

https://dspace.org/

Evaluating the reliability of quantification results

Feni, Democracy Yongama

2021
https://hdl.handle.net/10539/33051

Downloaded from a DSpace repository. https://dspace.org/

Evaluating the reliability of quantification results

Democracy Yongama Feni
(1455373)



School of Computer Science and Applied Mathematics
Faculty of Science
University of the Witwatersrand, Johannesburg.

Supervisor
Prof Benjamin Rosman

Co-Supervisor
Dr Pravesh Ranchod

Co-Supervisor
Ms Nyalleng Moorosi

A thesis submitted to the Faculty of Science, University of the Witwatersrand, in fulfillment of the requirements for the degree of Masters of Science

April 30, 2021

Declaration

I, Yongama Feni (student number: 1455373) declare that I am a registered student for MSc in Computer Science. I hereby declare that:

- I am aware that plagiarism (failure to acknowledge the ideas or writing of another as one's own) is a serious offence.
- The work submitted for assessment is my own except where I have explicitly indicated otherwise. I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signed: 

Date: 27/04/2021

Abstract

Categorical data analysis is often carried out using different techniques, one of which is quantification. Quantification refers to a collection of methods that assign values to the categories of the categorical variables. Assignment of these values means that categorical variables can now be treated as if they were continuous variables, that is, methods that are used for continuous data can be applied directly to this data. If these values are consistent in the way they are assigned to these categories, they are said to be stable. On the other hand, if these assigned values are not stable then inference, and specifically any subsequent classification will also be unstable.

Resampling methods such as bootstrapping have been used to assess the stability of quantification results in the past. These methods have typically been used for contingency tables whereas multiple-choice datasets have received less attention. Furthermore, the stability of quantification results has been evaluated in the context of visualization. This research investigated the stability of quantification results under different sampling strategies including subsampling and resampling. In addition to evaluating the stability in the context of visualizations, in this study stability was also assessed in terms of classification.

In this research, we used Dual Scaling (DS) to accomplish quantification. To investigate stability, we used different subsampling strategies, specifically, uniform and biased subsampling. In addition to subsampling, bootstrapping was used to further investigate the variation of results. Finally, we used Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDA) to perform classification on the quantified data to evaluate the effect of changing the data on quantification results.

The results showed that DS is stable under uniform subsampling and resampling. Quantification results were similar to the parent set for larger sample sizes in the uniform subsampling strategy whereas the sample size affected the stability of DS results for biased subsampling, that is, the results deviated largely from the parent data when the data is extremely biased. Contrary to our knowledge of larger sample sizes yielding better results, when subsampling a variable with two categories of unequal rows, our study showed that the category with fewer rows gave better results. This was tested on the “Gender”, “Income” and “Employment” variables and the same conclusions were reached. The same experiment was applied to the 2017 General Household Survey (GHS) data on “ Gender” variable and the results still yielded the same conclusions. The biased subsampling on evenly distributed categories yields better results than subsamples with a skewed distribution of categories.

Although results may indicate instability in the DS method for biased subsampling, the clustering and classification results showed that the quantification results were indeed stable since the classification results did not show a variation to any great extent and the clustering results retained the same groupings under difference subsampling strategies. Therefore, DS can be used for classification and other related techniques since it is stable.

Resampling methods, on the other hand, strengthen the results of either the parent sample or the subsamples, that is, resampling validates the underlying results of a study through the repetition of experiments. Finally, although resampling methods are useful for assessing the stability of quantification results, they should never be used as a substitution for gathering more data, this was shown in the experiment with intentions to investigate the effect of resampling up to a certain sample size greater than the sample size of the subsample.

Dedication

I dedicate this dissertation to my family, my wife and my friends for giving me tireless support and encouragement when I needed it most.

Acknowledgements

I would like to thank God for all the strength He has given me to cope with the pressure and to work earnestly towards the completion of this dissertation. I would have never finished this thesis without the guidance of my supervisors, help from friends, colleagues and support from my family.

I would like to express my deepest and sincere gratitude to my supervisor, Dr Benjamin Rosman for providing me with all the guidance and the valuable inputs he had given me throughout the course of this dissertation and the excellent skills and wisdom he has given to me about the research in general.

It would not have been possible for me to finish this dissertation without the suggestions and the technical guidance of Ms Nyalleng Moorosi. I would like to express my sincere gratitude to Nyalleng for her tireless support throughout this masters.

I would sometimes be convinced that I am articulate with my work and it is understood but I would miss minor things and with disappointment I would find that some people do not understand. Dr Pravesh Ranchod always brought a new perspective in my work and for that, I would like to thank him.

Venturing to Computer Science for me was one adventure in which I left my comfort zone and I had my supervisors, colleagues and friends at the Council for Scientific and Industrial Research (CSIR) to give me all the help I cried for. The data science team in CSIR, the Robotics Autonomous Intelligence Learning (RAIL) at Wits has gave me valuable feedback in my work and I will forever be grateful for their inputs.

This work would not have been possible without the financial support I got from CSIR. I will forever be grateful for being chosen to be part of the studentship programme.

Contents

Declaration i

Abstract ii

Dedication iv

Acknowledgements v

Contents vi

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Problem Statement . . . 5
- 1.3 Hypotheses Statements . . . 6
- 1.4 Research Questions and Objectives . . . 6
  - 1.4.1 Questions . . . 6
  - 1.4.2 Objectives . . . 7
- 1.5 Delineation and Limitations . . . 7
- 1.6 Significance . . . 8
- 1.7 Brief Chapter Overviews . . . 9

2 Background and Related Work 10
- 2.1 Introduction . . . 10
  - 2.1.1 Chapter Organization . . . 10
- 2.2 Quantification Methods . . . 10
- 2.3 Singular Value Decomposition . . . 11
- 2.4 Dual Scaling . . . 13
  - 2.4.1 DS for contingency tables . . . 13
  - 2.4.2 DS for multiple-choice data . . . 17
  - 2.4.3 DS for rank-order data . . . 21
    - 2.4.3.1 Example of DS for rank-order data . . . 23
- 2.5 Correspondence Analysis . . . 25
  - 2.5.1 CA for contingency tables . . . 25
  - 2.5.2 CA for multiple-choice data . . . 27
  - 2.5.3 CA for rank-order data . . . 28
- 2.6 Relationship between DS and CA . . . 31
- 2.7 Resampling Techniques . . . 32

2.7.1 Bootstrap . . . 32
2.7.2 Relationship between Bootstrap and Jackknife . . . 35
2.7.3 Permutation Tests Resampling . . . 35
2.8 Related Research . . . 37
2.9 Conclusion . . . 40

3 Research Methodology **42**
3.1 Introduction . . . 42
3.1.1 Chapter Organization . . . 44
3.2 Data . . . 44
3.2.1 Data Generation . . . 44
3.2.1.1 Probabilistic Graphical Models (PGMs) . . . 44
3.2.2 Data Description . . . 46
3.3 Quantification . . . 49
3.4 Subsampling . . . 50
3.5 Resampling . . . 50
3.6 Classification . . . 51
3.7 Experiments . . . 52
3.7.1 Uniform subsampling vs the original data . . . 53
3.7.2 Biased subsampling vs original data . . . 54
3.7.3 Uniform vs biased subsampling . . . 55
3.7.4 Resampling parent data vs the original data . . . 56
3.7.5 Resampling vs biased Subsampling . . . 57
3.7.6 Resampling vs uniform subsampling . . . 58
3.7.7 Similarity and Distance Measures . . . 59
3.7.8 Hierarchical clustering . . . 60
3.8 Conclusion . . . 61

4 Experimentation **62**
4.1 Introduction . . . 62
4.1.1 Chapter Organization . . . 62
4.2 DS of the parent sample . . . 63
4.2.1 Similarity and Distance Measures . . . 65
4.3 Classification of the parent data . . . 69
4.4 Experiment 1: Uniform subsampling vs the original data . . . 70
4.4.1 Visual inspection . . . 70
4.4.2 Comparisons . . . 75
4.5 Experiment 2: Biased subsampling vs original data . . . 77
4.5.1 Visual Inspection . . . 78
4.5.2 Comparisons . . . 81
4.6 Experiment 3: Uniform subsampling vs biased subsampling . . . 84
4.6.1 Visual Inspection . . . 84
4.6.2 Comparisons . . . 86
4.7 Experiment 4: Resampling vs parent sample . . . 90
4.7.1 Visual Inspection . . . 90
4.7.2 Comparisons . . . 91
4.8 Experiment 5: Resampling vs Biased Subsampling . . . 92

4.8.1 Visual Inspection . . . 92
4.8.2 Comparisons . . . 94
4.9 Experiment 6: Resampling vs uniform Subsampling . . . 97
4.9.1 Visual Inspection . . . 98
4.9.2 Comparisons . . . 99
4.10 Summary of results . . . 101
4.11 Conclusion . . . 102

5 Summary and Discussion 103
5.1 Introduction . . . 103
5.1.1 Chapter Organization . . . 103
5.1.2 Subsampling . . . 103
5.1.3 Resampling . . . 106
5.2 Discussion . . . 107
5.2.1 Extension of DS to other data types . . . 108
5.2.2 Classification . . . 109

