contain each Garbage component after replacing 10% of the observations in the variable Food with zeros.

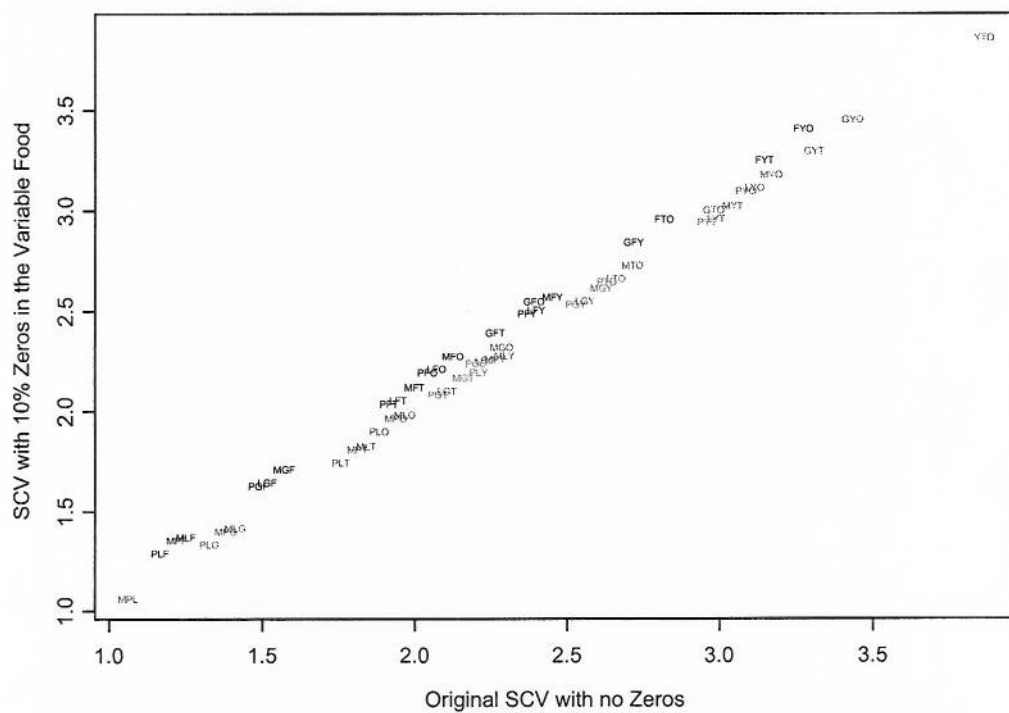


Figure 35. Changes in the Sum of Coefficients of Variation after replacing 10% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

**Replacing 10% of the observations in the variable
Paper with zeros**

The component Paper has the largest mean and smallest Coefficient of Variation in the Garbage compositional data. After replacing 10% of the observations in the variable Paper with zeros, correlation coefficients between Sum of Coefficients of Variation and Total Variability of all 3-part subcompositions dropped to 0.451, 0.737, 0.745, 0.157, 0.643, 0.451, 0.451, and 0.272 across the eight zero treatment techniques respectively. In addition, Table (20) shows that the amount and order of Total Variability computed for all 3-part subcompositions changes dramatically with 10% of zeros in the data and even more than the changes occurred after replacing 10% of the observations in the variable Food with zeros. Correlation coefficients between original compositional Total Variability for 3-part subcompositions and the new Total Variability after replacing zeros ranged between 0.230 and 0.810 . In the original data with no zeros, out of the top five subcompositions with largest Total Variability, only one of them contains the variable Paper. However, Table (20) below shows that of the top five 3-part subcompositions that retained most of the variability in the data after replacing zeros, number of subcompositions that includes the variable Paper increased to 5, 2, 2, 5, 4, 5, 5, and 5 across the eight techniques respectively.

In contrast, we don't see the same amount of change in the SCV before and after zeros. The correlation coefficient between SCV of all 3-part subcompositions of the original data and SCV of the new data after replacing zeros is 0.992. In addition, None of the top five 3-part subcompositions with largest Sum of Coefficients of Variation include the variable Paper which is comparable to the results in the original data with no zeros.

Table 20. Sum of Coefficients of Variation and Total Variability after replacing 10% of the observations in the variable Paper with zeros

	3-part Subcomposition	SCV	MR $r =$ $1.3E^{-4}$	MR $r =$ $1.0E^{-3}$	AA $r =$ $1.0E^{-3}$	AA $r =$ $1.0E^{-5}$	AZR $r =$ $4.9E^{-4}$	AZR $r =$ $1.3E^{-4}$	RZRO $r =$ $1.3E^{-4}$	Rank
1	Metal, Paper, Plastic	1.447	4.699	2.821	2.82	7.683	3.416	4.699	4.699	1.782
2	Metal, Paper, Glass	1.667	5.054	3.165	3.164	8.051	3.764	5.054	5.054	1.804
3	Metal, Paper, Food	1.490	4.613	2.791	2.791	7.528	3.367	4.613	4.613	1.685
4	Metal, Paper, Yard	2.570	7.419	5.722	5.732	10.181	6.254	7.419	7.419	2.495
5	Metal, Paper, Text	2.055	5.715	3.888	3.887	8.634	4.465	5.714	5.715	1.882
6	Metal, Paper, Other	2.174	6.937	5.137	5.137	9.824	5.705	6.937	6.935	2.281
7	Metal, Plastic, Glass	1.652	0.532	0.532	0.532	0.532	0.532	0.532	0.532	0.205
8	Metal, Plastic, Food	1.475	0.424	0.424	0.424	0.424	0.424	0.424	0.424	0.168
9	Metal, Plastic, Yard	2.554	3.679	3.679	3.746	3.680	3.679	3.679	3.679	1.109
10	Metal, Plastic, Text	2.039	1.380	1.380	1.392	1.380	1.380	1.38	1.380	0.344
11	Metal, Plastic, Other	2.159	2.577	2.577	2.583	2.577	2.577	2.577	2.582	0.710
12	Metal, Glass, Food	1.695	0.732	0.732	0.732	0.732	0.732	0.732	0.732	0.221
13	Metal, Glass, Yard	2.775	3.815	3.815	3.884	3.816	3.815	3.815	3.815	1.118
14	Metal, Glass, Text	2.260	1.674	1.674	1.687	1.674	1.674	1.674	1.674	0.397
15	Metal, Glass, Other	2.379	3.095	3.095	3.102	3.095	3.095	3.095	3.101	0.829
16	Metal, Food, Yard	2.598	3.950	3.950	4.015	3.950	3.950	3.950	3.950	1.143
17	Metal, Food, Text	2.083	1.677	1.677	1.690	1.677	1.677	1.677	1.677	0.392
18	Metal, Food, Other	2.202	2.865	2.865	2.871	2.865	2.865	2.865	2.870	0.770
19	Metal, Yard, Text	3.163	5.046	5.046	5.133	5.047	5.046	5.046	5.046	1.344
20	Metal, Yard, Other	3.282	5.974	5.974	6.036	5.975	5.974	5.974	5.977	1.693

	3-part Subcomposition	SCV	MR $r = 1.3E^{-4}$	MR $r = 1.0E^{-3}$	AA $r = 1.0E^{-3}$	AA $r = 1.0E^{-5}$	AZR $r = 4.9E^{-4}$	AZR $r = 1.3E^{-4}$	RZRO $r = 1.3E^{-4}$	Rank
21	Metal, Text, Other	2.767	3.310	3.310	3.326	3.310	3.310	3.310	3.316	0.858
22	Paper, Plastic, Glass	1.612	5.026	3.141	3.140	8.016	3.739	5.026	5.026	1.814
23	Paper, Plastic, Food	1.435	4.562	2.745	2.744	7.471	3.319	4.562	4.562	1.695
24	Paper, Plastic, Yard	2.515	7.419	5.726	5.735	10.175	6.257	7.419	7.419	2.526
25	Paper, Plastic, Text	1.999	5.568	3.746	3.745	8.482	4.322	5.568	5.568	1.863
26	Paper, Plastic, Other	2.119	6.757	4.962	4.961	9.638	5.528	6.757	6.755	2.238
27	Paper, Glass, Food	1.656	4.960	3.132	3.131	7.881	3.710	4.960	4.960	1.715
28	Paper, Glass, Yard	2.735	7.645	5.942	5.953	10.415	6.476	7.645	7.645	2.501
29	Paper, Glass, Text	2.220	5.952	4.120	4.119	8.879	4.699	5.952	5.952	1.883
30	Paper, Glass, Other	2.340	7.365	5.559	5.560	10.26	6.129	7.365	7.364	2.323
31	Paper, Food, Yard	2.558	7.424	5.788	5.795	10.111	6.299	7.424	7.424	2.444
32	Paper, Food, Text	2.043	5.600	3.835	3.833	8.444	4.391	5.600	5.600	1.797
33	Paper, Food, Other	2.163	6.779	5.041	5.041	9.591	5.587	6.779	6.778	2.183
34	Paper, Yard, Text	3.123	8.570	6.930	6.946	11.262	7.442	8.570	8.570	2.637
35	Paper, Yard, Other	3.242	9.49	7.876	7.874	12.150	8.380	9.490	9.486	2.994
36	Paper, Text, Other	2.727	7.274	5.531	5.528	10.092	6.080	7.274	7.273	2.263
37	Plastic, Glass, Food	1.640	0.693	0.692	0.693	0.693	0.692	0.693	0.693	0.210
38	Plastic, Glass, Yard	2.720	3.827	3.827	3.895	3.828	3.827	3.827	3.827	1.127
39	Plastic, Glass, Text	2.205	1.539	1.539	1.553	1.540	1.539	1.539	1.539	0.356
40	Plastic, Glass, Other	2.324	2.927	2.927	2.934	2.927	2.927	2.927	2.933	0.764

	3-part Subcomposition	SCV	MR $r = 1.3E^{-4}$	MR $r = 1.0E^{-3}$	AA $r = 1.0E^{-3}$	AA $r = 1.0E^{-5}$	AZR $r = 4.9E^{-4}$	AZR $r = 1.3E^{-4}$	RZRO $r = 1.3E^{-4}$	Rank
41	Plastic, Food, Yard	2.543	3.939	3.939	4.003	3.94	3.939	3.939	3.939	1.152
42	Plastic, Food, Text	2.028	1.520	1.520	1.533	1.520	1.520	1.520	1.520	0.351
43	Plastic, Food, Other	2.147	2.674	2.674	2.681	2.674	2.674	2.674	2.680	0.706
44	Plastic, Yard, Text	3.107	4.940	4.940	5.026	4.941	4.940	4.940	4.940	1.323
45	Plastic, Yard, Other	3.227	5.835	5.835	5.894	5.835	5.835	5.835	5.838	1.649
46	Plastic, Text, Other	2.712	3.024	3.024	3.041	3.024	3.024	3.024	3.030	0.765
47	Glass, Food, Yard	2.763	4.118	4.118	4.184	4.119	4.118	4.118	4.118	1.158
48	Glass, Food, Text	2.248	1.857	1.857	1.870	1.857	1.857	1.857	1.857	0.402
49	Glass, Food, Other	2.367	3.235	3.235	3.242	3.235	3.235	3.235	3.241	0.822
50	Glass, Yard, Text	3.328	5.105	5.105	5.194	5.106	5.105	5.105	5.105	1.329
51	Glass, Yard, Other	3.447	6.224	6.224	6.287	6.225	6.224	6.224	6.228	1.720
52	Glass, Text, Other	2.932	3.571	3.571	3.590	3.572	3.571	3.571	3.578	0.880
53	Food, Yard, Text	3.151	5.328	5.328	5.412	5.329	5.328	5.328	5.328	1.387
54	Food, Yard, Other	3.270	6.214	6.214	6.272	6.214	6.214	6.214	6.217	1.723
55	Food, Text, Other	2.755	3.430	3.430	3.447	3.430	3.430	3.430	3.436	0.855
56	Yard, Text, Other	3.835	6.704	6.704	6.782	6.704	6.704	6.704	6.707	1.807

Table 20. Garbage Compositional Data: All 3-part subcompositions and the corresponding Sum of Coefficients of Variation and total variability after replacing 10% of the observations in the variable Paper with zeros.