6 Conclusion and Recommendations 110
6.1 Introduction . . . 110
6.2 Conclusion . . . 110
6.3 Recommendations . . . 111

A Supplementary Results 113
A.1 Additional Experiment Results . . . 113
A.1.1 Experiment 1 supplementary results . . . 113
A.1.2 Supplementary Biased Subsampling . . . 116
A.1.3 Employment . . . 119

B Application to 2017 GHS data 122
B.1 Application on General Household Survey 2017 data . . . 122
B.2 Data Description . . . 122
B.2.1 Dual scaling of the 2017 GHS data . . . 124
B.2.2 Experiment 1: Uniform subsampling vs parent data (GHS) . . . 127
B.2.3 Experiment 2: Biased subsampling vs parent data (GHS) . . . 130
B.2.4 Experiment 3: Uniform subsampling vs Biased subsampling . . . 132
B.2.5 Experiment 4: Resampling vs parent data (GHS) . . . 135
B.2.6 Experiment 5: Resampling vs Biased subsampling . . . 136
B.2.7 Experiment 6: Resampling vs Uniform subsampling . . . 138
B.2.8 Summary of results . . . 141

C Similarity and Distance Measures 142
C.1 Cosine Similarity of the original parent data . . . 142
C.2 Euclidean distance of the parent data . . . 142

Bibliography 151

Chapter 1

Introduction

1.1 Background

People are often asked to evaluate different services, sort objects according to their similarities, evaluate lectures, training sessions and workshops, etc. All of these lead to the data collected being categorical. Categorical data can be divided into two groups: nominal and ordinal. Nominal refers to a set of categories without any order, while ordinal refers to a set of categories with some order, such as the Likert scale. An example of a variable from a Likert scale is one with questions such as "Do you want free WiFi services for everyone?", where the responses are *strongly agree, agree, no comment, disagree, strongly disagree*. The distinction between these kinds of data is important because each may require a specific analytic method in order to be analyzed successfully (Johansson Fernstad, 2011).

Whether it be questionnaires, evaluation forms or a census, this data is very important across various fields such as computer science, statistics, behavioural sciences, market research and much more. It is common for questionnaires to require biographic details of respondents, details such as the gender, race, ethnicity, citizenship and the town and province of birth. The gender of a person can either be female or male, while the race is Black, White, Asian or Coloured. Similarly, the province of birth is categorical and can take as many discrete values as relevant. As an example, South Africa has nine provinces, therefore we may expect to get nine provinces in the questionnaire data in South Africa's 2011 census. We may also have questions with choices that have a natural ordering, such as the level of education. The level of education, in this case, may take three categories: primary, secondary and tertiary. In this manner we may know that secondary is a level which is higher than primary, however, this knowledge does not give

us an understanding as to how much more important the secondary is over the primary level.

Distance measures between continuous variables are very important in many data mining tasks and knowledge discovery (Boriah et al., 2008). Unfortunately, categorical variables do not have distance metrics, and as a result, there is a need to understand the different techniques that can be used to analyze this kind of data. On the other hand, it is easy to find similarities between continuous variables since there exist a variety of distance measures, such as euclidean (Cha, 2007), Mahalanobis and correlation distance (Boriah et al., 2008; Anandan and Uthra, 2017). Most of these distance measures can be converted to similarities or vice versa (Singh, 2017).

We must understand how different variables are associated with each other, as this provides a similarity measure which is important for data mining tasks such as clustering. Consider an example of the nominal variable, province. An interesting question may be to find out how provinces relate to each other. This would be easy with distance metrics, that is, the closer the physical distance between two provinces the more similar they are. This is currently difficult with categorical variables (Boriah et al., 2008).

Some statistical procedures that can be used for analyzing categorical data include Fisher's exact test (McHugh, 2013) and Pearson's chi-squared test (Connelly, 2016). These two tests are used to examine whether there is a relationship between two categorical variables. The former is used when the sample size is small and the latter is used for larger sample sizes. Both methods may not be sufficient when we wish to know the strength of the association, i.e., whether the existing relationship is strong or weak (Johansson Fernstad, 2011). Additionally, both methods are highly dependent on assumptions such as sample size, independence, etc. Furthermore, they cannot handle datasets with more than two variables as they are used specifically to analyze a two contingency table [1]. To mitigate this, other common methods have been developed to understand categorical data. Quantification refers to a collection of such methods. Quantification is defined as a set of related methods that are used to analyze categorical data by assigning values to the rows and columns of a data table. These values are computed in such a way that there is high discrimination between dissimilar categories while similar categories are grouped. In addition, quantification methods are used to find a lower-rank matrix of an appropriately transformed data table.

Quantification can be traced back to the 1940's, pioneered by Guttman (1941), Hayashi (1950) and Nishisato and Sheu (1980). These methods are convenient in several settings

[1] Contingency tables are discussed in Section (2.5.1). These tables can also be seen as the summary of multiple-choice data, where the rows show a distribution of one variable and the columns show a distribution of another.

of interest, whether the objective is reducing the dimensionality of data, maximizing the correlation, or performing categorical classification. Quantification methods assign values to the categories of categorical variables, by maximizing the variance between the categories, that is, they assign similar values to the respondents with similar answers to the questions. In this way, methods that are used for continuous numeric data may be used directly on the assigned values. In other words, quantification methods are used to give a numerical interpretation of some categorical data. Quantification techniques can thus enable us to visualize the relationships between the categorical variables.

Multiple quantification techniques are used to understand categorical data and these methods are conceptually related and lead to one solution. For this study, we focus on Dual Scaling (DS), which is a multivariate exploratory method for computation of numerical values for rows and columns of the data table. This method has been widely used in the literature (Johansson Fernstad, 2011; Nishisato, 2014; Greenacre, 2007; Saporta and Keita, 2006; Jose and Scholar, 2014; Blasius and Greenacre, 2006; van de Velden et al., 2017) and it is generally intuitive and applicable to a variety of categorical data (Nishisato, 1998).

Quantification results may largely be affected by some properties of the data such as sample size and data imbalance or biased sampling. For example, if the sample size is very small then the interpretation of results change dramatically, however, for larger sample sizes the interpretation remains similar. If the quantification methods are consistent in the way they assign numerical values, then they are said to be stable. On the other hand, if the quantification methods are not consistent in the way they assign numerical values then they are said to be unstable. If the assigned values are not stable then any further inference specifically, the classification may not be stable, that is, small unimportant changes in the data may cause changes in classification results (Gifi, 1990; Nishisato, 2006). Classification is highly dependent on several parameters such as sample size, the number of features and class imbalance (Hua et al., 2004; Sordo and Zeng, 2005; Chu et al., 2012).

Stability of quantification results has been studied by many researchers, and different tools have been used by researchers in order to assess the stability of quantification results (Linting et al., 2007; Markus, 1994; Lebart, 2006; Nishisato and Hillsdale, 1994; Takane, 1994).

The general approach to assessing the stability of quantification results is through the use of resampling techniques such as Bootstrapping and Jackknife. Resampling techniques are statistical methods that are used to draw samples from a data table, and then to assess the stability of solutions. In this study, we use Bootstrapping together with some subsampling strategies, such as uniform and biased to get more insight into the stability

of quantification under different sampling conditions. This is a desirable facet of stability since it provides a wide understanding of different conditions where the DS results are stable and where they may have limitations. To strengthen the results, permutation tests are used to validate results. The permutation test is an alternative resampling method which tests the stability of the quantification method by shuffling the resamples and reapplying DS technique on shuffled datasets.

Despite the considerable amount of work that has been devoted to addressing the issue of stability of quantification results, the conditions under which the quantification methods give stable and reliable results are still not clear. Hence in this research, we use Bootstrapping and subsampling strategies to investigate the effect of different sampling strategies on quantification.

The past research assessed the stability of quantification results on two kinds of categorical data, that is, rank-ordered data (Xu, 1999) and contingency tables (Greenacre, 2007; Lombardo et al., 2012). In this research, the results of quantification are extended to another kind of categorical data, specifically, multiple-choice data. Furthermore, stability has been studied in a context of visualization, i.e, that is how visualization is affected by mixing of rows in the data.

In this research, DS is used to accomplish quantification on multiple-choice data. After obtaining the quantification Bootstrapping and subsampling strategies to test the stability of the quantification results. Specifically, Bootstrap is used to draw several samples from the given data, and after that, quantification is estimated on these samples. Moreover, subsampling is used to draw several subsamples that have smaller sample sizes compared to the original data, and then apply DS on each of these subsamples. The focus is given to the quantification results as well as the investigation of whether quantification methods yield consistent results regardless of the changes made in the data. The changes, in this case, refer to the mixing of rows and columns, i.e, the alteration to the original data table through subsampling and resampling. To minimize the sampling error and possible variation in the results, the 95 % Confidence intervals together with error bars for all repeated experiments.

To evaluate the effect of different sampling strategies, Support Vector Machines (SVM) and Fisher's Linear Discriminant Analysis (LDA) are further used to perform classification on the quantified data. Fisher's LDA has been used in quantification in the past while SVM is one of the most effective classification algorithms (Rebentrost et al., 2014).

Finally, the classification results of LDA and SVM on the subsampled and resampled data are compared using precision, accuracy and recall measures for classification performance. Additionally, the similarity and distance measures are also used as a way of directly applying the machine learning techniques to a quantified data. In this study Cosine similarity and Euclidean distance was used along with the one-mode plot. Cosine similarity is defined as a measure of similarity between two high-dimension vectors (Kalhori et al., 2018). Euclidean distance is a measure of dissimilarity between two vectors (Wang et al., 2019). The relationship between these two is indirectly proportional, that is as the distance between two vectors increases the similarity decrease. These two metrics are generally intuitive and simple to interpret. It is worth noting that Hierarchical clustering was applied on the column weights to aid interpretation of distance and similarity measures [2]. Hierarchical clustering is a clustering method which finds nested groupings in data by using either furthest or shortest distances of some distance measures such as Euclidean distance (Govender and Sivakumar, 2020).

1.2 Problem Statement

There are multiple problems associated with quantification, and although it has been considerably researched, to the best of our knowledge, there are no sufficient methods that describe the conditions under which quantification methods, specifically, Dual Scaling (DS) give stable results. Stable results refer to consistent results, such that the patterns observed in the sample can be generalized to a larger population. More precisely, the question is whether the patterns observed in the data are real or they occurred by mere chance (Greenacre, 2007). Stable quantification results refer to those results which are not affected by changes in the data. The changes in this study refer to the mixing of rows and columns of a data table using either subsampling or resampling. The resampling strategies allow for testing of the stability of solutions across different samples, that is, how would the quantification results change if the quantification was applied to the samples which were collected similarly as the data at hand, that is, samples that are drawn from the same distribution as the sample at hand, that is, what would be the quantification of similar datasets with some minor differences from the data at hand.

However, resampling does not uncover other aspects of sampling such as the effects of different sample sizes, biased sampling and uniform sampling on the quantification results. Furthermore, although resampling methods have been applied to contingency tables (Lombardo et al., 2012; Michailidis and de Leeuw, 1998), multiple-choice data has

[2] Hierarchical clustering is discussed in Chapter 3.

not been given enough attention. Multiple-choice datasets present a challenge because of their nature, as they are normally stored in large tables with thousands of rows, and it would be interesting to predict the results of a population given the available data. There are no benchmarks that indicate the stability of quantification methods across different sampling conditions. Thus, subsampling can give us an insight into different conditions of the data that may affect the interpretation of DS results.

1.3 Hypotheses Statements

The hypotheses of this research are as follows:

- There is a significant difference between the quantification results of the subsampled and the original parent data.

  The quantification results that are considered in this study include the eigenvalues and the optimal weights. The hypothesis will be tested on different subsampling strategies, in particular, uniform and biased subsampling. The eigenvalue is also known as the squared correlation ratio, η^2, which is a measure of information loss by representing high-dimensional data into lower-dimensional data. The optimal weights refer to the quantification of rows and columns, and emphasis will be placed on the quantification of columns, i.e., the quantification of categories since they are easy to interpret.

- There is a significant difference between the quantification results of the resampled and the original parent data.

  The results considered in this case also will be the eigenvalues and optimal weights, specifically the column weights.

1.4 Research Questions and Objectives

1.4.1 Questions

- Can we discover some sampling conditions such as uniform subsampling, biased subsampling and resampling, which affect the stability of quantification results?

  That is, using Bootstrapping and some subsampling can we discover the effect of different sampling conditions on the stability of quantification results?

- How do different sampling conditions such as uniform subsampling, biased subsampling and resampling affect the classification results of LDA and SVM on quantified data?

1.4.2 Objectives

The objective of this research is to:

- Investigate under what sampling conditions quantification methods, and DS, in particular, provide stable results. The sampling conditions that are considered in this research are as follows:
  - Uniform subsampling
  - Biased subsampling
  - Resampling
- Compare the classification results of LDA and SVM on the quantified results of the subsampled data and the resampled data.

  That is, compare how the classification results change as a function of data, i.e, the amount of data, as the stability in the classification may indicate the stability of the quantification results. The change in the data, in this case, refers to any alteration in the data by subsampling and resampling.
- Investigate how different properties of the data can affect the stability of quantification results.

  The specific property that is considered in this study is the sample size

  The sample size is a desirable property since it affects the classification results (Popovici et al., 2010; Chu et al., 2012; Sordo and Zeng, 2005; Figueroa et al., 2012). In theory, it is expected that the larger the sample, the closer the results will be to the results of the full population data. However, instabilities may occur in the quantification results if the sample size does not represent or characterize the population (Linting et al., 2007).

1.5 Delineation and Limitations

Since CA, Multiple Correspondence Analysis (MCA) and DS are conceptually related, in this research one quantification method: DS, in particular, is considered. This method

can handle a wider variety of data types than MCA and CA. However, the results of this research may be generalized to other quantification methods, specifically MCA and CA. Quantification methods can quantify categorical variables, categories of these variables and the individual respondents, i.e, rows, depending on the needs of a researcher. In this study, only the quantification of categories and rows is considered.

This study only focuses on multiple-choice data since the stability of DS for multiple-choice data has not been well studied as compared to other data types, specifically, contingency tables and rank-ordered data.

1.6 Significance

Stability of results is not a new concept in multivariate analysis. However, the investigation of stability for quantification results has also been focused on contingency tables. This research aims to draw on and extend the stability analysis to other types of categorical data, in particular, multiple-choice data. The considerable research on the stability of quantification results has been focused on one context of stability i.e., visualization. In this study, we evaluate quantification in the context of classification.

Thus, this study not only intends to assess stability in terms of visualization, it also evaluate the stability of quantification results in the context of classification. Contingency tables, in essence, can also be seen as the summary of the multiple-choice data. Hence the attention in this research is through employing resampling methods together with subsampling strategies to assess the stability of DS in different ways.

To summarize the contributions of this research they are divided into the following components:

- Datasets - In addition to what has been contributed in the field of quantification, this research examines the stability of quantification results in multiple-choice data.
- Unsupervised learning application - This study employs some similarity and distance measures to the quantified data to get more insights into the data. Moreover, also as a way of assessing stability. These similarity/distance measures are further backed up by hierarchical clustering method which aids in interpretation.
- Context - This research further explores the use of classification methods as a way of evaluating the stability of quantification results. The study does not aim to diminish the work on visualization but further looks on classification as an alternative context.

- Subsampling - In addition to using resampling methods such as Bootstrapping, this study further investigates different sampling strategies, such as the effect of sampling uniformly and biased sampling.

1.7 Brief Chapter Overviews

The stability of quantification results opens new opportunities for the use of the results in the advanced analysis of multivariate data. To explore and investigate the stability of quantification, several experiments and perturbations on the categorical datasets are undertaken.

Thus Chapter 2 introduces the background on the methods that are used in this study. It begins by describing the quantification methods on different types of categorical data. Because all these methods employ the Singular Value Decomposition (SVD), a brief introduction to SVD is discussed in the background section. The background of the resampling methods is also discussed as is the relationship between the different types of quantification methods. Both quantification methods, i.e., Correspondence Analysis and Dual Scaling are described as used in this research. Once the background information and the related research is thoroughly discussed, the existing literature both in resampling and quantification is considered and reviewed.

Chapter 3 presents the methodology that is followed in this research to answer the aforementioned research questions. In this chapter, the datasets used in the research are defined. Following the data section, the overall methodology and experimental setup are discussed. The last part of this chapter presents the evaluation for the stability of quantification by employing classification.

Chapter 4 presents the experimentation and analysis of results. This entails the results of quantification for different categorical datasets and the analysis of quantification across the subsamples and over the resampled datasets. Again the last part of this chapter emphasizes the evaluation of the stability through the use of classification, specifically binary classification using LDA and SVM on the quantified data from the parent sample, across the subsamples and on the resampled datasets.

Chapter 5 presents the discussion and the summary of results, and the general findings of the stability of quantification results. Finally, chapter 6 suggests some recommendations and conclusions for the entire study.

Chapter 2

Background and Related Work

2.1 Introduction

The analysis of categorical data is often carried out in different ways depending on the objective of the analysis. One of the insightful methods of analyzing categorical data is through quantification techniques. The purpose of this section is to review the different quantification methods and describe the methods used to assess the stability of these quantification methods.

2.1.1 Chapter Organization

This chapter is organized as follows. The quantification methods are defined in Section 2.2. Since these methods are heavily dependant on the Singular Value Decomposition (SVD), it is introduced in Section 2.3. Section 2.4 describes the technical details of Dual Scaling while Section 2.5 describes the technical details of Correspondence Analysis (CA). Following this, resampling methods are introduced and discussed in Section 2.7 and finally the current state of literature pertaining to the stability of quantification results is considered in Section 2.8. It is worth noting that in this study we shall use Bootstrap as a resampling method. Finally, a summary of the state of the art and hence identify the literature gaps is given in Section 2.9.

2.2 Quantification Methods

Quantification is a set of methods that can be used to analyze categorical variables by assigning meaningful weights to the rows and the columns of categorical data. The

weights are computed in such a way that a certain criterion is optimized, such as maximizing the correlation ratio, η^2. Quantification can be traced back to Guttman's paper (Guttman, 1941) and later proposed as quantification theory by Hayashi (Hayashi, 1950, 1951). Several quantification methods were then popularized in different schools, including Correspondence Analysis (Greenacre, 1984), Dual Scaling (Nishisato and Hillsdale, 1994), homogeneity analysis (Michailidis and de Leeuw, 1998) and Non-Linear Principal Component Analysis (NLPCA) (Linting et al., 2007).