Replacing 10% of the observations in the variable
Yard with zeros

The variable Yard has the largest Coefficient of Variation. As expected, slight changes happened in the Sum of Coefficients of Variation and the compositional Total Variability after replacing 10% of the observations in the variable Yard with zeros. Correlation coefficients between SCV and Total Variability of all 3-part subcompositions after replacing zeros are 0.959, 0.963, 0.964, 0.931, 0.964, 0.959, 0.959, and 0.922 across the eight zero treatment techniques respectively. In addition, correlation coefficients between the original Total Variability for 3-part subcomposition and the new Total Variability after zeros ranged between 0.963 and 0.998 across the eight methods. From Table (16) in Chapter 3, all top five subcompositions with largest Sum of Coefficients of Variations or largest Total Variability contains the variable Yard. Same results we found after replacing 10% of the observations in the variable Yard with zeros as shown in Table (21)

Table 21. Sum of Coefficients of Variation and Total Variability after replacing 10% of the observations in the variable Yard with zeros

	3-part Subcomposition	SCV	MR $r =$ $1.3E^{-4}$	MR $r =$ $1.0E^{-3}$	AA $r =$ $1.0E^{-3}$	AA $r =$ $1.0E^{-5}$	AZR $r =$ $4.9E^{-4}$	AZR $r =$ $1.3E^{-4}$	RZRO $r =$ $1.3E^{-4}$	Rank
1	Metal, Paper, Plastic	1.066	0.236	0.236	0.236	0.236	0.236	0.236	0.236	0.105
2	Metal, Paper, Glass	1.379	0.554	0.554	0.554	0.554	0.554	0.554	0.554	0.162
3	Metal, Paper, Food	1.213	0.386	0.386	0.386	0.386	0.386	0.386	0.386	0.117
4	Metal, Paper, Yard	2.554	4.330	3.516	3.515	6.014	3.741	4.330	4.330	1.488
5	Metal, Paper, Text	1.804	1.444	1.444	1.462	1.444	1.444	1.444	1.444	0.321
6	Metal, Paper, Other	1.932	2.772	2.772	2.781	2.772	2.772	2.772	2.780	0.701
7	Metal, Plastic, Glass	1.418	0.532	0.532	0.532	0.532	0.532	0.532	0.532	0.204
8	Metal, Plastic, Food	1.252	0.424	0.424	0.424	0.424	0.424	0.424	0.424	0.171
9	Metal, Plastic, Yard	2.593	4.338	3.501	3.500	6.050	3.734	4.338	4.338	1.583
10	Metal, Plastic, Text	1.843	1.380	1.380	1.398	1.380	1.380	1.380	1.380	0.356
11	Metal, Plastic, Other	1.971	2.577	2.577	2.586	2.577	2.577	2.577	2.585	0.700
12	Metal, Glass, Food	1.566	0.732	0.732	0.732	0.732	0.732	0.732	0.732	0.222
13	Metal, Glass, Yard	2.906	4.513	3.667	3.667	6.235	3.904	4.513	4.513	1.584
14	Metal, Glass, Text	2.157	1.674	1.674	1.694	1.674	1.674	1.674	1.674	0.403
15	Metal, Glass, Other	2.284	3.095	3.095	3.106	3.095	3.095	3.095	3.104	0.816
16	Metal, Food, Yard	2.740	4.549	3.726	3.725	6.243	3.954	4.549	4.549	1.588
17	Metal, Food, Text	1.990	1.677	1.677	1.693	1.677	1.677	1.677	1.677	0.401
18	Metal, Food, Other	2.118	2.865	2.865	2.874	2.865	2.865	2.865	2.873	0.762
19	Metal, Yard, Text	3.331	5.238	4.532	4.542	6.791	4.720	5.238	5.238	1.669
20	Metal, Yard, Other	3.459	6.226	5.524	5.530	7.772	5.711	6.226	6.227	2.011

	3-part Subcomposition	SCV	MR		MR		AA		AA		AZR		AZR		RZRO		Rank
			$r = 1.3E^{-4}$	$r = 1.0E^{-3}$	$r = 1.0E^{-3}$	$r = 1.0E^{-3}$	$r = 1.0E^{-5}$	$r = 4.9E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	$r = 1.3E^{-4}$	
21	Metal, Text, Other	2.709	3.310	3.310	3.327	3.310	3.310	3.310	3.310	3.310	3.310	3.310	3.310	3.310	3.316	3.316	0.852
22	Paper, Plastic, Glass	1.331	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.547	0.157
23	Paper, Plastic, Food	1.165	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.114
24	Paper, Plastic, Yard	2.505	4.378	3.548	3.547	3.547	6.082	3.779	4.378	4.378	4.378	4.378	4.378	4.378	4.378	4.378	1.509
25	Paper, Plastic, Text	1.756	1.319	1.319	1.336	1.319	1.319	1.319	1.319	1.319	1.319	1.319	1.319	1.319	1.319	1.319	0.290
26	Paper, Plastic, Other	1.883	2.614	2.614	2.623	2.614	2.614	2.614	2.614	2.614	2.614	2.614	2.614	2.614	2.622	2.622	0.644
27	Paper, Glass, Food	1.478	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.717	0.167
28	Paper, Glass, Yard	2.819	4.605	3.767	3.767	3.767	6.319	4.001	4.605	4.605	4.605	4.605	4.605	4.605	4.605	4.605	1.513
29	Paper, Glass, Text	2.069	1.666	1.666	1.686	1.666	1.666	1.666	1.666	1.666	1.666	1.666	1.666	1.666	1.666	1.666	0.339
30	Paper, Glass, Other	2.197	3.185	3.185	3.196	3.185	3.185	3.185	3.185	3.185	3.185	3.185	3.185	3.185	3.194	3.194	0.761
31	Paper, Food, Yard	2.653	4.559	3.743	3.743	3.743	6.245	3.969	4.559	4.559	4.559	4.559	4.559	4.559	4.559	4.559	1.506
32	Paper, Food, Text	1.903	1.587	1.587	1.603	1.587	1.587	1.587	1.587	1.587	1.587	1.587	1.587	1.587	1.587	1.587	0.326
33	Paper, Food, Other	2.030	2.873	2.873	2.882	2.873	2.873	2.873	2.873	2.873	2.873	2.873	2.873	2.873	2.881	2.881	0.697
34	Paper, Yard, Text	3.244	5.255	4.556	4.566	4.556	6.799	4.742	5.255	5.255	5.255	5.255	5.255	5.255	5.255	5.255	1.578
35	Paper, Yard, Other	3.371	6.340	5.646	5.652	5.646	7.878	5.830	6.340	6.340	6.340	6.340	6.340	6.340	6.342	6.342	1.930
36	Paper, Text, Other	2.622	3.324	3.324	3.341	3.324	3.324	3.324	3.324	3.324	3.324	3.324	3.324	3.324	3.330	3.330	0.779
37	Plastic, Glass, Food	1.517	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.693	0.210
38	Plastic, Glass, Yard	2.858	4.551	3.690	3.689	3.690	6.293	3.932	4.551	4.551	4.551	4.551	4.551	4.551	4.551	4.551	1.597
39	Plastic, Glass, Text	2.108	1.539	1.539	1.559	1.539	1.540	1.539	1.539	1.539	1.539	1.539	1.539	1.539	1.539	1.539	0.363
40	Plastic, Glass, Other	2.236	2.927	2.927	2.938	2.927	2.927	2.927	2.927	2.927	2.927	2.927	2.927	2.927	2.936	2.936	0.750

	3-part Subcomposition	SCV	MR $r = 1.3E^{-4}$	MR $r = 1.0E^{-3}$	AA $r = 1.0E^{-3}$	AA $r = 1.0E^{-5}$	AZR $r = 4.9E^{-4}$	AZR $r = 1.3E^{-4}$	RZRO $r = 1.3E^{-4}$	Rank
41	Plastic, Food, Yard	2.692	4.565	3.725	3.725	6.279	3.959	4.565	4.565	1.602
42	Plastic, Food, Text	1.942	1.520	1.520	1.536	1.520	1.520	1.520	1.520	0.363
43	Plastic, Food, Other	2.069	2.674	2.674	2.683	2.674	2.674	2.674	2.682	0.698
44	Plastic, Yard, Text	3.283	5.159	4.437	4.446	6.731	4.630	5.159	5.159	1.655
45	Plastic, Yard, Other	3.410	6.113	5.395	5.400	7.678	5.587	6.112	6.114	1.971
46	Plastic, Text, Other	2.661	3.024	3.024	3.040	3.024	3.024	3.024	3.030	0.760
47	Glass, Food, Yard	3.005	4.782	3.934	3.934	6.506	4.171	4.782	4.782	1.600
48	Glass, Food, Text	2.256	1.857	1.857	1.875	1.857	1.857	1.857	1.857	0.406
49	Glass, Food, Other	2.383	3.235	3.235	3.245	3.235	3.235	3.235	3.244	0.810
50	Glass, Yard, Text	3.596	5.363	4.632	4.644	6.945	4.829	5.363	5.363	1.650
51	Glass, Yard, Other	3.724	6.541	5.815	5.822	8.117	6.010	6.540	6.543	2.034
52	Glass, Text, Other	2.974	3.571	3.571	3.591	3.572	3.571	3.571	3.578	0.869
53	Food, Yard, Text	3.430	5.487	4.779	4.786	7.042	4.968	5.487	5.487	1.685
54	Food, Yard, Other	3.557	6.431	5.728	5.733	7.980	5.915	6.431	6.433	2.018
55	Food, Text, Other	2.808	3.430	3.43	3.443	3.43	3.430	3.430	3.435	0.847
56	Yard, Text, Other	4.149	6.515	5.929	5.933	7.921	6.075	6.515	6.515	1.977

Table 21. Garbage Compositional Data: All 3-part subcompositions and the corresponding Sum of Coefficients of Variation and total variability after replacing 10% of the observations in the variable Yard with zeros.

**Replacing 20%, 30%, or 40% of the observations
with zeros**

**Replacing 20%, 30%, or 40% of the observations
in the variable Food with zeros**

Table (22) presents the number of 3-part subcompositions out of the top five with largest Sum of Coefficients of Variation and compositional Total Variability that include the variable Food after replacing 10%, 20%, 30%, and 40% of the observations in the variable Food with zeros. At 20% zeros in the data, all top five 3-part subcompositions with largest Total Variability include the variable Food. In contrast, using Sum of Coefficients of Variation technique, only two out of the top five include the variable Food. The same results were found when we replaced 30% of the observations with zeros. At 40%, Sum of Coefficients of Variations gave three out of the top five that include the variable Food while all top five 3-part subcompositions with largest Total Variability include the variable Food. Figure (36) summarizes the number of 3-part subcompositions out of the original top five remain in the top five subcompositions after replacing 10%, 20%, 30%, and 40% of the observations in the variable Food with zeros using compositional Total Variability and Sum of Coefficients of Variation. For example at 10% zeros, Sum of Coefficients of Variation keeps four in the top five subcompositions compared to only one subcomposition remains in the top five using the Total Variability in five of the eight techniques. The only subcomposition that remains in the top five with targets Total Variability is the one that includes the component Food. At 20% replacement, Sum of Coefficients of Variation keeps four in top five subcompositions compared to only one subcomposition using Total Variability in all zero replacement techniques and similarly at 30%. Finally, at 40% replacement Sum of Coefficients of Variation keeps three in top five subcompositions compared to one subcomposition using Total Variability in

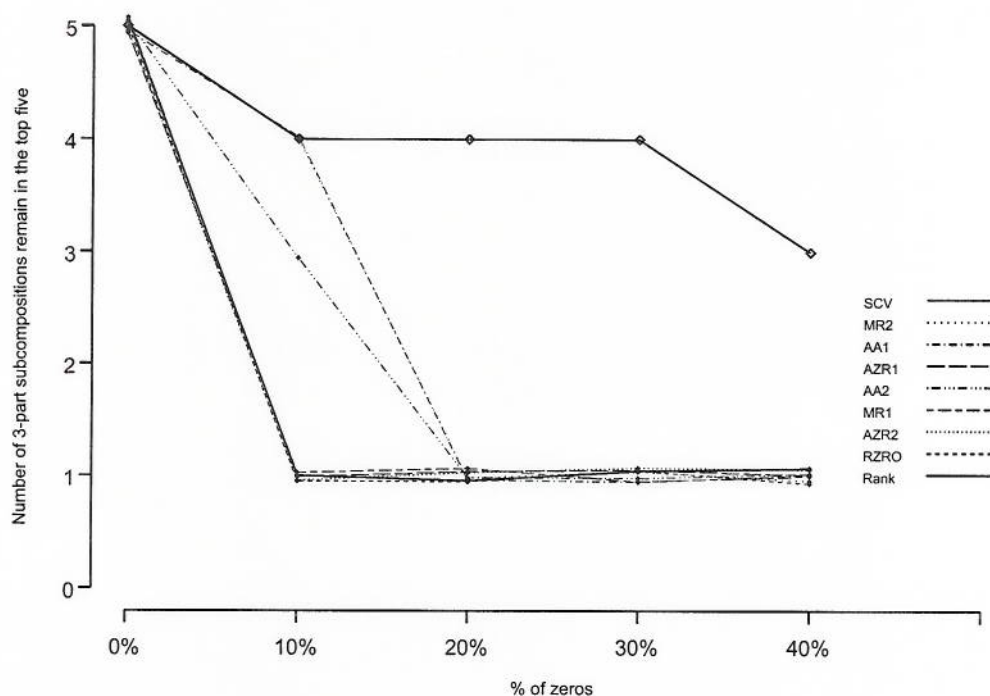


Figure 36. Number of 3-part subcompositions remain in the top five after replacing 10%, 20%, 30%, and 40% of the observations in the variable Food with zeros

all zero replacement techniques. Figures (37) and (38) display scatter plots of the original Total Variability and the new Total Variability after replacing 20% zeros and Figures (39) and (40) display scatter plots of the original Total Variability and the new Total Variability after replacing 30% zeros. All the plots are given with the same scale except for the Rank technique. Figures (41) and (42) display scatter plots of the original Sum of Coefficients of Variation and the new Sum of Coefficients of Variation after replacing 20% and 30% zeros.