The detailed historical background on quantification methods is discussed in Nishisato's paper (Nishisato, 1996). Not only does this paper discuss a detailed history, but it also clarifies the mathematical differences in these formulations.

Quantification methods are convenient in several settings of interest, whether the objective is reducing the dimensionality of data, maximizing the correlation or performing categorical classification. In this way, methods that are used for continuous numeric data may be used directly on the assigned values. In other words, quantification methods are used to give numerical meaning to some categorical data. Quantification techniques can thus enable us to visualize the relationships between categorical variables.

Quantification may also be considered as the "PCA for qualitative data", where the quantification aims to extract important information from the table of data and represent it in a lower dimension space, represented as a linear combination of categories that are orthogonal to each other. This generalization of PCA gives rise to modern quantification methods such as correspondence analysis. It is thus important to note that quantification depends on some eigendecomposition or Singular Value Decomposition (SVD) (see Section 2.3) of rectangular matrices, as does PCA.

2.3 Singular Value Decomposition

Singular Value Decomposition (SVD) remains one of the most important procedures in multivariate data analysis. As the method has been used in different applications such as PCA, cluster analysis and other ordination techniques [1], the purpose of this section is to describe SVD in relation to quantification methods. SVD is a factorization method that is used to find a lower rank of a matrix. This property is used in quantification to obtain the eigenvalues and eigenvectors of a data matrix. Thus, eigenvalues and eigenvectors help us understand the structure of the data and finally helps us compute the row and column weights, which are referred to as optimal weights.

[1] Ordination techniques in statistics are methods which are used primarily for exploratory data analysis, i.e, descriptive statistics as opposed to inferential statistics (hypothesis tests)

Let $\mathbf{A}$ be a real $m \times n$ matrix, often with $m \geq n$.

$\mathbf{A}$ can be decomposed into three sub-matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ in the following manner,

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{2.1}$$

where $\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathbf{T}} = \mathbf{I_n}$ and $\mathbf{\Sigma} = \mathbf{diag}(\sigma_{\mathbf{1}}, ..., \sigma_{\mathbf{n}})$.

In which case $\mathbf{U}$ contains n orthonormalised eigenvectors associated with the n largest eigenvalues of $\mathbf{A}\mathbf{A}^{\mathbf{T}}$ and $\mathbf{V}$ contains the orthonormalised eigenvectors of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, and the diagonal elements of $\mathbf{\Sigma}$ are the singular values of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$. It is important to note that the singular values in $\mathbf{\Sigma}$ are in an ascending order, that is
$\sigma_1 \geq \sigma_2, ..., \geq \sigma_n \geq 0$.
SVD has the interesting property of giving an optimal low rank approximation of a matrix, that is a matrix that has a lower dimension but with the maximum variance. For the rest of this study, the singular values σ_n are referred to as either ρ_k or η_k. The data or data tables $\mathbf{A}$ referred to $\mathbf{F}$ for the sake of readability. The SVD can also be expressed as a bi-linear expansion in the following equation:

$$f_{ij} = \frac{f_{i.}f_{.j}}{f_t}(1 + \rho_1 y_{1i}x_{1j} + \rho_2 y_{2i}x_{2j} + \rho_3 y_{3i}x_{3j} + ... + \rho_K y_{Ki}x_{Kj}), \tag{2.2}$$

where $f_{i.}$, $f_{.j}$ and f_t are row totals, column totals and grand total, respectively.

This decomposition results in the following dual equations,

$$y_{jk} = \frac{1}{\rho_k}\frac{\sum_{j=1}^{m} f_{ij}x_{jk}}{f_{i.}} \qquad \text{and} \qquad x_{ik} = \frac{1}{\rho_k}\frac{\sum_{i=1}^{n} f_{ij}y_{ik}}{f_{.j}}. \tag{2.3}$$

Multiplying both sides of the equations by singular values, ρ_k, we then obtain the means of weighted responses which are also called projected weights or principal coordinates, $\rho_k y_{jk}$ and $\rho_k x_{ik}$, respectively. The row and column weights can be interpreted from a Euclidean space, which means that SVD allows us to explore. The Young-Householder theorem as depicted in the SVD ensures that the projected weights are orthogonal and preserve the Euclidean distance (Kaplan, 2004; Clavel and Nishisato, 2012).

In the following sections, the quantification methods in particular Dual Scaling (DS), Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA) is discussed and all these methods employ the SVD algorithm for computation of rows and columns weights.

2.4 Dual Scaling

Dual Scaling (DS) is a multivariate descriptive method that falls under the umbrella of quantification. The name *Dual Scaling* was first coined by (Nishisato, 1980) although the theory can be traced back to original work by Guttman (Guttman, 1941).

DS is best known for handling diverse categorical data. The basic table that is analyzed by DS is called a contingency table, a case when we have at most two categorical variables. There exist some other tables or kinds of categorical data including multiple-choice data, sorting data and rank-order data. In the following sub-sections, we explain exactly how these tables differ and how to handle them using both the correspondence analysis and dual scaling framework.

2.4.1 DS for contingency tables

Consider a contingency table in Table 2.1 and let $\mathbf{F} = (\mathbf{f_{ij}})$ be a data table with m rows and n columns, where the entries $\mathbf{f_{ij}}$ represent the non-negative frequency data points.

TABLE 2.1: Contingency table of teachers' performance being evaluated by students

	Good	Average	Poor
White	1	3	6
Brown	3	5	2
Green	6	3	0

Let $\mathbf{r}$ represent the vector of row totals, $\mathbf{r} = \begin{pmatrix} 10 \\ 10 \\ 9 \end{pmatrix}$ and $\mathbf{c}$ represent the vector of column totals, $\mathbf{c} = \begin{pmatrix} 10 \\ 11 \\ 8 \end{pmatrix}$

Let $\mathbf{D_r}$ and $\mathbf{D_c}$ in Table 2.2 and Table 2.3 represent a diagonal matrix of row and column totals, respectively.

TABLE 2.2: $\mathbf{D_r}$: A diagonal matrix of row totals

	White	Brown	Green
White	10.0	0.0	0.0
Brown	0.0	10.0	0.0
Green	0.0	0.0	9.0

TABLE 2.3: $\mathbf{D_c}$: A diagonal matrix of column totals

	Good	Average	Poor
Good	10.0	0.0	0.0
Average	0.0	11.0	0.0
Poor	0.0	0.0	8.0

Then let $\mathbf{f_t}$ represent the grand total

$$\mathbf{f_t} = \sum_{\mathbf{i=1}}^{\mathbf{m}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{f_{ij}} \tag{2.4}$$

In our example in Table 2.1, $\mathbf{f_t} = \mathbf{29}$. Let $\mathbf{y}$ and $\mathbf{x}$ be the vector of weights for the rows and columns of $\mathbf{F}$, respectively, $\mathbf{y} = (y_1, y_2, ..., y_n)$ and $\mathbf{x} = (x_1, x_2, ..., x_n)$.

There are various ways of computing the optimal weights such as SVD, the power method and the Alternating Least Squares (ALS) algorithm, and all these lead to the same solutions. We can find optimal weights that maximize certain criteria, such as maximizing the ratio of the between and within correlation, maximizing the variance and maximizing the product-moment correlation (Nishisato, 2014).

The purpose of applying DS to a contingency table is to find a sound way to assign a reasonable score to each of the rows while assigning scores for columns, and hence it is sometimes called the problem of bi-linear equations. DS and CA are both SVD methods as we will see later. Having described the SVD, we now show how the optimal weights are obtained using SVD. The weights of the DS solution are obtained by applying SVD to $\mathbf{S}$, which is the weighted Chi-squared ($\chi^{\mathbf{2}}$) distance matrix. $\mathbf{S}$ is defined as follows:

$$\mathbf{S} = \mathbf{D_r}^{-\frac{\mathbf{1}}{\mathbf{2}}} \mathbf{F} \mathbf{D_c}^{-\frac{\mathbf{1}}{\mathbf{2}}}. \tag{2.5}$$

In our example in Table 2.4, $\mathbf{S}$ becomes,

TABLE 2.4: S matrix, indented with column and row names

	Good	**Average**	**Poor**
White	0.100000	0.286039	0.670820
Brown	0.300000	0.476731	0.223607
Green	0.632456	0.301511	0.000000

Once the data is pre-processed, SVD can be applied directly to $\mathbf{S}$, as

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}. \tag{2.6}$$

Here $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular matrices of the SVD matrix and $\mathbf{\Sigma}$ is the diagonal matrix of singular values. Although there exist other methods [2] of obtaining the optimal weights, SVD is a more flexible way of obtaining optimal weights and they all come to similar weights. It is worth noting that the SVD usually produces the trivial solution, $\mathbf{1}$, which is removed before displaying the results, $\mathbf{1}$ is the result of the singular value 1 as we will see later on that this is the same for CA. Then the optimal weights associated with the eigenvalues obtained in the SVD are obtained in this way:

$$\mathbf{y} = \mathbf{f_t^{\frac{1}{2}}}\mathbf{D_r^{-\frac{1}{2}}}\mathbf{U} \tag{2.7}$$

and

$$\mathbf{x} = \mathbf{f_t^{\frac{1}{2}}}\mathbf{D_c^{-\frac{1}{2}}}\mathbf{V}, \tag{2.8}$$

respectively.

Here $\mathbf{x}$ and $\mathbf{y}$ are column and row optimal weights for $\mathbf{F}$, respectively. These optimal weights maximize the product-moment correlation, $\eta^{\mathbf{2}}$. While these optimal weights are often called normed weights, the projected weights may be obtained by pre-multiplying equation (2.7) and equation (2.8) by $\mathbf{\Sigma}$, that is,

$$\mathbf{projected_y} = \mathbf{f_t^{\frac{1}{2}}}\mathbf{D_r^{-\frac{1}{2}}}\mathbf{U}\mathbf{\Sigma} \tag{2.9}$$

and

$$\mathbf{projected_x} = \mathbf{f_t^{\frac{1}{2}}}\mathbf{D_c^{-\frac{1}{2}}}\mathbf{V}\mathbf{\Sigma}. \tag{2.10}$$

If we take $\mathbf{\Sigma}^T$ and multiply it by $\mathbf{\Sigma}$, then we get the corresponding eigenvalues (η^2). There are $K = min(m,n)\text{-}1$ total number of components that can be extracted by SVD. The sum of eigenvalues (correlation ratios) is defined as the total information in the contingency table and is thus distributed over K nontrivial components excluding the trivial component which is 1 (Nishisato, 2006). Therefore total information, $\mathbf{v_t}$ is obtained as follows:

$$\mathbf{v_t} = \sum_{\mathbf{k=1}}^{\mathbf{K}} \eta_{\mathbf{k}}^{\mathbf{2}} = \mathbf{tr}(\mathbf{S^T S}) = \frac{\chi^{\mathbf{2}}}{\mathbf{f_t}} - \mathbf{1}, \tag{2.11}$$

[2] Some methods that yield the same optimal weights include the power method which is a numerical method that extracts dominant eigenvalues iteratively, the eigendecomposition, the method of reciprocal averages (MRA). These methods are discussed in (Nishisato, 2014).

where χ^2 is a chi-squared statistic corresponding to the table being studied. Since the total information is defined, we need to know how the information is distributed over the K components and this information can be captured by a δ statistic,

$$\delta_j = \frac{100\eta_j^2}{\sum_{k=1}^{K} \eta_k^2}. \tag{2.12}$$

The columns and row optimal weights are presented in Table 2.5 and Table 2.6, respectively. Where the column weights are as follows:

TABLE 2.5: Column weights (x) for teachers data

	DS axis1	DS axis2
Good	1.076020	-0.861499
Average	0.092163	1.275880
Poor	-1.471749	-0.677461

and the row weights are,

TABLE 2.6: Row weights (y) for teachers data

	DS axis1	DS axis2
White	-1.232015	-0.618174
Brown	0.122802	1.372924
Green	1.232458	-0.838611

and the corresponding projected weights are as follows:

TABLE 2.7: Projected column weights ($proj_x$)

	DS axis1	DS axis2
Good	0.653114	-0.153107
Average	0.055940	0.226751
Poor	-0.893311	-0.120399

TABLE 2.8: Projected row weights ($proj_y$)

	DS axis1	DS axis2
White	-0.747799	-0.109863
Brown	0.074538	0.243998
Green	0.748068	-0.149039

As we can see in Table 2.7 and Table 2.8, the projected weights (principal coordinates in CA framework) are obtained by multiplying η_k, the singular value corresponding to the k component by the normed weight, x_j for columns and y_i for rows, that is, $\eta_k \times x_j$ or $\eta_k \times y_i$ for columns and rows respectively. It is known that even these projected weights are orthogonal to each other in Euclidean space (Nishisato, 1996). Furthermore the weights are scaled in such a way that their average equals zero, i.e. $\bar{x}$ and $\bar{y}$ are zero (van de Velden and Neudecker, 2000; Nishisato, 2014).

Finally Table 2.9 presents a summary of DS results, including the eigenvalues, the variance accounted for and the cumulative percentage of variance accounted for by the 2 eigenvalues. The results shown in Table 2.9 show that there are two possible solutions of our teachers' data in 2.1, the first eigenvalue, 0.032 accounts for 92 % of the total variances while the second eigenvalue, 7.896 accounts for the remaining 8 % of the total variance.

TABLE 2.9: Summary of DS solution (teachers)

	DS Axis 1	DS Axis 2
Eigen_value	0.368415	0.031585
Delta	92.103768	7.896232
cum Delta	92.103768	100.000000

2.4.2 DS for multiple-choice data

Contingency tables take at most two categorical variables, and when we have more than two categorical variables, then this is called multiple-choice data. Consider **F** as an N by n multiple-choice dataset, where N and n indicate the number of respondents and number of questions, respectively. Since this table contains only categorical variables, it can be further transformed into an indicator or binary matrix. The columns of this new table represent the categories of each question. In our example, if the response of a respondent is (3, 1, 2, 1) then this can be expressed as follows, [(001), (100), (010), (100)]. DS for multiple-choice data has the same formulation with the DS for contingency tables, that is, the same DS algorithm in Section 2.4.1 is used for the multiple-choice

data, the only difference is that there is a change of inputs in the SVD algorithm, that is, the matrix of interest is the indicator matrix instead of the direct data table. It is worth noting that the indicator matrix can also be seen as one-hot encoding in statistics and machine learning.

Dual scaling for multiple-choice data spans a larger space compared to the DS for a contingency table, that is, the data to be decomposed by SVD has many dimensions, where the columns represent the total number of categories in the full dataset. DS for multiple-choice data analyses $\mathbf{X}$, an indicator matrix of $\mathbf{F}$ in Table 2.10. Consider equation (2.5), $\mathbf{X}$ replaces $\mathbf{F}$ and the standard dual scaling algorithm may be applied to our new $\mathbf{S}$. There exist some restrictions on the analysis of multiple-choice data since we are dealing with the expanded space. Firstly, we examine the number of components that can be extracted by DS on the multiple-choice data. In the case of an indicator matrix, we have N respondents and m options (total number of categories), but because the trivial solution is removed from the solutions, then the total number of components, K, is given as follows:

$$K = m\text{-}n, \tag{2.13}$$

where m is the number of categories and n is the total number of questions. In the example in Table 2.10, we can have at most 12-4 = 8 non-trivial solutions.

The statistics discussed in Section (2.4.1) are all obtained in a similar way, except that the statistic for the total variance of δ is affected by the total number of solutions. We now have expanded space and the resulting eigenvalues are very small (Nishisato, 2006). To illustrate this we will show how the total variance is calculated in equation (2.14),

$$\delta_{\mathbf{k}} = \frac{\mathbf{100}\eta_{\mathbf{k}}^{\mathbf{2}}}{\bar{\mathbf{m}} - \mathbf{1}}. \tag{2.14}$$

where $\bar{m}$ is the average number of options (categories). In Table 2.10 $\bar{m} = 3$ since we have 3 options per questions.

Thus the first solution corresponding to the largest eigenvalue cannot explain more than 50% of the total variance (Gaul and Pfeifer, 2013; Nishisato, 2014). The denominator of equation (2.14) is called the total variance (total information), $\mathbf{v_t}$, that is,

$$\mathbf{v_t} = \bar{\mathbf{m}} - \mathbf{1}. \tag{2.15}$$

In multiple-choice data, this statistic $\mathbf{v_t}$ is defined as the sum of all correlation ratios of non-trivial solutions, and thus is defined,

$$\mathbf{v_t} = \eta_{\mathbf{1}}^{\mathbf{2}} + \eta_{\mathbf{2}}^{\mathbf{2}} + ... + \eta_{\mathbf{m-n}}^{\mathbf{2}}. \tag{2.16}$$

As it was with the contingency table, the objective of DS in multiple-choice data is to determine the row and column weights that maximize the correlation ratio, $\eta^{\mathbf{2}}$. Individuals with similar answers to the questions are expected to be close to each other in a numerical space, that is, we assign similar scores to individuals who chose the same options for a question (Guttman, 1941). Consider multiple-choice data with 4 questions, each with 3 options to choose from, that is to say, 4 categorical variables and 3 categories in each. The dataset in Table 2.10 is used with great thanks to Nishisato (2014). The dataset in Table 2.10 is a multiple-choice data with 4 questions, each with 3 options and 23 respondents. The first column in Table 2.10 contains the questions and their corresponding options. Dual scaling was applied to this data and the summary of results are reported in Table 2.11.

TABLE 2.10: Singapore dataset, multiple-choice data extracted from Nishisato (2014). This data was collected from a Dual Scaling conference and it consists of 23 responses and 4 categorical variables, these variables were 4 close-ended questions and are shown in the index column. In this data, all the questions comprise of exactly 3 options to choose from. For instance, item 1, which is the first question was concerned with the age of the respondents and there were 3 options to choose from, namely: (1) 20-29, (2) 30-39 and (3) 40 or older. When processing the data for Dual Scaling the index column is ignored since it was used in this table for illustrative purpose only.