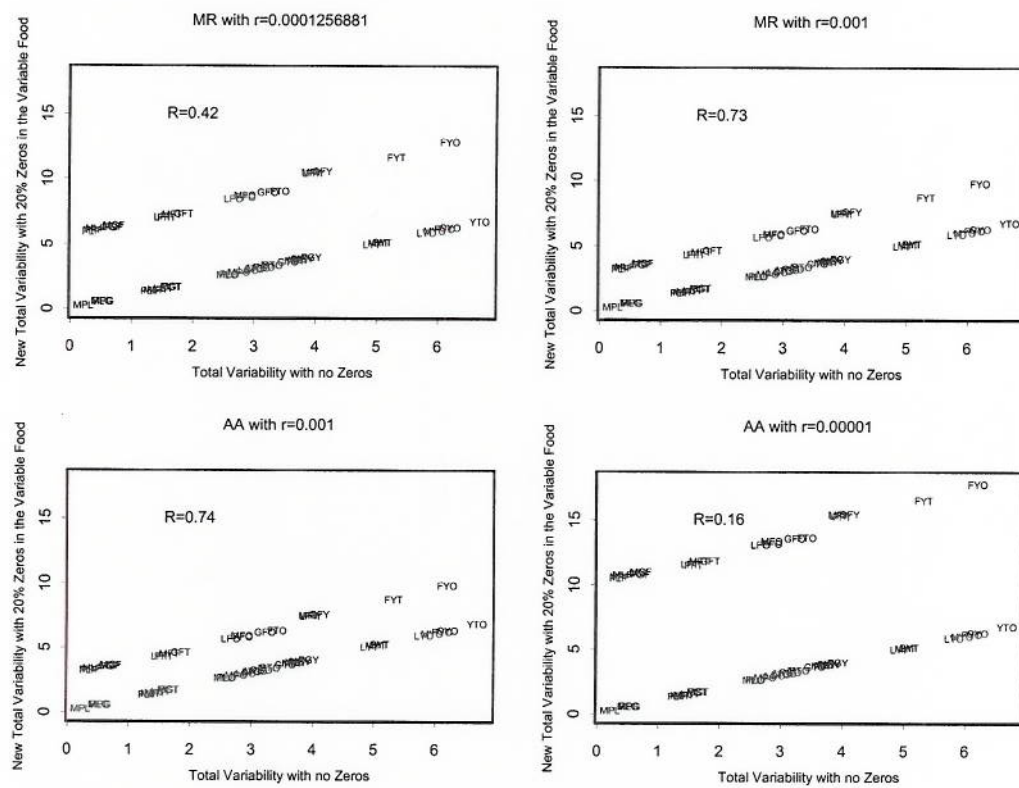


Figure 37. Changes in Total Variability after replacing 20% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

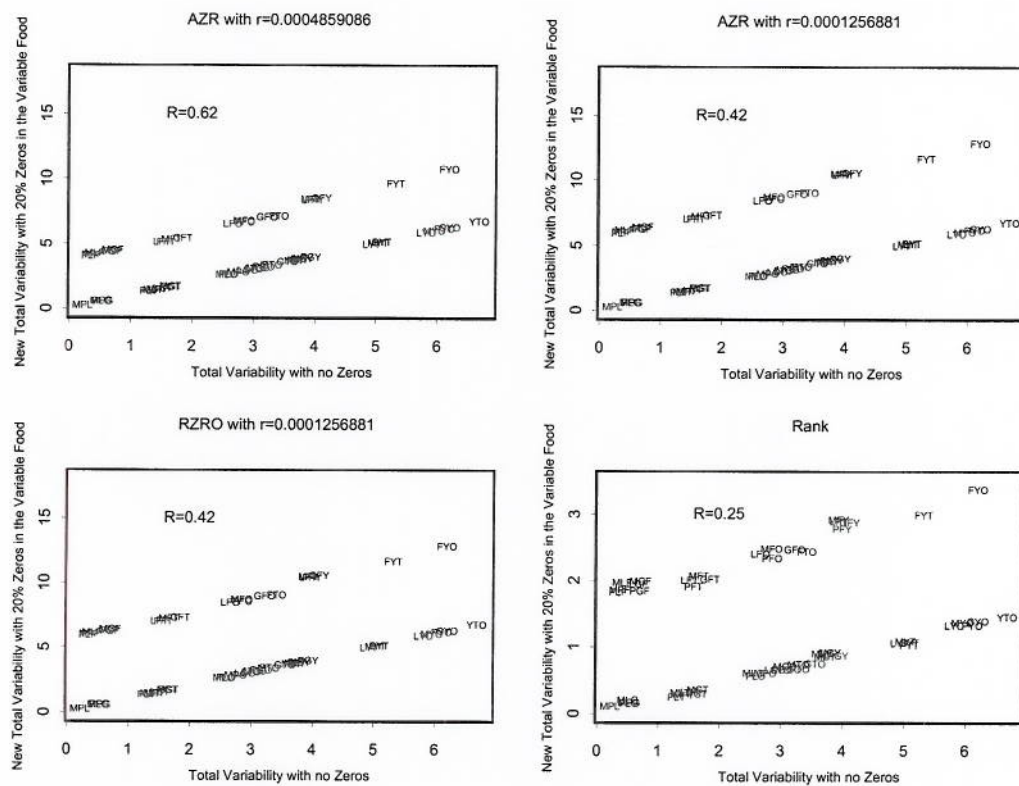


Figure 38. Changes in Total Variability after replacing 20% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

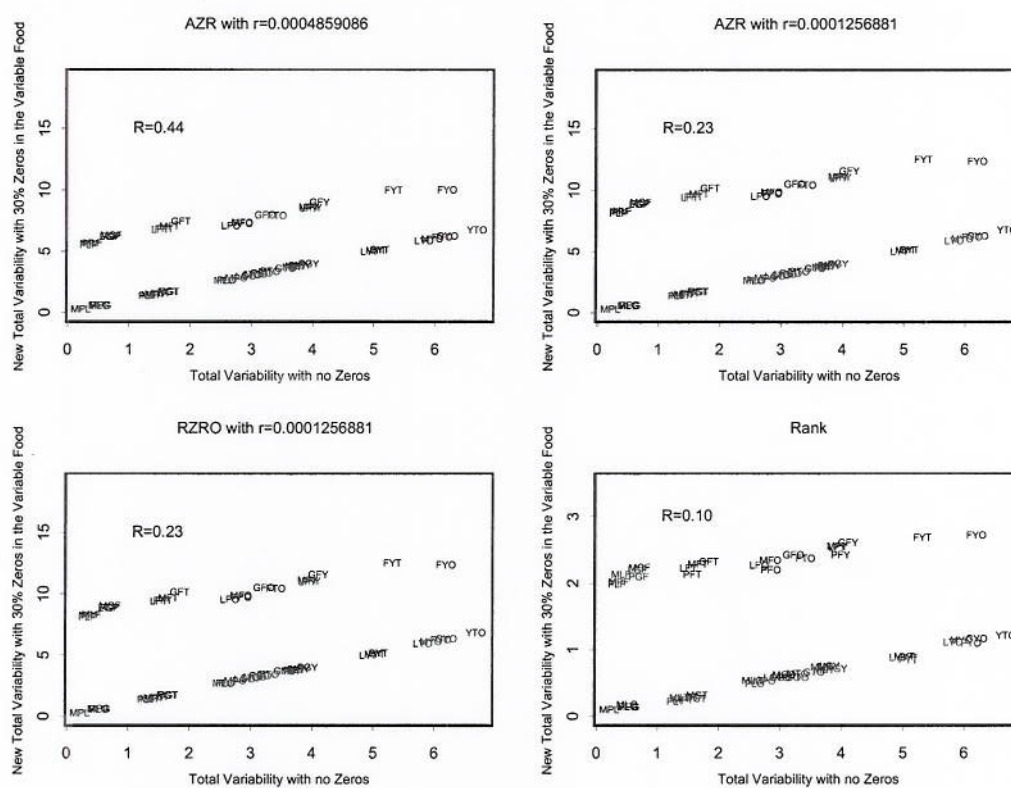


Figure 40. Changes in Total Variability after replacing 30% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

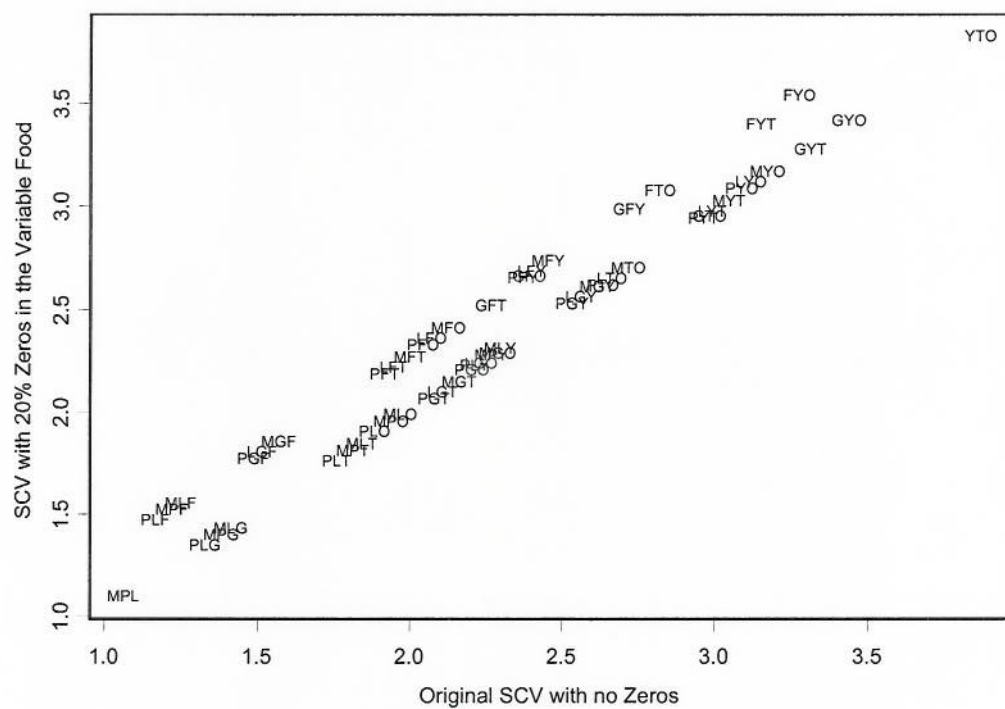


Figure 41. Changes in the Sum of Coefficients of Variation after replacing 20% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

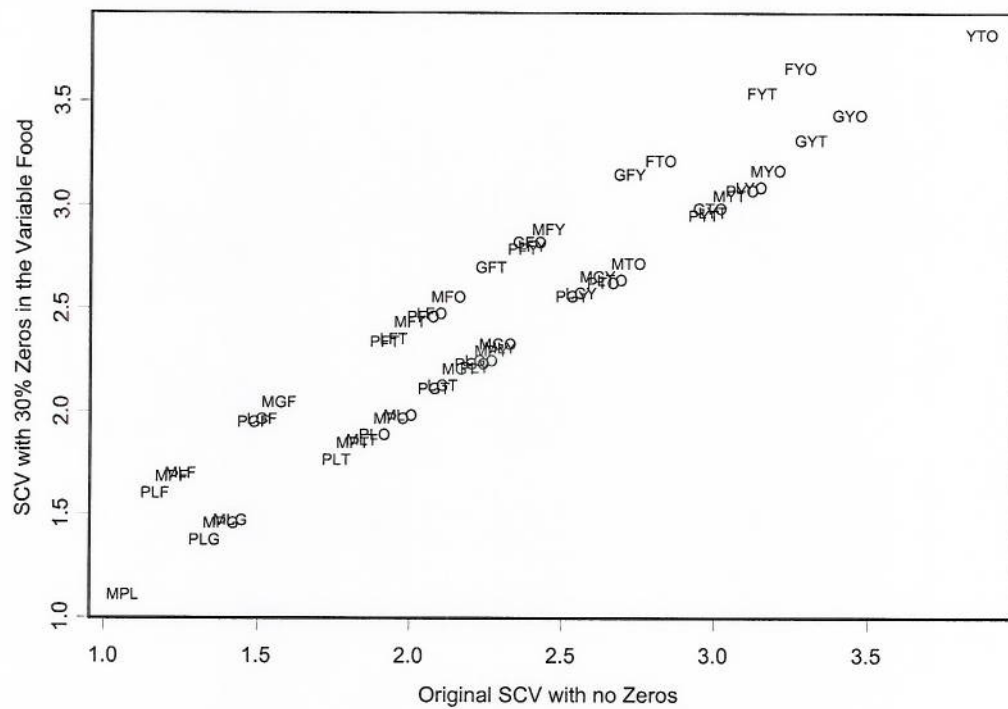


Figure 42. Changes in the Sum of Coefficients of Variation after replacing 30% of the observations in the variable Food with zeros. M: Metal, P: Paper, L: Plastic, G: Glass, F: Food, Y: Yard, T: Text, O: Other

**Replacing 20%, 30%, or 40% of the observations
in the variable Paper with zeros**

Table (23) presents number of 3-part subcompositions out of the top five with largest Sum of Coefficients of Variation and compositional Total Variability that include the variable Paper after replacing 10%, 20%, 30%, and 40% of the observations with zeros. The results are comparable to what we found using the variable Food. Compositional Total Variability is strongly affected by zeros while we observe a slow change in the Sum of Coefficients of Variation as the percentage of zeros increases in the data.

**Replacing 20%, 30%, or 40% of the observations
in the variable Yard with zeros**

Table (24) presents number of 3-part subcompositions out of the top five with largest Sum of Coefficients of Variation and compositional Total Variability that include the variable Yard after replacing 10%, 20%, 30%, and 40% of the observations in the variable with zeros. The results are comparable to what we found at 10% of zeros. Small changes in both Sum of Coefficients of Variations and the Total Variability occurred since all top five 3-part subcompositions in the original data with no zeros include the variable Yard.

% of Zeros	SCV	MR	MR	AA	AA	AZR	AZR	RZRO	Rank
10%	2	5	2	2	5	3	5	5	5
20%	2	5	5	5	5	5	5	5	5
30%	2	5	5	5	5	5	5	5	5
40%	3	5	5	5	5	5	5	5	5

Table 22. Number of 3-part subcompositions of the top five with largest Sum of Coefficients of Variation and largest Total Variability that include the variable Food at different percentages of zeros in the variable Food.

% of Zeros	SCV	MR	MR	AA	AA	AZR	AZR	RZRO	Rank
10%	0	5	2	2	5	4	5	5	5
20%	1	5	5	5	5	5	5	5	5
30%	2	5	5	5	5	5	5	5	5
40%	3	5	5	5	5	5	5	5	5

Table 23. Number of 3-part subcompositions of the top five with largest Sum of Coefficients of Variation and largest Total Variability that include the variable Paper at different percentages of zeros in the variable Paper.