	Data				
		Item			
	Respondent	1	2	3	4
Item 1					
How old are you?					
(1) 20-29	1	3	1	2	1
(2) 30-39	2	2	1	3	2
(3) 40 or older	3	1	1	2	2
Item 2	4	3	2	2	3
Children today are	5	3	1	2	2
not as disciplined as when	6	1	3	1	2
I was a child	7	2	1	2	2
(1) Agree	8	2	1	1	2
(2) Disagree	9	1	2	3	1
(3) I cannot tell	10	3	1	2	1
Item 3	11	1	2	2	3
Children,today are not	12	2	1	1	1
as fortunate as	13	2	1	3	3
when I was a child	14	3	1	2	1
(1) Agree	15	1	1	2	3
(2) Disagree	16	3	1	2	1
(3) I cannot tell	17	3	1	1	1
Item 4	18	2	3	2	2
Religion should be	19	3	1	2	1
taught at school	20	2	1	2	2
(1) Agree	21	1	3	3	3
(2) Disagree	22	2	1	2	2
(3) Indifferent	23	1	3	3	3

To illustrate how DS handles the multiple-choice data, the summary of results for the example in Table 2.10 is reported in Table 2.11, where the eigenvalues were obtained

from the SVD solution in equation (2.5) and δ refers to the variance accounted for by component k (in percentages). The total number of components is 8 as shown before, the first eigenvalue 0.648 accounts for 32.38 % of the total variance. The eigenvalues are arranged in descending order, the first eigenvalue cannot account for more than 50 % of the total variance, δ.

TABLE 2.11: A summary of DS results for Singapore dataset

	DS Axis 1	DS Axis 2	DS Axis 3	DS Axis 4	DS Axis 5	DS Axis 6	DS Axis 7	DS Axis 8
Eigenvalues	0.647561	0.440666	0.317039	0.213612	0.184268	0.115704	0.052759	0.028392
δ	32.378033	22.033307	15.851955	10.680599	9.213385	5.785180	2.637960	1.419583
Cumδ	32.378033	54.411339	70.263294	80.943893	90.157277	95.942457	98.580417	100.000000

2.4.3 DS for rank-order data

Rank-order data is often encountered in marketing, social science and psychological research, where individuals may be asked to rank different items. It may be different cars, different government services or brands. The rank-order data is usually stored in a rectangular matrix with rows being the judgments and columns represent the objects being ranked. It becomes difficult to interpret this data matrix when it is high-dimensional. Thus DS can be used to summarize this data and give a quantitative interpretation of ranks. Most of the attention is given to obtaining weights for the columns, that is weights for items being ranked (Nishisato, 1996). Items with the highest preference are assigned the largest weights, for example in Table 2.1. Dual scaling for rank-order data is flexible and unbiased in extracting patterns, in a sense that the existing patterns are not dependent on subjective perceptions but are extracted by the method itself from the data.

Let $\mathbf{R}$ be an N by n data table of ranks. Dual scaling in this dataset aims to produce a set of weights for the items being ranked. The lower-rank approximation for this data contains several solutions or components where each component accounts for a proportion of variance. The proportion of the total variance is referred to as inertia, more generally it is the eigenvalue corresponding to a component. The optimal weights can be obtained using the SVD. It is common to plot the resulting optimal weights in a lower-dimensional space using biplots and joint display.
In this section, we will show how the optimal weights are obtained step by step. Let $\mathbf{R} = \mathbf{r_{ij}}$ be a rank order data matrix, where $\mathbf{r_{ij}}$ is the rank assigned by judge i on item $\mathbf{X_j}$. Then we use the transformation $\mathbf{e_{ij}}$ to obtain $\mathbf{E}$, where $\mathbf{E} = e_{ij}$ (Nishisato, 2014):

$$e_{ij} = n + 1 - 2 * r_{ij} \tag{2.17}$$

where $i = 1, 2, ..., N$ and $j = 1, 2, ..., n$. $\mathbf{E}$ is called a dominance matrix.
To obtain the optimal weight, vector $\mathbf{x}$ that maximizes the squared correlation, $\eta^{\mathbf{2}}$, we then solve the the following eigen equation:

$$(\mathbf{H}_n - \lambda \mathbf{I})\mathbf{x} = 0 \tag{2.18}$$

where

$$\mathbf{H}_n = \frac{1}{Nn(n-1)^2} * \mathbf{E^T E} \tag{2.19}$$

and the squared correlation ratio can be given as follows:

$$\eta^2 = \frac{1}{Nn(n-1)^2} \frac{\mathbf{x^T E^T E x}}{\mathbf{x^T}} = \frac{\mathbf{x^T H}_n \mathbf{x}}{\mathbf{x^T x}}. \tag{2.20}$$

In a more general way, we can use SVD on

$$\mathbf{S} = \mathbf{D_r}^{-\frac{1}{2}} \mathbf{E} \mathbf{D_c}^{-\frac{1}{2}} \tag{2.21}$$

where

$$\mathbf{D}_r = n(n-1)\mathbf{I}_N \tag{2.22}$$

and

$$\mathbf{D}_c = N(n-1)\mathbf{I}_n \tag{2.23}$$

where $\mathbf{I}_N$ and $\mathbf{I}_n$ are identity matrices corresponding to N respondents and n items being ranked, respectively. This is a minor difference from DS of contingency tables and multiple-choice data since the row totals of the matrix $\mathbf{E}$ are equal to zero (Nishisato, 1978). When an individual judges the item, that particular item is compared with the remaining $n - 1$ items hence each row is based on $n(n - 1)$ responses and each column is based on $N(n - 1)$ responses.

It is generally the case that the resulting eigenvalues are ordered in descending order by SVD, where $\mathbf{S}$ is decomposed into orthogonal factors such that

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{2.24}$$

From this decomposition, the optimal weights can be found in a similar way to that of DS for contingency tables as explained in Section 2.4.1, except that the grand total, $\mathbf{f_t}$, becomes $Nn(n-1)$. The optimal solution can uncover existing patterns in the rank-order dataset. The corresponding components of the SVD solution explain a proportion of variance in the data (Nishisato, 1978).

Each component accounts for a certain amount of variance, called *inertia* in the CA framework, which is an eigenvalue corresponding to that component. Thus the total amount of variance in the data is equivalent to the sum of eigenvalues defined as follows: consider, $\mathbf{H}_n$ in equation (2.19), in order for us to extract the amount of variance accounted for by each component we use the following equation:

$$\delta_k = \frac{\lambda_k}{tr(\mathbf{H}_n)} \tag{2.25}$$

where δ_k is the inertia or percentage of variance accounted for by component k, λ_k is an eigenvalue corresponding to component k and $tr(\mathbf{H}_n) = \sum_k \lambda_k$, that is, $tr(\mathbf{H}_n)$ is the summation of all the diagonal elements of $\mathbf{H}_n$.

2.4.3.1 Example of DS for rank-order data

Consider a rank-order dataset of 10 different government services to be judged by 35 people. The aim of DS is to find the patterns in these 10 services concerning the 35 people who judged them. We use Table 2.12 to get the dominance matrix, $\mathbf{E}$ and $\mathbf{H}_n$. The total number of possible components, k, is equal to 9. Applying SVD to $\mathbf{H}_n$ yields the optimal weights and the summary of DS. Table 2.13 shows a summary of DS results, where, λ, δ and $cum\delta$, is the eigenvalue, percentage of inertia and cumulative percentage of inertia, respectively. Furthermore, Table 2.13 shows that the first two eigenvalues account for 60.4 % of the total variance, and this means that the rows and column weights corresponding to these eigenvalues may be used for describing the rank-order dataset in Table 2.12.

TABLE 2.12: Government services data, rank-order data

Respondent	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1	7	9	10	2	6	3	8	5	4
2	6	10	9	5	3	1	7	2	4	8
3	9	8	4	3	5	6	10	2	1	7
4	2	10	5	6	3	1	4	8	7	9
5	2	10	6	7	4	1	5	3	9	8
6	1	3	5	6	7	8	2	4	10	9
7	7	10	1	6	5	3	8	4	2	9
8	2	10	6	7	4	1	5	3	9	8
9	2	10	5	8	4	1	6	3	7	9
10	2	10	5	9	8	7	4	1	3	6
11	9	10	7	6	5	1	4	2	3	8
12	6	10	7	4	2	1	3	9	8	5
13	1	10	3	9	6	4	5	2	7	8
14	8	6	5	3	10	7	9	2	1	4
15	8	10	9	6	4	1	3	2	5	7
16	3	5	10	4	6	9	8	2	1	7
17	1	10	8	9	3	5	2	6	7	4
18	5	4	9	3	10	8	7	2	1	6
19	2	10	6	7	8	1	5	4	3	9
20	1	4	2	10	9	7	6	3	5	8
21	2	10	5	7	3	1	4	6	8	9
22	6	3	9	4	10	8	7	2	1	5
23	6	9	10	4	8	7	5	2	1	3
24	5	2	1	9	10	4	8	6	3	7
25	2	10	6	7	9	1	3	4	5	8
26	7	10	9	5	2	6	3	1	4	8
27	7	7	10	3	5	9	4	2	1	6
28	8	8	6	7	5	10	9	2	6	1
29	3	10	7	9	4	1	5	3	6	8
30	2	10	9	1	4	7	5	3	6	8
31	2	10	9	7	5	1	3	2	6	8

TABLE 2.13: DS for the government services data by Nishisato

	DS Axis 1	DS Axis 2	DS Axis 3	DS Axis 4	DS Axis 5	DS Axis 6	DS Axis 7	DS Axis 8	DS Axis 9
λ	0.156	0.090	0.0534	0.0423	0.020	0.0170	0.0112	0.009	0.008
δ	38.277	22.123	13.114	10.390	4.908	4.169	2.746	2.213	1.964
$cum\delta$	38.277	60.400	73.513	83.904	88.812	92.981	95.727	97.940	99.904

Table 2.13 shows that the first eigenvalue, 0.156, accounts for 38.28 % of the total variance. Furthermore, the first two components account for 60.4 % of the total variance which suggests that these two components can be used to understand the structure of the 10 ten government services.