% of Zeros	SCV	MR	MR	AA	AA	AZR	AZR	RZRO	Rank
10%	5	5	5	5	5	5	5	5	5
20%	5	5	5	5	5	5	5	5	5
30%	5	5	5	5	5	5	5	5	5
40%	5	5	5	5	5	5	5	5	5

Table 24. Number of 3-part subcompositions of the top five with largest Sum of Coefficients of Variation and largest Total Variability that include the variable Yard at different percentages of zeros in the variable Yard.

CHAPTER 5

REAL COMPOSITIONAL DATA WITH ZERO OBSERVATIONS

In this chapter, we evaluate the performance of the new method based on the Sum of Coefficients of Variation and Total Variability (based on the logratio transformations) using two real compositional data sets with zero observations. Total Variability is obtained after applying different zero treatment techniques presented in chapter 2.

Glacial Data Set

Consider the Glacial data set included in Aitchison (1986) and discussed in Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2000). It has 92 samples of pebbles of glacial tills sorted into four categories: red sandstone, gray sandstone, crystalline, and miscellaneous. The components x_1, x_2, x_3 and x_4 represent the corresponding percentages by weight of these four categories. There are 6 zero observations in x_3 , 30 zero observations in x_4 and 6 zero observations in both x_3 and x_4 . The Sum of Coefficients of Variation was computed for all 2-part and 3-part subcompositions. The Total Variability was computed after employing three treatments (1) Aitchison Additive approach with two different δ values $\delta_1 = 0.001$ and $\delta_2 = 0.0005$, (2) Multiplicative Replacement approach with $r_1 = 0.001$ and $r_2 = 0.0005$ and (3) Rank across variables and cases. Table (25) presents summary statistics of the components and Tables (26) and (27) present 2-part and 3-part subcompositions formed

Components	Mean	Standard Deviation	Coefficient of Variation
x_1	0.585	0.314	0.536
x_2	0.378	0.311	0.822
x_3	0.016	0.020	1.267
x_4	0.021	0.040	1.904

Table 25. Summary Statistics of the Glacial Compositional Data

from this data set and the corresponding Aitchison Total Variability. The results in these Tables show that the amount and the order of the Total Variability change when different strategies are employed as well as when the same strategy is employed with different replacements. For example, Total Variability for the 2-part subcomposition (x_1, x_2) is computed after applying Aitchison Additive replacement strategy with $\delta_1 = 0.001$ is 1.920, putting this subcomposition in the third place. With $\delta_2 = 0.0005$, Total Variability is 1.896 which is now in the second place. For the same subcomposition, Total Variability after applying Multiplicative Replacement with $r_1 = 0.001$ is 1.878 in the fifth place and with $r_2 = 0.0005$ is 1.878 but in the third place. Figures (43) and (44) are scatter plots of Total Variability of all 2-part and 3-part subcompositions obtained after using Aitchison Additive replacement strategy with two different δ s, $\delta_1 = 0.001$ and $\delta_2 = 0.0005$. Clearly the amount and the order of the Total Variability change with different replacements and that Aitchison Additive Strategy is sensitive to the changes in δ . Similarly, Figures (45) and (46) show the changes in the order and the amount of the Total Variability of the 2-part subcompositions and in the amount of the 3-part subcompositions obtained after applying Multiplicative replacement strategy with two different replacements $r_1 = 0.001$ and $r_2 = 0.0005$.

In contrast, there is a small change in the amount and no change in the order of the Sum of Coefficients of Variation for all 2-part and 3-part subcompositions when it was computed using the original data and the replaced data sets using different

2-part Subcomposition	MR $r = 0.001$	MR $r = 0.0005$	AA $\delta = 0.001$	AA $\delta = 0.0005$	Rank
x_1, x_2	1.878	1.878	1.920	1.896	0.450
x_1, x_3	1.023	1.206	1.317	1.566	0.753
x_1, x_4	1.980	2.512	2.787	3.469	1.660
x_2, x_3	1.291	1.508	1.614	1.908	0.858
x_2, x_4	1.652	2.160	2.404	3.076	1.486
x_3, x_4	1.619	2.292	2.604	3.504	1.837

Table 26. Glacial Compositional Data: 2-part subcompositions and the corresponding Total Variability

3-part Subcomposition	MR $r = 0.001$	MR $r = 0.0005$	AA $\delta = 0.001$	AA $\delta = 0.0005$	Rank
x_1, x_2, x_3	2.795	3.061	3.234	3.580	2.061
x_1, x_2, x_4	3.673	4.366	4.740	5.628	3.596
x_1, x_3, x_4	3.082	4.007	4.472	5.692	4.251
x_2, x_3, x_4	3.042	3.973	4.415	5.659	4.182

Table 27. Glacial Compositional Data: 3-part subcompositions and the corresponding Total Variability

zero replacement strategies and with different replaced values. Tables (28) and (29) present Sum of Coefficients of Variation obtained using the original data with zeros and the the replaced data sets using Aitchison Additive and Multiplicative strategies. Figures (47 - 49) show the same findings graphically.

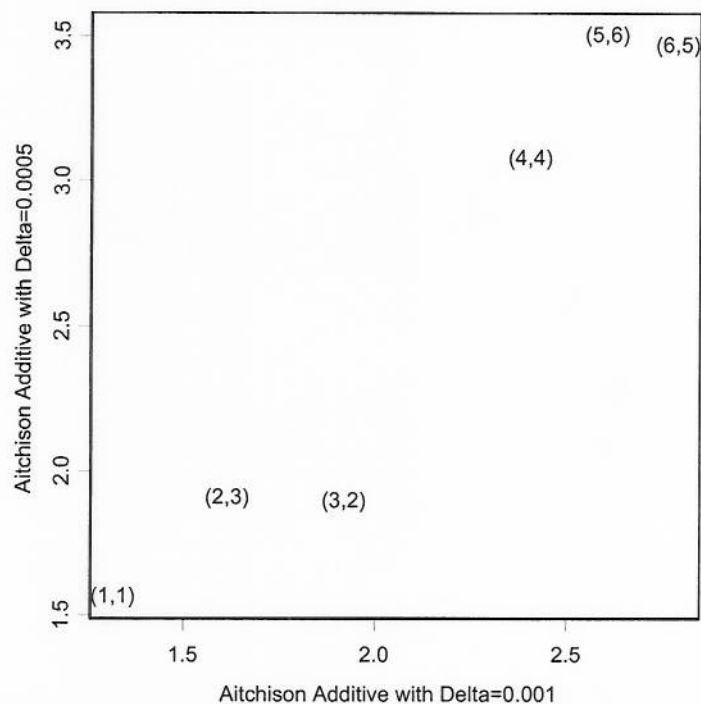


Figure 43. Plot of Total Variability for all 2-part subcompositions obtained after employing Aitchison Additive Replacement Strategy with $\delta_1 = 0.001$ and $\delta_2 = 0.0005$

2-part Subcomposition	SCV Original Data	SCV MR $r = 0.001$	SCV MR $r = 0.0005$	SCV AA $\delta = 0.001$	SCV AA $\delta = 0.0005$
x_1, x_2	1.358	1.358	1.358	1.358	1.358
x_1, x_3	1.803	1.786	1.795	1.800	1.802
x_1, x_4	2.440	2.396	2.418	2.424	2.432
x_2, x_3	2.089	2.072	2.081	2.086	2.088
x_2, x_4	2.726	2.682	2.704	2.710	2.718
x_3, x_4	3.171	3.110	3.141	3.152	3.162

Table 28. Glacial Compositional Data: 2-part subcompositions and the corresponding Sum of Coefficients of Variation obtained using the original and the replaced data sets

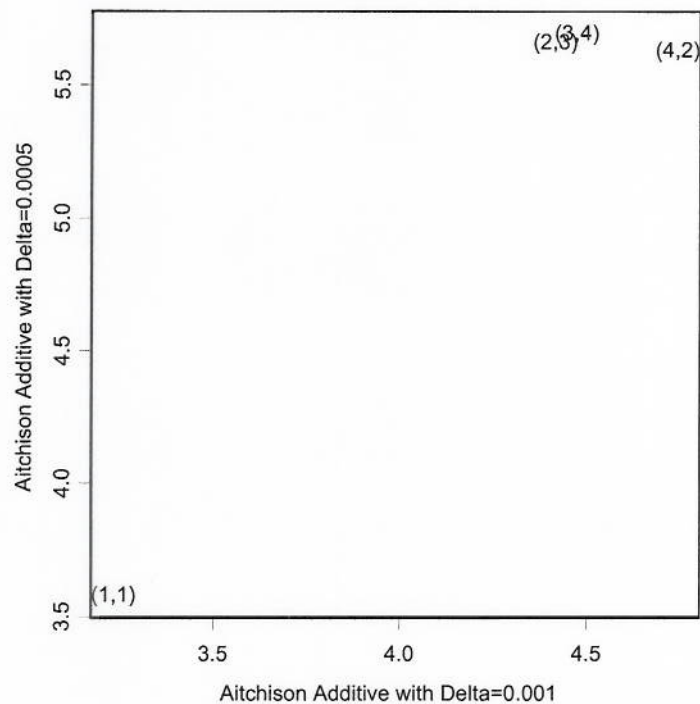


Figure 44. Plot of Total Variability for all 3-part subcompositions obtained after employing Aitchison Additive Replacement Strategy with $\delta_1 = 0.001$ and $\delta_2 = 0.0005$

3-part Subcomposition	SCV Original Data	SCV MR $r = 0.001$	SCV MR $r = 0.0005$	SCV AA $\delta = 0.001$	SCV AA $\delta = 0.0005$
x_1, x_2, x_3	2.625	2.608	2.617	2.622	2.624
x_1, x_2, x_4	3.262	3.218	3.240	3.246	3.254
x_1, x_3, x_3	3.707	3.646	3.677	3.688	3.698
x_2, x_3, x_4	3.993	3.932	3.963	3.975	3.984

Table 29. Glacial Compositional Data: 3-part subcompositions and the corresponding Sum of Coefficients of Variation obtained using the original and the replaced data sets

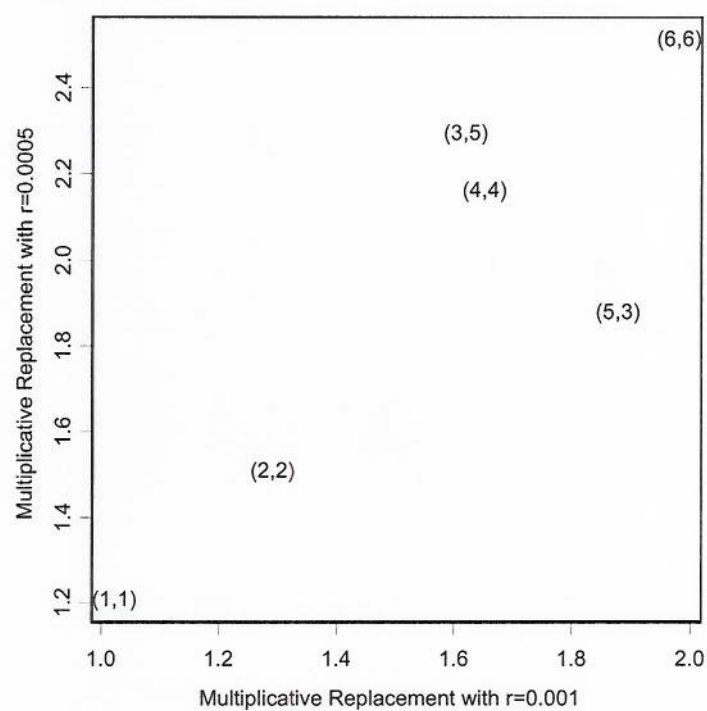


Figure 45. Plot of Total Variability for all 2-part subcompositions obtained after employing Multiplicative Replacement Strategy with $r_1 = 0.001$ and $r_2 = 0.0005$

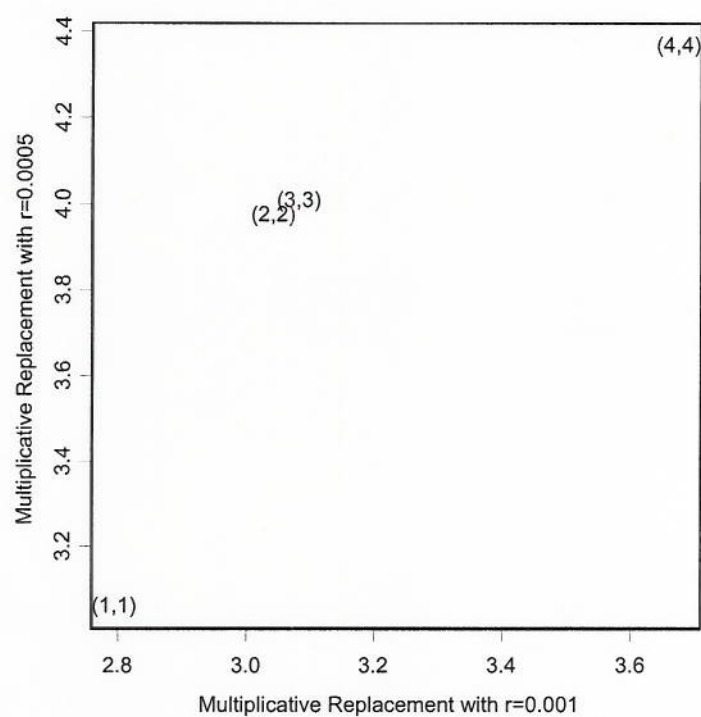


Figure 46. Plot of Total Variability for all 3-part subcompositions obtained after employing Multiplicative Replacement Strategy with $r_1 = 0.001$ and $r_2 = 0.0005$