2.5 Correspondence Analysis

Correspondence Analysis (CA) is another popular quantification technique which aims to decompose the chi-squared statistic (χ^2) into orthogonal components. One of the interesting aspects to study in CA are the row profiles, that is, $\mathbf{P}_{ij}/\mathbf{P}_{.j}$ and the column profiles, $\mathbf{P}_{ij}/\mathbf{P}_{i.}$, where, $\mathbf{P}_{ij}$ is a non-negative entry of a contingency table, $\mathbf{P}_{.j}$ is a column total and $\mathbf{P}_{i.}$. The aim of CA is to investigate how these profiles relate: they need to be analyzed in a lower-dimensional space, that is, on a two-dimensional space rather than the original nominal space. In our example in Table 2.1, the row profile for White is 1/10, 3/10, 6/10, for Brown is 3/10, 5/10, 2/10 and for Green is 6/9, 3/9, 0. In short, the profiles can be seen as conditional proportions. In CA the total variance is often referred to as *inertia* and this statistic (inertia) is proportional to the independence χ^2 statistic.

In this section, we will discuss in-depth how CA quantifies the categorical data and reduces the high-dimensional data into a lower rank approximation.

2.5.1 CA for contingency tables

Let $\mathbf{F} = (\mathbf{f_{ij}})$ be the m by n data matrix (non-negative matrix data), whose element $\mathbf{f_{ij}}$ represents the number of observations that belong to the ith level of first categorical variables and the the jth level of the second categorical variable. Consider Table 2.14, in this table the row and column totals are highlighted in green and yellow, respectively. Furthermore, the f_{ij} entries correspond to the remaining data entries, these values are the frequencies corresponding to the combination of the first and second question. In this example, $\mathbf{f_{11}} = 1$, that is, there is one student who rated Mr Green good. This is the same table as in Table 2.1. Furthermore, let $\mathbf{f_t}$ be the grand total of $\mathbf{F}$ that is, the sum of all the frequencies in a contingency table, thus for this example, the total number of students, $\mathbf{f_t}$ who rated teachers is 29.

TABLE 2.14: A contingency table annotated with totals

	Good	Average	Poor	Total
White	1	3	6	10
Brown	3	5	2	10
Green	6	3	0	9
Total	10	11	8	29

The first step in correspondence analysis is to transform the data into probabilities by dividing every element by the total, $\mathbf{f_t}$, so that the elements sum up to 1. This matrix is called the correspondence matrix, given by:

$$\mathbf{P} = \mathbf{F}/\mathbf{f_t}. \tag{2.26}$$

Let $\mathbf{r}$ and $\mathbf{c}$, respectively, be row and column marginal totals as described in Section 2.4.1. Similar to Table 2.1, let $\mathbf{D_r}$ and $\mathbf{D_c}$ be the diagonal matrix of $\mathbf{r}$ and $\mathbf{c}$, respectively.

Once $\mathbf{D_r}$ and $\mathbf{D_c}$ are defined we can compute standardized residuals which is just a weighted χ^2-distance matrix, $\mathbf{S}$,

$$\mathbf{S} = {\mathbf{D_r}}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathbf{T}}){\mathbf{D_c}}^{-1/2}. \tag{2.27}$$

Decompose $\mathbf{S}$ using SVD:

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad \text{where} \quad \mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{2.28}$$

The standardized coordinates are also called the normed weights in the DS literature. Equation (2.29) and equation (2.30) are used to compute the weights for the rows and columns, respectively:

$$\mathbf{Y} = {\mathbf{D_r}}^{-1/2}\mathbf{U} \tag{2.29}$$

and

$$\mathbf{X} = {\mathbf{D_c}}^{-1/2}\mathbf{V}, \tag{2.30}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are row and column standardized coordinates, respectively. Once the standardized coordinates are obtained we can now compute the principal coordinates

which are more insightful and are computed directly from the singular values. The principal coordinates are preferred over standardized weights simply because they contain information about the original dataset and they can also determine the structure of multidimensional data (Nishisato, 2006). These coordinates are also called projected weights in other quantification literature (Nishisato, 2014; Greenacre, 1984). To compute the principal coordinates we use the following equations:

$$\mathbf{F} = \mathbf{D_r}^{-1/2}\mathbf{U\Sigma}, \tag{2.31}$$

and

$$\mathbf{G} = \mathbf{D_c}^{-1/2}\mathbf{V\Sigma}, \tag{2.32}$$

where $\mathbf{F}$ and $\mathbf{G}$ are the rows and columns principal coordinates, respectively. The total inertia of a contingency table is closely related to the χ^2 distances: it is a summation of all the distances and can be computed as follows,

$$inertia = tr(\mathbf{S^T S}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j}. \tag{2.33}$$

2.5.2 CA for multiple-choice data

Multiple Correspondence Analysis (MCA) is an extension of simple correspondence analysis to handle more than two categorical variables. The manner in which MCA handles categorical variables is through taking a super-indicator matrix, $\mathbf{X}$ as input and applying the CA algorithm. $\mathbf{X}$ can be expressed as $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_p}]$, where the sub-matrices are expressed as follows:

$$X_p = \begin{cases} 1 & \text{if the category is chosen} \\ 0 & \text{otherwise} \end{cases} \tag{2.34}$$

, where p is the number of categorical variables

MCA pursues the quantification of categorical variables through the CA algorithm. In order to show the direct relationship between CA and MCA, consider the standardized residuals matrix, $\mathbf{S}$ in equation (2.27), $\mathbf{P}$ is computed from $\mathbf{X}$, the indicator matrix, instead of a non-negative contingency table, $\mathbf{F}$. It is worth noting that there is an alternative of computing MCA from the what is called the *Burt matrix*, $\mathbf{Z}$, where $\mathbf{Z} = \mathbf{X^T X}$.

The Burt matrix is just a concatenation of two-way cross-tabulations between the pairs of variables (Nenadic and Greenacre, 2007). There are many approaches to MCA, including:

1. *Indicator*: compute MCA through decomposing the indicator matrix by simple correspondence analysis.
2. *Burt*: Decompose the *Burt* matrix by some eigen-decomposition method (Greenacre, 2007).
3. *Adjusted*: This approach was proposed by Greenacre (1984) and Nenadic and Greenacre (2007), the inertias are adjusted by re-scaling the coordinates of the solution to best fit the main diagonal of the *Burt* matrix (Nenadic and Greenacre, 2007).

The approach that is related to the DS is the one that decomposes the indicator matrix. This is a more general way of handling multiple-choice data as it is also used in DS. Thus the subsequent statistics that are used in the CA literature for contingency table are also used in the MCA, except that there are some minor differences in terms of the marginal totals, that is, row and column totals, respectively.

2.5.3 CA for rank-order data

Correspondence analysis can also handle rank-ordered data. In this section, we describe how this is done. We first discuss one approach that leads to the same solution as that of DS. This approach decomposes a doubled matrix (van de Velden and Neudecker, 2000) instead of $\mathbf{E}$ or $\mathbf{H_n}$ as seen in Section 2.4.3. The dominance matrix $\mathbf{E}$ which is $(N \times n)$ can also be computed as follows: $\mathbf{E} = \mathbf{L} - \mathbf{M}$, where $\mathbf{L}$ is the matrix of preference and $\mathbf{M}$ is the matrix of dis-preference counts. Therefore dominance matrix$\mathbf{E}$ can also be seen as the difference between the preference and dis-preference on a rank-order matrix $\mathbf{R_{ij}}$.

Furthermore, the matrix of preferences can be defined as $\mathbf{L} = (L_{ij})$ and $L_{ij} = R_{ij} - 1$, where $\mathbf{R} = (R_{ij})$ is an $N \times n$ matrix of ranks. Lastly, the matrix of dis-preferences $\mathbf{M}$ is defined as $M_{ij} = n - 1 - L_{ij}$. CA analyses the doubled matrix where $\mathbf{L}$ is stacked on top of $\mathbf{M}$ (Torres and Greenacre, 2002).

Consider Table 2.15 as an example of rank-order data, in this example, there are 6 items ranked by 10 respondents, where 1 represent the most preferred and 6 represent the least preferred. The original data matrix has 10 rows but the transformed matrix in Table

2.16 has $2N$ rows, where the first N rows represent dis-preference, for an example on column 1, row 1 in Table 2.15, A is ranked 6th by the first respondent, therefore on the first row, the first column of first 10 rows, A is therefore dis-preferred compared to 5 other. The doubled matrix gives such quantifiable differences between the ranks.