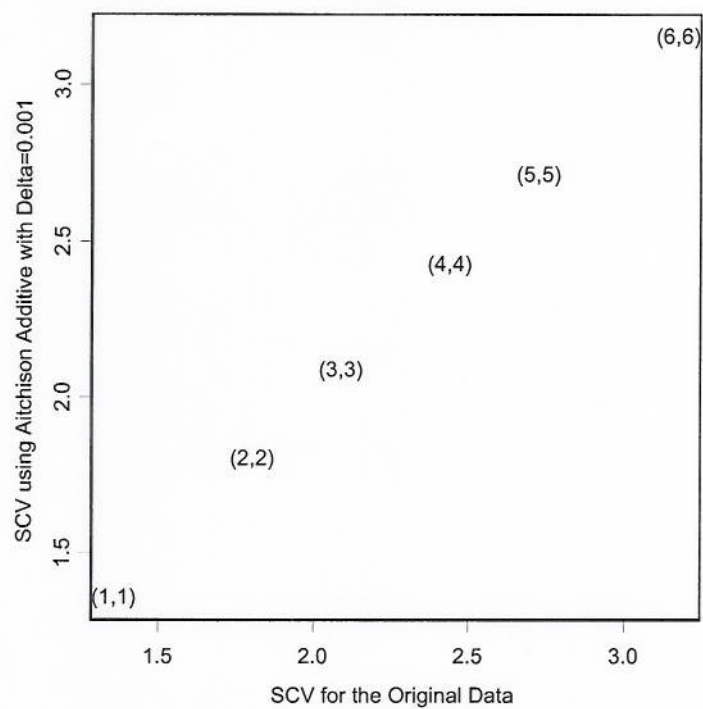


Figure 47. Plot of Sum of Coefficients of Variation for all 2-part subcompositions obtained using the original and the replaced data sets using Aitchison Additive Replacement Strategy with $\delta = 0.001$

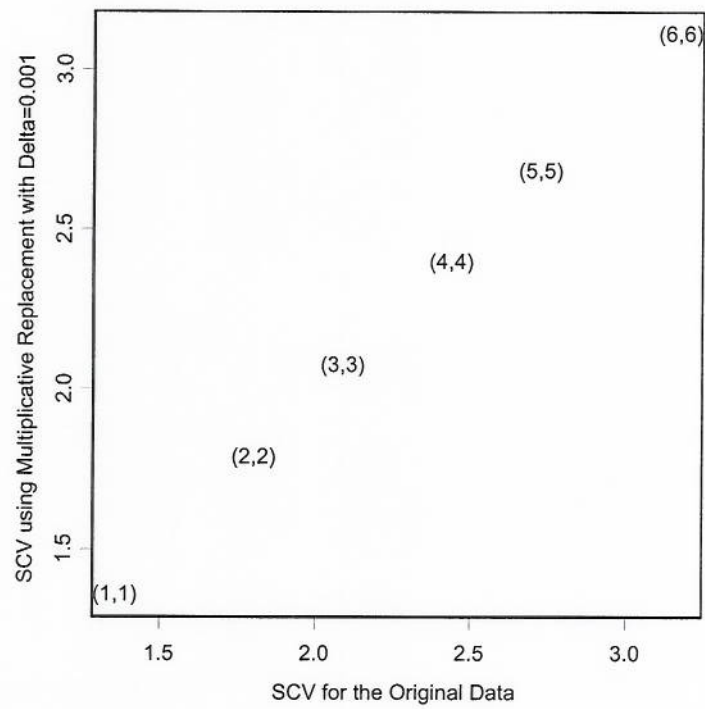


Figure 48. Plot of Sum of Coefficients of Variation for all 2-part subcompositions obtained using the original data and the replaced data sets using Multiplicative Replacement Strategy with $r = 0.001$

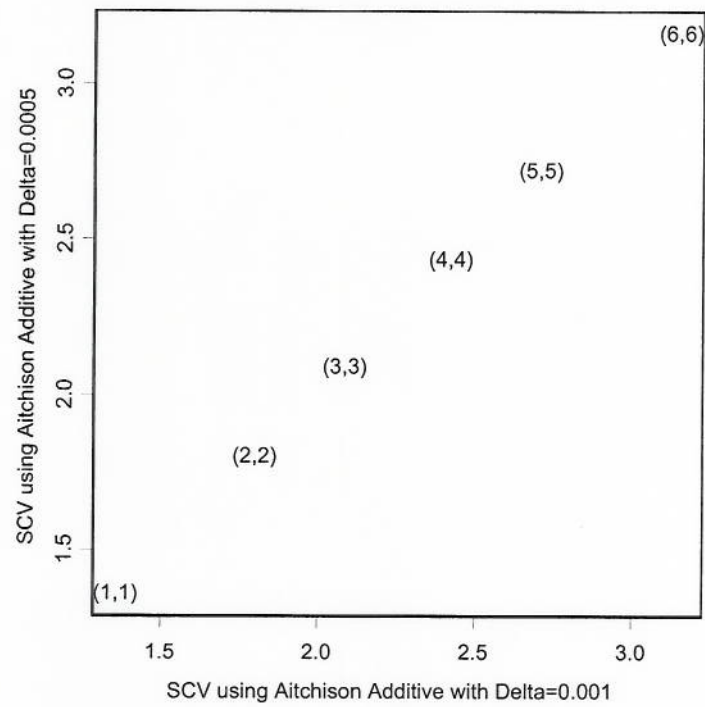


Figure 49. Plot of the Sum of Coefficients of Variation for all 2-part subcompositions obtained using the replaced data sets using Aitchison Additive Replacement Strategy with $\delta_1 = 0.001$ and $\delta_2 = 0.0005$

Archaeological Glass

The second compositional data we investigate is a set of archaeological glass compositions for a particular colorless Romano-British glass vessel type, facet-cut beakers. The data consist of 12 major and minor oxides for 63 samples : Al_2O_3 , Fe_2O_3 , MgO , CaO , Na_2O , K_2O , TiO_2 , P_2O_5 , MnO , Sb_2O_5 , PbO , and Other. These data are given in Baxter, Cool and Jackson (2005), where the archaeological background is discussed. The research was based on large samples from the four typologically distinct groups of vessels, Type 1 is the cast colorless bowl, Type 2 is the externally ground facet-cut beaker, Type 3 is the wheel-cut beaker, and Type 4 is the cylindrical cup.

Table (30) presents summary statistics of the 12 oxides for the facet-cut type. The 11th oxide , PbO , contains 14 zero observations. The Sum of Coefficients of Variation and Total Variability were computed for all 4-part subcompositions. Total Variability computed after employing (1) Aitchison Additive approach with two different values $r_1 = 0.0000076$ and $r_2 = 0.0001$ and (2) Multiplicative Replacement approach with $r_1 = 0.000055$ and $r_2 = 0.0001$.

Figure (50) is a scatter plot of the top 20 4-part subcompositions with largest Total Variability obtained after using Aitchison Additive replacement strategy with $r_1 = 0.0000076$ and Total Variability for the same 20 subcompositions after employing Aitchison Additive replacement strategy with $r_2 = 0.0001$. Clearly the amount and the order of the Total Variability of the top 20 4-part subcompositions changed after using a different replacement value for the zero observations. Table (31) presents the top 20 4-part subcomposition with largest Total Variability after employing Aitchison's Additive replacement Strategy with $r_1 = 0.0000076$ and the top 20 4-part subcomposition with largest Total Variability after employing Aitchison's Additive replacement Strategy with $r_1 = 0.0001$. It is clear that the two zero

Components	Mean	Standard Deviation	Coefficient of Variation
Al_2O_3	0.0184	0.0031	0.1664
Fe_2O_3	0.0037	0.0008	0.2247
MgO	0.0038	0.0007	0.1773
CaO	0.0537	0.0098	0.1818
Na_2O	0.1794	0.0084	0.0468
K_2O	0.0054	0.0010	0.1882
TiO_2	0.0006	0.0002	0.2659
P_2O_5	0.0004	0.0001	0.2549
MnO	0.0002	0.0002	0.8877
Sb_2O_5	0.0145	0.0053	0.3682
PbO	0.0017	0.0019	1.0928
<i>Other</i>	0.7181	0.0156	0.0217

Table 30. Summary Statistics of the Archaeological Glass Compositional Data

replacement scenarios produce different groups of subcompositions. Similar results in Figure(51) after employing Multiplicative replacement strategy with $r_1 = 0.000055$ and $r_2 = 0.0001$.

Consistent with what we found in the Glacial data, there was no change in the order of the top 20 4-par subcompositions with targets Sum of Coefficients of Variation computed for the original data and for the replaced data using any of the zero replacement strategies and with any replaced value as it appears in Figure (52).

	Top 20 4-part Subcomposition AA with $r=0.0000076$	Top 20 4-part Subcomposition AA with $r=0.0001$
1	(AL,Ca,Mn,Pb) Totvar= 4.620	(AL,Ca,Mn,Pb) Totvar= 1.909
2	(AL,Ca,Ti,Pb) Totvar= 4.568	(AL,Mn,Pb, Other) Totvar= 1.852
3	(AL,Ca,P,Pb) Totvar= 4.566	(AL,K,Mn,Pb) Totvar= 1.845
4	(AL,Mg,Ca,Pb) Totvar= 4.564	(AL,Na,Mn,Pb) Totvar= 1.844
5	(AL,Ca,K,Pb) Totvar= 4.561	(AL,Ca,P,Pb) Totvar= 1.839
6	(AL,Ca,Pb,Other) Totvar= 4.588	(AL,Ca,Ti,Pb) Totvar= 1.836
7	(AL,Fe,Ca,Pb) Totvar= 4.557	(AL,Ca,K,Pb) Totvar= 1.835
8	(AL,Ca,Na,Pb) Totvar= 4.552	(Ca,Mn,Pb, Other) Totvar= 1.834
9	(AL,Mn,Pb,Other) Totvar= 4.546	(AL,Ca,Sb,Pb) Totvar= 1.832
10	(AL,Na,Mn,Pb) Totvar= 4.539	(AL,Fe,Ca,Pb) Totvar= 1.832
11	(AL,K,Mn,Pb) Totvar= 4.535	(AL,Mg,Ca,Pb) Totvar= 1.827
12	(Ca,Mn,Pb,Other) Totvar= 4.531	(AL,Ca,Pb,Other) Totvar= 1.827
13	(AL,Mg,Mn,Pb) Totvar= 4.525	(Ca,K,Mn,Pb) Totvar= 1.826
14	(Ca,Na,Mn,Pb) Totvar= 4.522	(Ca,Na,Mn,Pb) Totvar= 1.825
15	(AL,Ca,Sb,Pb) Totvar= 4.520	(AL,Mg,Mn,Pb) Totvar= 1.824
16	(Ca,K,Mn,Pb) Totvar= 4.518	(AL,Ca,Na,Pb) Totvar= 1.821
17	(AL,Ti,Mn,Pb) Totvar= 4.512	(AL,Ti,Mn,Pb) Totvar= 1.816
18	(AL,P,Mn,Pb) Totvar= 4.508	(AL,P,Mn,Pb) Totvar= 1.815
19	(AL,Mg,K,Pb) Totvar= 4.504	(AL,Fe,Mn,Pb) Totvar= 1.810
20	(AL,Mg,Pb,Other) Totvar= 4.504	(AL,Mn,Sb,Pb) Totvar= 1.802

Table 31. Top 20 4-part subcompositions with largest Total Variability computed after employing Aitchison Additive zero replacement strategy (AA) with two different values $r_1 = 0.0000076$ and $r_2 = 0.0001$

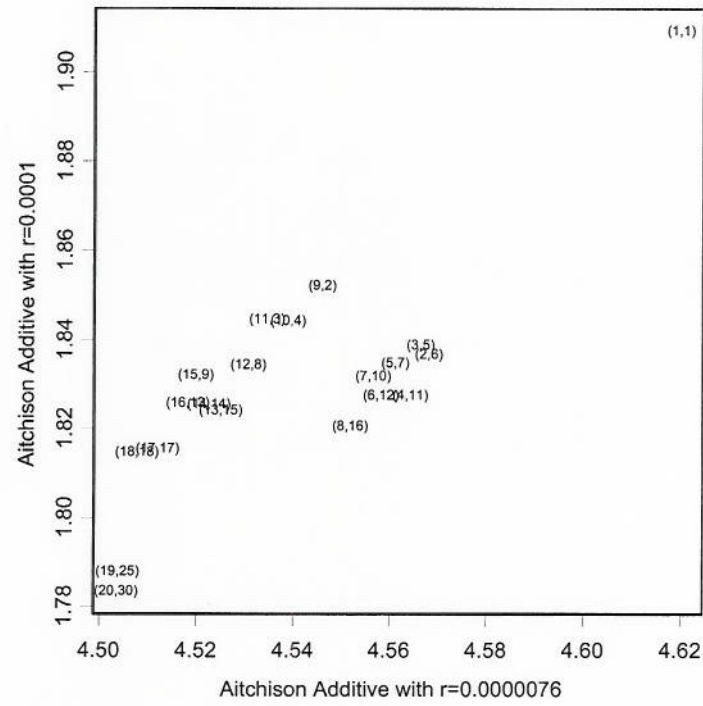


Figure 50. Plot of the top 20 4-part subcompositions with largest Total Variability obtained after employing Aitchison Additive Replacement Strategy with $r_1 = 0.0000076$ and the Total Variability for the same subcompositions after using $r_2 = 0.0001$

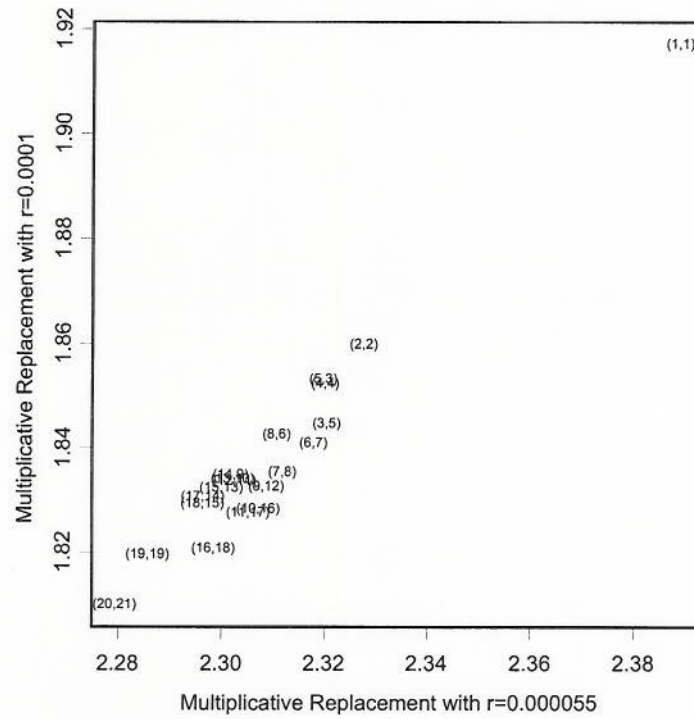


Figure 51. Plot of the top 20 4-part subcompositions with largest Total Variability obtained after employing Multiplicative Replacement Strategy with $r_1 = 0.000055$ and Total Variability for the same subcompositions after using $r_2 = 0.0001$

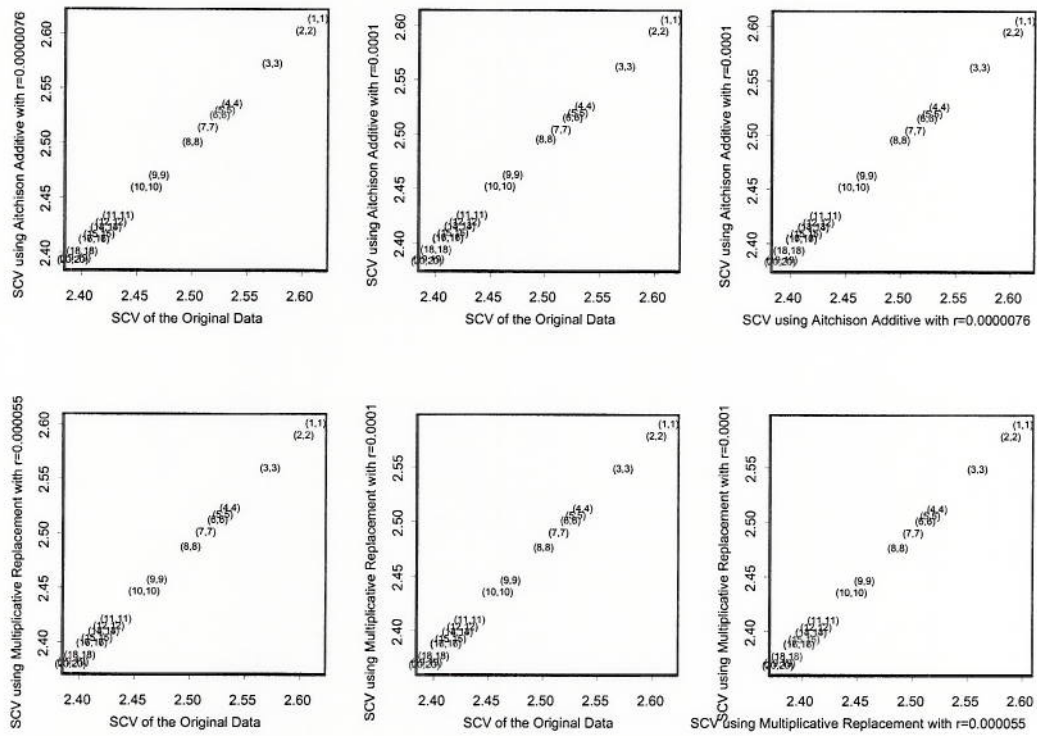


Figure 52. Plot of the top 20 4-part subcompositions with largest Sum of Coefficients of Variation obtained using the original data and Sum of Coefficients of Variation for the same subcompositions in the replaced data sets using Aitchison Additive Replacement Strategy with $r_1 = 0.0000076$ and $r_2 = 0.0001$ and Multiplicative Replacement Strategy with $r_1 = 0.000055$ and $r_2 = 0.0001$

CHAPTER 6

CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

Conclusion

We have introduced a new and simple measure of compositional data variability based on the Sum of Coefficients of Variation of the subcompositions. This is offered as an alternative to Aitchison's Total Variability measure based on the logratio models. For Dirichlet distribution generated from similar independent Gamma random variables, we demonstrated theoretically that the two measures are functionally related.

In a wide range of both numerical simulations and real compositional datasets we illustrate a strong correlation between the Sum of Coefficients of Variation and compositional Total Variability computed using Aitchison's logratio transformations. However, the two approaches perform differently when applied to real data sets with zeros. Total Variability varied in magnitude and subcompositional ordering when different zero treatment techniques were employed. Accounting for zeros is not necessary when using the new technique based on the Sum of Coefficients of Variation. The new approach when applied to data where zeros were replaced using different strategies, yielded nearly identical results to those without any special zero treatment whatsoever. The new measure of compositional data variability avoids the complicated construction of the logratio models and resolves many of the challenges

in measuring variability in the subcompositions.

Future Research

Missing values and outliers in compositional data are active fields of research. Hron et al. (2008) introduced two imputation methods for estimating missing values in compositional data. The first method was the k-nearest neighbors procedure based on Aitchison Distance. They indicated that such a method is not robust against outliers. The second method is based on an iterative regression, accounting for the whole multivariate data information. Martin-Fernandez et al. (2003) considered a generalization of the multiplicative replacement strategy as a substitution method for missing data sets. Filzmoser and Hron (2007) developed an outlier detection tool using Mahalanobis distance of logratios of the compositional data. Applying the Sum of Coefficients of Variation technique to compositional datasets with outliers or missing values present is likely to find new results in many applications. We would like to conduct a comparison between the performance of Sum of Coefficients of Variation and compositional Total Variability in such data sets.

Hijazi and Jernigan (2009) introduced a Dirichlet regression technique to model compositional data in the presence of a covariate. They demonstrated that the Dirichlet regression is an informative alternative to logratio covariate models. We would like to extend the application of Sum of Coefficients of Variation technique to model compositional data in the presence of an observed covariate. For example, a weighted Sum of Coefficients of Variation could be implemented to adjust for the household size in modeling the Garbage compositional data.

There are some applications where components with low absolute percentages can have a great importance. Small changes in the proportion of even low absolute percentage components can lead to significant changes in the structure of a composition. Consider the example of salt as a compositional component in a bowl of soup.

If salt is absent, the soup may be bland and unsatisfying. A small proportion of salt can greatly enhance the taste of the soup. But an overuse of the salt can result in soup that is unpalatable or even inedible. The Total Variability or the Sum of Coefficients of Variation measures would be preferred in this setting.

But it is essential to understand the structure and setting of the data for proper measurement of compositional variability. Aitchison's Total Variability or the Sum of Coefficients of Variation approach put greater emphasis on components with high relative variation. These can be components with simultaneous low absolute percentages that don't contribute to a possible causal understanding of the data and can therefore produce unsatisfactory results. This was illustrated by Baxter, Beardah, Cool, and Jackson (2005) and Baxter, Cool, and Jackson (2005) with the glass compositional dataset. Baxter et al. (2005) and Baxter and Freestone (2006) illustrated that even bivariate analysis and crude principal component analysis can produce more interpretable results than logratio analysis. Beardah et al. (2003) and Greenacre (2002) suggested some form of weighted logratio analysis could down-weight the influence of those components with low absolute percentages. For such settings and data, measuring compositional variability by Total Variability or even with the closely correlated Sum of Coefficients of Variation approach may not be ideal. Other approaches need to be investigated.

APPENDIX A
GARBAGE COMPOSITIONAL DATA

	Garbage Compositional Data							
	Metal	Paper	Plastic	Glass	Food	Yard	Text	Other
1	0.1013	0.2240	0.0251	0.0799	0.0967	0.0353	0.0046	0.4331
2	0.0521	0.3791	0.0706	0.1733	0.1843	0.0005	0.0230	0.1172
3	0.0931	0.3460	0.0793	0.1638	0.1605	0.0087	0.0181	0.1304
4	0.0792	0.2314	0.0743	0.1291	0.0782	0.0165	0.0593	0.3319
5	0.0538	0.3125	0.0785	0.2262	0.2258	0.0054	0.0197	0.0781
6	0.0959	0.3178	0.0826	0.1137	0.0667	0.2091	0.0164	0.0977
7	0.0884	0.3129	0.0389	0.0234	0.4040	0.0032	0.0275	0.1017
8	0.0725	0.2318	0.0619	0.1179	0.1953	0.0966	0.0043	0.2198
9	0.0697	0.4833	0.1028	0.0589	0.1326	0.0039	0.0243	0.1244
10	0.0532	0.1795	0.0591	0.4972	0.0768	0.1086	0.0186	0.0070
11	0.0734	0.2937	0.0659	0.0722	0.2095	0.0158	0.0083	0.2613
12	0.0883	0.2515	0.0540	0.1094	0.0795	0.2978	0.0941	0.0255
13	0.0617	0.4563	0.0656	0.0937	0.1621	0.0224	0.0127	0.1255
14	0.0957	0.2705	0.1362	0.1343	0.1420	0.0792	0.0425	0.0995
15	0.0666	0.1449	0.0450	0.1172	0.1589	0.1382	0.0045	0.3247
16	0.0622	0.2429	0.0385	0.0841	0.1044	0.0568	0.0112	0.3999
17	0.0646	0.2947	0.0790	0.0463	0.0465	0.3910	0.0104	0.0676
18	0.0762	0.3190	0.0524	0.1024	0.1448	0.0056	0.0634	0.2361
19	0.0826	0.3519	0.0378	0.3642	0.1123	0.0006	0.0500	0.0006
20	0.1139	0.4306	0.0674	0.1114	0.2044	0.0003	0.0304	0.0418
21	0.1939	0.3020	0.0273	0.0870	0.2908	0.0865	0.0048	0.0078
22	0.0493	0.5200	0.0555	0.0802	0.2090	0.0210	0.0062	0.0588
23	0.0682	0.4299	0.1515	0.0081	0.3140	0.0006	0.0149	0.0129
24	0.0998	0.7956	0.0365	0.0219	0.0389	0.0024	0.0024	0.0024
25	0.0870	0.2871	0.0624	0.2949	0.2376	0.0004	0.0293	0.0013
26	0.0761	0.4106	0.0623	0.0542	0.2773	0.0193	0.0065	0.0937

	Metal	Paper	Plastic	Glass	Food	Yard	Text	Other
27	0.0954	0.2452	0.0680	0.1050	0.0902	0.0002	0.0129	0.3832
28	0.0495	0.4866	0.0738	0.1010	0.2356	0.0208	0.0267	0.0059
29	0.0421	0.4454	0.0628	0.1353	0.2133	0.0554	0.0054	0.0402
30	0.0741	0.3328	0.0484	0.1861	0.1525	0.2024	0.0016	0.0021
31	0.0680	0.5956	0.0577	0.0687	0.1860	0.0214	0.0006	0.0019
32	0.0633	0.4544	0.0386	0.1304	0.2458	0.0309	0.0024	0.0343
33	0.1469	0.5552	0.0423	0.0670	0.1863	0.0006	0.0012	0.0006
34	0.0518	0.2717	0.0587	0.2699	0.0529	0.1735	0.1043	0.0172
35	0.0747	0.4075	0.0885	0.1148	0.2673	0.0002	0.0233	0.0237
36	0.1998	0.4272	0.0509	0.1345	0.1146	0.0376	0.0349	0.0006
37	0.0855	0.3536	0.0804	0.2308	0.2296	0.0172	0.0023	0.0006
38	0.0575	0.3371	0.0613	0.2725	0.1804	0.0008	0.0051	0.0853
39	0.0904	0.5039	0.0878	0.0988	0.0978	0.0005	0.0904	0.0303
40	0.0448	0.3849	0.0729	0.0848	0.3873	0.0172	0.0057	0.0024
41	0.0628	0.5186	0.068	0.1383	0.1656	0.0005	0.0184	0.0278
42	0.0902	0.4511	0.0496	0.0626	0.1449	0.1628	0.0220	0.0169
43	0.0836	0.2179	0.0909	0.1932	0.2233	0.1798	0.0007	0.0107
44	0.0586	0.3716	0.0856	0.2230	0.1734	0.0766	0.0090	0.0023
45	0.1777	0.4029	0.0701	0.0778	0.0459	0.0371	0.0032	0.1853
46	0.1547	0.4306	0.1129	0.1735	0.0697	0.0005	0.0043	0.0538
47	0.0721	0.3667	0.0891	0.0981	0.2535	0.0004	0.0252	0.0950
48	0.1203	0.4019	0.0779	0.2570	0.0629	0.0766	0.0007	0.0027
49	0.0996	0.2907	0.1014	0.0465	0.1809	0.1971	0.0123	0.0715
50	0.1008	0.4582	0.0784	0.0972	0.1689	0.0394	0.0151	0.0420
51	0.0850	0.1942	0.0969	0.2264	0.0540	0.0022	0.0540	0.2873
52	0.1295	0.4471	0.1127	0.1363	0.1081	0.0008	0.0206	0.0449
53	0.0719	0.1918	0.0734	0.0872	0.2281	0.0197	0.0374	0.2906
54	0.0351	0.3170	0.0813	0.1276	0.0862	0.1236	0.0561	0.1730
55	0.0859	0.3298	0.0770	0.2483	0.1586	0.0354	0.0502	0.0148
56	0.0599	0.3990	0.0574	0.0779	0.3259	0.0091	0.0295	0.0413
57	0.1039	0.5060	0.0649	0.0556	0.2019	0.0004	0.0085	0.0588
58	0.0638	0.3789	0.0604	0.0607	0.3804	0.0050	0.0076	0.0432
59	0.0755	0.3064	0.1021	0.0755	0.3455	0.0009	0.0018	0.0924
60	0.0540	0.3228	0.0455	0.3150	0.1715	0.0004	0.0337	0.0572
61	0.1228	0.3306	0.0520	0.3262	0.1595	0.0072	0.0009	0.0009
62	0.1127	0.2229	0.0632	0.2588	0.2878	0.0145	0.0009	0.0393

APPENDIX B

S-PLUS PROGRAMS

This chapter consists of major S-Plus codes used through the study.

Main Functions

Compositional Total Variability

```
comp.var<-function(h)
{
  m <- dim(h)
  a <- matrix(0, m[2], m[2])
  for(i in 1:m[2]) {
    a[, i] <- apply(log(sweep(h, 1, h[, i], "/")), 2, var)
  }
  a
}
```

```
comp.r2<-function(x, sc) {
  n <- dim(x)
  tt <- comp.var(x)
  b <- sum(tt)
  a <- sum(comp.var(x)[sc, sc])
  r2 <- (a/length(sc))/(b/n[2])
  c(a, r2)
}
```

Closure Operation

```
closure<-function(x) {
  sweep(x, 1, apply(x, 1, sum), "/")
}
```

The following S-plus code used to calculate different combinations

```
combinations<-function(n, k, set = 1:n) {
  fun <- function(n, k, set)
  {
    if(k <= 0)
      vector(mode(set), 0)
    else if(k >= n)
      set
    else rbind(cbind(set[1], Recall(n - 1, k - 1, set[-1])), Recall(n - 1,
      k, set[-1]))
  }
  fun(n, k, set)
}
```

Triangle Plot

```
triangle_function(p, a = c(1, 2, 3), pch = 1, add = F, covar, r = 1,
cex = 1) {
  q <- as.data.frame(p)
  x <- sweep(p[, a], 1, apply(p[, a], 1, sum), "/")
  at <- c(0, 2/sqrt(3), 1/sqrt(3), 0)
  bt <- c(0, 0, 1, 0)
  par(pty = "s")
  if(!add) {
    plot(at, bt, type = "l", axes = F, xlab = "", ylab = "", xlim
      = c(0, 1.2), ylim = c(0, 1.2))
    text(-0.02, 0, names(q)[a[2]])
    text(2/sqrt(3) + 0.05, 0, names(q)[a[3]])
    text(1/sqrt(3), 1.05, names(q)[a[1]])
  }
  if(missing(covar)) {
    points((x[, 1] + 2 * x[, 3])/sqrt(3), x[, 1], pch = pch)
  }
  else {
    if(is.numeric(covar)) {
      cv <- round(covar, r)
    }
    else {
      cv <- covar
    }
    text((x[, 1] + 2 * x[, 3])/sqrt(3), x[, 1], cv, cex = cex)
  }
}
```

This function used to generate random samples from Additive Logistic Normal Distribution

```
radlognorm_function(m, d, mean = rep(0, d), cov = diag(d))
{
  y <- rmvnorm(m, mean = mean, cov = cov)
  x <- exp(y)/(apply(exp(y), 1, sum) + 1)
  x <- cbind(x, 1 - apply(x, 1, sum))
  x
}
```

Additive logratio transformation

```
alr_function(x) {
  n <- dim(x)
  y <- log(sweep(x, 1, x[, n[2]], "/"))
  y[, 1:(n[2] - 1)]
}
```

Estimates of Sum of Coefficients of Variation and Total Variability

Computing CV using equation (3.4) for D=3 and different alphas ($\alpha = 10, \dots, 40$)

```
smry.alpha<-function(N){
  out.smry<-matrix(0,ncol=1,nrow=N,byrow=F)
  for(i in 1:N)
  { alph<-i out.smry[i]<-sqrt(2/(3*i+1))
  }
  return(out.smry)
}
smry.cv.out<-smry.alpha(40)[10:40]
```

Computing Compositional Total Variability using equation (3.14) for D=3 and different alphas ($\alpha = 10, \dots, 40$)

```
Alpha<-(10:40)
2*trigamma(alpha)
var.tiagamma<-c(0.21, 0.19, 0.174, 0.16, 0.148, 0.138, 0.129, 0.121,
```



```
0.114, 0.108, 0.103, 0.098, 0.093, 0.089, 0.085, 0.082, 0.078,
0.075, 0.073, 0.07, 0.068, 0.066, 0.063, 0.062, 0.06, 0.058, 0.056,
0.055, 0.053, 0.052, 0.051)
```

```
# Computing Sum of Coefficients of Variation and Compositional Total Vari-
ability using Simulated Data from random Gamma with D=3, n=100 and different
alphas ( $\alpha = 10, \dots, 40$ )
```

```
find.dist1<-function(N,m,col.cnt){
find.out.mean<-matrix(0,ncol=col.cnt,nrow=m,byrow=F)
for(j in 10:m){
cat("S:", j, "out of", m, fill = T)
matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F) for(i in 1:N){
randomgam1_rgamma(100,j)
randomgam2_rgamma(100,j)
randomgam3_rgamma(100,j)
compgam123_cbind(randomgam1,randomgam2,randomgam3)
compgam2c_closure(compgam123)
dircv_apply(compgam2c,2,stdev)/apply(compgam2c,2,mean)
dirscv_sum(apply(compgam2c,2,stdev)/apply(compgam2c,2,mean))
dirmcv_mean(apply(compgam2c,2,stdev)/apply(compgam2c,2,mean))
var.tot_comp.var(compgam2c)
atotvar_(sum(comp.var(compgam2c)))/6
smry.all<-cbind(dirscv,dirmcv, atotvar)
matr.smry[i,]<-smry.all
}
find.out.mean[j,]=apply(matr.smry,2,mean)
}
return(find.out.mean)
}

find.out1<-find.dist1(100,40,3)
find.out.fin1<-find.out1[10:40,]
```

```
# Plots of Sum of Coefficients of Variation and Total Variability computed
using the derived formulas and the simulated data
```

```
par(mfrow=c(1,1),pty="s")
plot(3*smry.cv.out,find.out.fin1[,1],xlab="Sum of Coefficients of
Variation computed using Alpha", ylab="Sum of Coefficients of
```



```

Variation computed using the Standard Formula" )

plot(Alpha, find.out.fin1[,1],ylab="Sum of Coefficients of
Variation" ,cex=0.8)
points(Alpha,3*smry.cv.out, pch=2,cex=0.8)
legend(22,0.7,c("SCV Computed using Standard Formula", "SCV computed
using Alpha"), marks =c(1,2),cex=0.8)

plot(var.tiagamma, find.out.fin1[,3], xlab="Total Variability using
Trigamma Function ", ylab="Total Variability using Aitchison
Logratio Transformation")

plot(Alpha, find.out.fin1[,3],ylab="Total Variability" ,cex=0.8)
points(Alpha,var.tiagamma, pch=2,cex=0.8)
legend(17,0.18,c("TOTVAR_Computed using Aitchison Logratio
Transformation", "TOTVAR_computed using the Trigamma Function"),
marks =c(1,2),cex=0.7)

```

Relationship between Total Variability and Sum of Coefficients of Variation

For D=3, N=100 and $\alpha=10$

```

find.dist2<-function(N,col.cnt){
  matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F)
  for(i in 1:N){
    randomgam4_rgamma(100,10)
    randomgam5_rgamma(100,10)
    randomgam6_rgamma(100,10)
    gamcv4_stdev(randomgam4)/mean(randomgam4)
    gamcv5_stdev(randomgam5)/mean(randomgam5)
    gamcv6_stdev(randomgam6)/mean(randomgam6)
    gamscv_sum(gamcv4,gamcv5,gamcv6)
    compgam456_cbind(randomgam4,randomgam5,randomgam6)
    compgam3c_closure(compgam456)
    dircv_apply(compgam3c,2,stdev)/apply(compgam3c,2,mean)
    dirscv_sum(apply(compgam3c,2,stdev)/apply(compgam3c,2,mean))
    var.tot_comp.var(compgam3c)
    atotvar_(sum(comp.var(compgam3c)))/6
    tot.cv2_2*trigamma((2-(dirscv/3)^2)/(3*(dirscv/3)^2))
    smry.all<-cbind(dircv[1],dircv[2],dircv[3], dirscv,atotvar,tot.cv2)
    matr.smry[i,<-smry.all
  }
}

```

```

return(matr.smry)
}

find.out2<-find.dist2(1000,6)
apply(find.out2,2,mean)

plot(find.out2[,5],find.out2[,6], xlab="Total Variability using
Aitchison Logratio Transformation",ylab="Total Variability using
Trigamma Function")

plot(find.out2[,4],find.out2[,6], xlab="Sum of Coefficients of
Variation", ylab="Total Variability using Trigamma Function")

# For D=5, N=100 and Alpha=10

find.dist8<-function(N,col.cnt){
  matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F)
  for(i in 1:N){
    randomgam7_rgamma(100,10)
    randomgam8_rgamma(100,10)
    randomgam9_rgamma(100,10)
    randomgam10_rgamma(100,10)
    randomgam11_rgamma(100,10)
    compgam4_cbind(randomgam7,randomgam8,randomgam9,randomgam10,randomgam11)
    compgam4c_closure(compgam4)
    dircv_apply(compgam4c,2,stdev)/apply(compgam4c,2,mean)
    dirscv_sum(apply(compgam4c,2,stdev)/apply(compgam4c,2,mean))
    var.tot_comp.var(compgam4c)
    atotvar_(sum(comp.var(compgam4c)))/10
    tot.cv2_4*trigamma((4-(dirscv/5)^2)/(5*(dirscv/5)^2))
    smry.all<-cbind(dircv[1],dircv[2],dircv[3],dircv[4], dircv[5],
    dirscv,atotvar,tot.cv2)
    matr.smry[i,<-smry.all
  }
  return(matr.smry)
}

find.out8<-find.dist8(1000,8)
apply(find.out8,2,mean)

plot(find.out8[,7], find.out8[,8], xlab="Total Variability using
Aitchison Logratio Transformation", ylab="Total Variability using
Trigamma Function")

```

```

plot(find.out8[,6],find.out8[,8], xlab="Sum of Coefficients of
Variation", ylab="Total Variability using Trigamma Function")

# For D=7, N=100 and Alpha=10

find.dist9<-function(N,col.cnt){
  matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F)
  for(i in 1:N){
    randomgam12_rgamma(100,10)
    randomgam13_rgamma(100,10)
    randomgam14_rgamma(100,10)
    randomgam15_rgamma(100,10)
    randomgam16_rgamma(100,10)
    randomgam17_rgamma(100,10)
    randomgam18_rgamma(100,10)
    compgam5_cbind(randomgam12,randomgam13,randomgam14,randomgam15,
    randomgam16, randomgam17, randomgam18)
    compgam5c_closure(compgam5)
    dirscv_apply(compgam5c,2,stdev)/apply(compgam5c,2,mean)
    dirscv_sum(apply(compgam5c,2,stdev)/apply(compgam5c,2,mean))
    var.tot_comp.var(compgam5c)
    atotvar_(sum(comp.var(compgam5c)))/14
    tot.cv2_6*trigamma((6-(dirscv/7)^2)/(7*(dirscv/7)^2))
    smry.all<-cbind(dircv[1],dircv[2],dircv[3],dircv[4], dircv[5],
    dircv[6], dircv[7], dirscv,atotvar,tot.cv2)

    matr.smry[i,<]=smry.all
  }
  return(matr.smry)
}

find.out9<-find.dist9(1000,10)

```

Relationship between Total Variability and Sum of Coefficients of Variation for different α s

```

# For D=3, N=100,  $\alpha_1=10$ ,  $\alpha_2=20$ , and  $\alpha_3=30$  using equation (3.16)

# (sqrt(5)+sqrt(2)+sqrt(1))^2 =21.62512
find.dist20<-function(N,col.cnt){
  matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F)

```



```

for(i in 1:N){
  randomgam201_rgamma(100,10)
  randomgam202_rgamma(100,20)
  randomgam203_rgamma(100,30)
  compgam6_cbind(randomgam201,randomgam202,randomgam203)
  compgam6c_closure(compgam6)
  dircv3_apply(compgam6c,2,stdev)/apply(compgam6c,2,mean)
  dirscv_sum(apply(compgam6c,2,stdev)/apply(compgam6c,2,mean))
  var.tot_comp.var(compgam6c)
  atotvar_(sum(comp.var(compgam6c)))/6
  alphahat_(21.62512-(dirscv^2))/(6*(dirscv^2))
  tot.cv2_(2/3)*(trigamma(alphahat)+trigamma(2*alphahat)+trigamma(3*alphahat))
  smry.all<-cbind(dirscv,atotvar,tot.cv2,dircv3[1],dircv3[2],dircv3[3])
  matr.smry[i,<]-smry.all
}
return(matr.smry)
}

find.out20<-find.dist20(1000,6)

# For D=3, N=100,  $\alpha_1=10$ ,  $\alpha_2=50$ , and  $\alpha_3=100$ 

#(sqrt(15)+sqrt(11/5)+sqrt(3/5))^2 =37.58695
find.dist40<-function(N,col.cnt){
  matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F)
  for(i in 1:N){
    randomgam401_rgamma(100,10)
    randomgam402_rgamma(100,50)
    randomgam403_rgamma(100,100)
    compgam7_cbind(randomgam401,randomgam402,randomgam403)
    compgam7c_closure(compgam7)
    dircv3_apply(compgam7c,2,stdev)/apply(compgam7c,2,mean)
    dirscv_sum(apply(compgam7c,2,stdev)/apply(compgam7c,2,mean))
    var.tot_comp.var(compgam7c)
    atotvar_(sum(comp.var(compgam7c)))/6
    alphahat_(37.58695-(dirscv^2))/(16*(dirscv^2))
    tot.cv2_(2/3)*(trigamma(alphahat)+trigamma(5*alphahat)+trigamma(10*alphahat))
    smry.all<-cbind(dirscv,atotvar,
    tot.cv2,dircv3[1],dircv3[2],dircv3[3])
    matr.smry[i,<]-smry.all
  }
  return(matr.smry)
}

```

```

find.out40<-find.dist40(1000,6)

# For D=3, N=100,  $\alpha_1=1$ ,  $\alpha_2=50$ , and  $\alpha_3=100$ 

#(sqrt(150)+sqrt(101/50)+sqrt(51/100))^2=206.8666
find.dist30<-function(N,col.cnt){
  matr.smry<-matrix(0,ncol=col.cnt,nrow=N,byrow=F)
  for(i in 1:N){
    randomgam301_rgamma(100,1)
    randomgam302_rgamma(100,50)
    randomgam303_rgamma(100,100)
    compgam8_cbind(randomgam301,randomgam302,randomgam303)
    compgam8c_closure(compgam8)
    dirscv3_apply(compgam8c,2,stdev)/apply(compgam8c,2,mean)
    dirscv_sum(apply(compgam8c,2,stdev)/apply(compgam8c,2,mean))
    var.tot_comp.var(compgam8c)
    atotvar_(sum(comp.var(compgam8c)))/6
    alphahat_(206.8666-(dirscv^2))/(151*(dirscv^2))
    tot.cv2_(2/3)*(trigamma(alphahat)+trigamma(50*alphahat)+trigamma(100*alphahat))
    smry.all<-cbind(dirscv, atotvar,
    tot.cv2,dirscv3[1],dirscv3[2],dirscv3[3])
    matr.smry[i,<-smry.all
  }
  return(matr.smry)
}

find.out30<-find.dist30(1000,6)

```

Relationship between Total Variability and Sum of Coefficients of Variation for correlated Variables

```

mvcorrelate<-function(x,d){
  p<-dim(x)
  z<-rmvnorm(p[1],cov=d)
  y<-x-x
  oz<-apply(z,2,order)
  for (j in 1:p[2]){
    y[oz[,j],j]<-sort(x[,j])
  }
  y
}

```



```

find.distnew1 <- function(N, col.cnt) {
  matr.smry <- matrix(0, ncol=col.cnt, nrow = N, byrow = F)
  for(i in 1:N) {
    randomgam19 <- rgamma(100, 5)
    randomgam20 <- rgamma(100, 5)
    randomgam21 <- rgamma(100, 5)
    gammax <- cbind(randomgam19, randomgam20, randomgam21)
    gammax.corr <- mvcorrelate(gammax, d)
    compgam9c <- closure(gammax.corr)
    dircv <- apply(compgam9c, 2, stdev)/apply(compgam9c, 2, mean)
    dirscv <- sum(apply(compgam9c, 2, stdev)/apply(compgam9c, 2, mean))
    var.tot <- comp.var(compgam9c)
    atotvar <- (sum(comp.var(compgam9c)))/6
    tot.cv2 <- 2 * trigamma((2 - (dirscv/3)^2)/(3 * (dirscv/3)^2))
    smry.all <- cbind(dircv[1], dircv[2], dircv[3], dirscv, atotvar, tot.cv2)
    matr.smry[i, ] <- smry.all
  }
  return(matr.smry)
}

```

```
find.outnew1 <- find.distnew1(1000, 6)
```

Correlated Gamma random variables using the correlation matrix of the
logratios of the color compositional dataset

```
alrcolour_alr(colour)
```

```

find.distnew7 <- function(N, col.cnt) {
  matr.smry <- matrix(0, ncol= col.cnt, nrow = N, byrow = F)
  for(i in 1:N) {
    randomgam1 <- rgamma(100, 10)
    randomgam2 <- rgamma(100, 10)
    randomgam3 <- rgamma(100, 10)
    randomgam4 <- rgamma(100, 10)
    randomgam5 <- rgamma(100, 10)
    gammax2 <- cbind(randomgam1, randomgam2, randomgam3, randomgam4, randomgam5 )
    gammax.corr2 <- mvcorrelate(gammax2, d1)
    compgam10c <- closure(gammax.corr2)
    dircv <- apply(compgam10c, 2, stdev)/apply(compgam10c, 2, mean)
    dirscv <- sum(apply(compgam10c, 2, stdev)/apply(compgam10c, 2, mean))
    var.tot <- comp.var(compgam10c)
    atotvar <- (sum(comp.var(compgam10c)))/10
    tot.cv2 <- 4 * trigamma((4 - (dirscv/5)^2)/(5 * (dirscv/5)^2))
  }
}

```

```

smry.all <- cbind(dircv[1],dircv[2], dircv[3], dircv[4], dircv[5],
dirscv, atotvar, tot.cv2)
matr.smry[i, ] <- smry.all
}
return(matr.smry)
}

```

```
find.outnew7 <- find.distnew7(1000, 8)
```

Correlated Additive Logistic Normal variables using covariance matrix of the logratios of the hongite compositional dataset

```

find.distnew9<- function(N, col.cnt) {
matr.smry <- matrix(0, ncol =col.cnt, nrow = N, byrow = F)
for(i in 1:N) {
xradlognorm20<- radlognorm(100,4,cov=d20)
dircv <- apply(xradlognorm20, 2, stdev)/apply(xradlognorm20, 2, mean)
dirscv <- sum(apply(xradlognorm20, 2, stdev)/apply(xradlognorm20, 2, mean))
var.tot <- comp.var(xradlognorm20)
atotvar <- (sum(comp.var(xradlognorm20)))/10
tot.cv2 <- 4 * trigamma((4 - (dirscv/5)^2)/(5 * (dirscv/5)^2))
smry.all <- cbind(dircv[1], dircv[2], dircv[3], dircv[4], dircv[5],
dirscv, atotvar, tot.cv2)
matr.smry[i, ] <- smry.all
}
return(matr.smry)
}

```

```
find.outnew9 <- find.distnew9(1000, 8)
```

Correlation between Sum of Coefficients of Variation and Subcompositional Total Variability using Garbage compositional data

3-part Subcompositional Analysis

```

# Sum of Coefficients of Variation
comb<-combinations(8,3)
ww3<-NULL for(i in 1:56){
sub<-cbind(garbage1c[,comb[i,1]],garbage1c[,comb[i,2]],
garbage1c[,comb[i,3]])
scv3<-sum(apply(sub,2,stdev)/apply(sub,2,mean))
ww3<-c(ww3,scv3)
}

```

```

}

# Total Variability
comb<-combinations(8,3)
garbage1ctot3<-matrix(0,ncol=1,nrow=56)
for(i in 1:56){
garbage1ctot3[i,1]_comp.r2(garbage1c,c(comb[i,1],comb[i,2],
comb[i,3]))[1]
}

# 4-part Subcompositional Analysis

# Sum of Coefficients of Variation
comb<-combinations(8,4) ww4<-NULL
for(i in 1:70){
sub<-cbind(garbage1c[,comb[i,1]],garbage1c[,comb[i,2]],
garbage1c[,comb[i,3]],garbage1c[,comb[i,4]])
scv<-sum(apply(sub,2,stdev)/apply(sub,2,mean))
ww4<-c(ww4,scv)
}

# Total Variability
comb<-combinations(8,4)
totalv4<-matrix(0,ncol=1,nrow=70)
for(i in 1:70){
totalv4[i,1]_comp.r2(garbage1c,c(comb[i,1],comb[i,2],
comb[i,3],comb[i,4]))[1]
}

# 5-part Subcompositional Analysis

# Sum of Coefficients of Variation
comb<-combinations(8,5)
ww5<-NULL for(i in 1:56){
sub<-cbind(garbage1c[,comb[i,1]],garbage1c[,comb[i,2]],
garbage1c[,comb[i,3]],garbage1c[,comb[i,4]],garbage1c[,comb[i,5]])
scv<-sum(apply(sub,2,stdev)/apply(sub,2,mean))
ww5<-c(ww5,scv)
}

# Total Variability
comb<-combinations(8,5)
totalv5<-matrix(0,ncol=1,nrow=56)
for(i in 1:56){

```

```
totalv5[i,1]_comp.r2(garbage1c,c(comb[i,1],comb[i,2],  
comb[i,3],comb[i,4],comb[i,5]))[1]  
}
```


REFERENCES

- Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika, 70:57–65.
- Aitchison, J. (1984). Reducing the Dimensionality of Compositional Data Sets. Mathematical Geology, 16:617–635.
- Aitchison, J. (1986, reprint 2003). The Statistical Analysis of Compositional Data. Chapman Hall, first edition.
- Aitchison, J. and Greenacre, M. (2002). Biplots for Compositional Data. Applied Statistics, 51:375–382.
- Bacon-Shone, J. (1992). Ranking methods for compositional data. Applied Statistics, 41:533–537.
- Balakrishnan, N. (1992). Hand Book of The Logistic Distribution. Marcel Dekker INC, New York.
- Barceló, C., Pawlowsky, V., and Grunsky, E. (1996). Some Aspect of Transformations of Compositional Data and the Identification of Outliers. Mathematical Geology, 28:501–518.
- Baxter, M., Beardah, C., Cool, H., and Jackson, C. (2005a). Compositional Data Analysis of Some Alkaline Glasses. Mathematical Geology, 37:183–196.
- Baxter, M., Cool, H., and Jackson, C. (2005b). Further Studies in the Compositional Variability of Colourless Romano-British Vessel Glass. Archaeometry, 47:47–68.
- Baxter, M. and Freestone, I. (2006). Log-ratio Compositional Data Analysis in Archaeometry. Archaeometry, 48:511–531.
- Beardah, C., Baxter, M., Cool, H., and Jackson, C. (2003). Compositional Data Analysis of Archaeological Glass: Problems and Possible Solutions. Archaeometry, 48:511–531.
- Bedecian, A. and Mossholder, K. (2000). On the Use of the Coefficient of Variation as a Measure of Diversity. Organizational Research Methods, 3 No. 3:285–297.

- Billheimer, D., Guttorp, P., and Fagan, W. (1998). Statistical Analysis and Interpretation of Discrete Compositional Data. NRCSE Technical Report Series, No. 011.
- Butler, J. C. (1979). Effects of Closure on the Measure of Similarity between Samples. Mathematical Geology, 11:431–440.
- Chayes, F. (1971). Ratio Correlation. University of Chicago Press, Chicago, Illinois, USA.
- Davis, J. (1986). Statistics and Data Analysis in Geology. Wiley and Sons, New York, second edition.
- Fry, J., Fry, T., and McLaren, K. (1996). Compositional Data Analysis and Zeros in Micro Data. Department of Econometrics, Monash University, G-120.
- Graf, M. (2006). Precision of Compositional Data in a Stratified Two-stage Cluster Sample: Comparison of the Swiss Earnings Structure Survey 2002 and 2004. Survey Research Methods Section, ASA, pages 3066–3072.
- Hijazi, R. and Jernigan, R. (2009). Modeling Compositional Data Using Dirichlet Regression Models. Journal of Applied Probability and Statistics.
- Jackson, D. (1997). Compositional Data in Community Ecology: The Paradigm or Peril of Proportions. Ecology, 78(3):929–940.
- Johnson, N., Kotz, S., and Balakrishnan, N. (1995). Continuous Univariate Distributions. John Wiley and Sons, New York, second edition.
- Martín-Fernández, J., Barceló-Vidal, C., and Pawłowsky-GlahnBeardah, V. (2000). Zero Replacement in Compositional Data Sets. In Kiers, H., Rasson, J., Groenen, P., and Shader, M., (Eds.): Studies in Classification, Data Analysis and Knowledge Organisation. Proceedings of 7th Conference of the International Federation of Classification Societies, pages 155–160.
- Martín-Fernández, J., Barceló-Vidal, C., and Pawłowsky-GlahnBeardah, V. (2003). Dealing with Zeros and Missing Values in Compositional Data Sets using Non-parametric Imputation. Mathematical Geology, 35:253–278.
- 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, 60:489–498.
- Rathje, W. (2005). The Garbage Project and The Archaeology of Us. Metamedia Lab at Stanford.

- Reed, G., Freyja, L., and Meade, B. (2002). Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. Clinical and Diagnostic Laboratory Immunology, 9(6):1235–1239.
- Rock, N. (1988). Numerical Petrology. Springer-Verlag, Berlin.
- Sarmanov, O. and Vistelius, A. (1959). On the Correlation of Percentage Values. SSSR, 126:22–5.
- Tauber, F. (1999). Spurious Clusters in Granulometric Data Caused by Logratio Transformation. Mathematical Geology, 31:491–504.
- Valls, R. (2008). Why and How We Should Use Compositional Data Analysis: A step-by-step Guid for the Field Geologists. Wikibooks, Toronto.
- Wu, J., Hung, W., and Lee, H. (2000). Some Moments and Limit Behaviors of the Generalized Logistic Distribution with Application. Proceedings of the National Science Council, ROC(A), 24(1):7–14.
- Zhang, L., Albaréde, S., Dumont, G., Campenhout, C., Libeer, J., and Albert, A. (2010). The Multivariate Coefficient of Variation for Comparing Serum Prorein Electrophoresis Techniques in External Quality Assessment Schemes. Accred Qual Assur, 15:351–357.