TABLE 2.15: Rank-order data extracted from Torres and Greenacre (2002)

Respondent	A	B	C	D	E	F
1	6	1	5	4	3	2
2	2	6	3	5	4	1
3	6	1	5	2	4	3
4	3	5	2	4	1	6
5	3	4	2	6	1	5
6	5	3	1	4	6	2
7	1	2	4	5	3	6
8	4	3	2	6	5	1
9	2	1	4	5	3	6
10	6	1	4	3	5	2

TABLE 2.16: Rank-order data in a doubled format (Torres and Greenacre, 2002)

	A	B	C	D	E	F
-1	5	0	4	3	2	1
-2	1	5	2	4	3	0
-3	5	0	4	1	3	2
-4	2	4	1	3	0	5
-5	2	3	1	5	0	4
-6	4	2	0	3	5	1
-7	0	1	3	4	2	5
-8	3	2	1	5	4	0
-9	1	0	3	4	2	5
-10	5	0	3	2	4	1
1	0	5	1	2	3	4
2	4	0	3	1	2	5
3	0	5	1	4	2	3
4	3	1	4	2	5	0
5	3	2	4	0	5	1
6	1	3	5	2	0	4
7	5	4	2	1	3	0
8	2	3	4	0	1	5
9	4	5	2	1	3	0
10	0	5	2	3	1	4

The correspondence matrix $\mathbf{P}$ is defined as

$$\mathbf{P} = \begin{bmatrix} \mathbf{L} \\ \mathbf{M} \end{bmatrix} / Nn(n-1)$$

The row and column masses of $\mathbf{P}$ for the doubled matrix are defined as follows:

$$\mathbf{r} = \frac{1}{2N}\mathbf{1} \quad \text{and} \quad \mathbf{c} = \frac{1}{n}\mathbf{1} \tag{2.35}$$

where $\mathbf{1}$ is a matrix of ones that corresponds either to the rows or columns. The diagonal matrices are defined as follows:

$$\mathbf{D_r} = \frac{1}{2N}\mathbf{I} \quad \text{and} \quad \mathbf{D_c} = \frac{1}{n}\mathbf{I}, \tag{2.36}$$

where $\mathbf{I}$ is an identity matrix that corresponds to either the rows or columns. The CA is a result of SVD of the residual matrix $\mathbf{S}$, where,

$$\mathbf{S} = \mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc^T})\mathbf{D_c^{-1/2}} \tag{2.37}$$

Substituting $\mathbf{D_r}, \mathbf{D_c}, \mathbf{r}, \mathbf{c}$ and $\mathbf{P}$ gives the following expression for $\mathbf{S}$:

$$\mathbf{S} = (2N)^{1/2}\left(\begin{bmatrix} \mathbf{L} \\ \mathbf{M} \end{bmatrix} / Nn(n-1) - \tfrac{1}{2Nn}\mathbf{1}\mathbf{1^T}\right) n^{1/2}. \tag{2.38}$$

Simplifying equation (2.38), yields:

$$S = \frac{1}{2^{1/2}(Nn)^{1/2}(n-1)} \begin{bmatrix} \mathbf{E} \\ -\mathbf{E} \end{bmatrix}, \tag{2.39}$$

Thus,

$$S = 2^{-1/2} c \begin{bmatrix} \mathbf{E} \\ -\mathbf{E} \end{bmatrix}. \tag{2.40}$$

where $\mathbf{E}$ is the dominance matrix as defined in Section (2.4.3) and $c = \sqrt{\frac{1}{Nn(n-1)^2}}$ is the same constant discussed in Section (2.4.3). Thus the dominance matrix is on top of the doubled matrix while its negative is located at the bottom of the same matrix. Then the SVD of $\mathbf{S}$ can be written as follows:

$$\mathbf{S} = \begin{bmatrix} \mathbf{U} \\ -\mathbf{U} \end{bmatrix} \mathbf{\Sigma} \mathbf{V}^T, \tag{2.41}$$

where $\mathbf{U^T U} + \mathbf{U^T} = \mathbf{V^T V} = \mathbf{I}$ and likewise, $\mathbf{\Sigma}$ contains the singular values in a descending order. Thus the corresponding optimal weights are computed as follows,

$$\mathbf{y} = \mathbf{D}_r^{-1/2}\mathbf{U} = (2N)^{1/2}\mathbf{U}. \tag{2.42}$$

This solution is associated with first N rows. To get the weights associated with the last N rows we will negate $\mathbf{y}$ and use $-\mathbf{y}$. The optimal weights for the columns are then obtained as follows:

$$\mathbf{x} = \mathbf{D}_c^{-1/2}\mathbf{V} = n^{1/2}\mathbf{V}. \tag{2.43}$$

2.6 Relationship between DS and CA

DS and CA are conceptually related when it comes to analyzing contingency table (Nishisato and Sheu, 1980; van de Velden and Neudecker, 2000). Several studies have shown that these methods lead to a similar solution, for example, van de Velden and Neudecker (2000) studied the eigenvalue property of CA, DS, homogeneity analysis and other related methods. Their study shows that mathematically all these methods are equivalent in terms of the eigenvalue decomposition: the singular values always lie between 0 and 1 (van de Velden and Neudecker, 2000). For instance, suppose $\mathbf{F}$ is an $n \times p$ non-negative matrix. Furthermore suppose $\mathbf{F}$ is a contingency table, where rows and columns are categories of two variables and the corresponding cells contain the frequency of occurrences or counts. For both methods (CA and DS) we have the row and column vectors, $\mathbf{r}$ and $\mathbf{c}$, respectively. The corresponding row and column profiles are obtained in the following way:

$$\mathbf{R} = \mathbf{D_r^{-1}F} \tag{2.44}$$

and

$$\mathbf{C} = \mathbf{D_c^{-1}F}, \tag{2.45}$$

where $\mathbf{R}$ and $\mathbf{C}$ are matrices that contain row and column profiles, respectively. Then their corresponding matrix for singular values is given as follows:

$$\mathbf{Q} = \mathbf{D_r^{-1/2}FD_c^{-1/2}}. \tag{2.46}$$

Where **F** is a contingency table. Alternatively, we can use **X**, the super-indicator matrix for multiple-choice data. Finally, CA and DS are both helpful exploratory methods and these two methods give similar results when analyzing the indicator matrix of multiple-choice data. We also noticed from section 2.4.3 and section 2.5.3 that DS has a very simple approach to rank order data than CA, therefore this study does not only focus on quantification but to DS, in particular since it handles a variety data sets simply.

2.7 Resampling Techniques

It is common practice in the case of multivariate studies to use resampling methods for assessing the stability of solutions. Several samples may be drawn from a sample with replacement. Techniques of interest may be applied across the resamples, such as those of quantification (MCA and DS). Resampling does not only help in the reproducibility of results but it can also be used to investigate instabilities caused by changing some parts of the data. It is usually expected that small unimportant changes in the data should lead to small unimportant changes in the results (Gifi, 1990). Moreover, real datasets may violate the assumptions of multivariate normality, making it difficult for parametric methods to be applied to the data for analysis, and even the underlying distribution of the dataset may not be directly known. Hence resampling methods play a major role in estimating those distributions.

2.7.1 Bootstrap

Bootstrap is a resampling technique that was developed by Efron (1992). It is often referred to as a non-parametric technique for inference. The general idea of Bootstrap is to take a sample and treat it as a virtual population and draw samples with replacement and estimate confidence intervals. Bootstrap can directly be utilized for inference by generalizing the parameter being studied (Mooney, 1996), the parameter involved in the study of quantification include eigenvalues, normed and projected coordinates. Bootstrap tends to strengthen the results of a study by replicating the sample and randomly sampling with replacement. This is done through the investigation of possible data arrangement of the same sample being studied, that is, given the sample at hand, and assuming that the sample is coming from an unknown full population dataset, Bootstrap aims to sample from the given sample possible samples of the data. Bootstrapping aims to estimate a certain estimate over multiple experiments, that is, measure how is this estimate changing on different samples called resamples. This estimate $\hat{\theta}$, can be the mean, $\bar{x}$, of the eigenvalues. The eigenvalues are computed for each run, thereafter the

mean of each set of eigenvalues is taken, finally, the average mean for different resamples is computed, this is referred to as Bootstrap mean. The Bootstrap means can be compared with the empirical that is computed from the original sample.

Some strengths of Bootstrap include that it is generally very simple to implement. The whole idea around Bootstrap is that the sample at hand is believed to be coming from the original population and therefore this sample represents the population. It is important to note that Bootstrap should not be used as a substitution for gathering more data. Bootstrap can be seen as a way of finding the distribution of a sample being studied. The distribution, in this case, is anything from spread and dispersion of the eigenvalues, that is, Confidence Intervals may be constructed around the statistics of interest, and this is normally practised in assessing the stability of quantification results.

Bootstrap can be used to construct confidence regions around the statistics of interest. One of the recent developments in Bootstrapping is the use of confidence intervals for various parameters such as the normal approximation method and percentiles (Fan and Wang, 1996).

Bootstrap has a flexible strategy in sampling, in that it can be used for any statistic. Moreover, Bootstrap samples are of equal size to the original sample being studied. The procedure of Bootstrap can be summarized as follows:

- Step 1: Any number of new samples can be drawn from the sample being studied by sampling with replacement, that is, after drawing randomly from the original sample the sampled observations may be selected again. Therefore, some observations may be selected more than once and some may not occur in the Bootstrap sample.
- Step 2: The second step involves calculating the Bootstrap distribution, which is a particular statistic of interest, $\hat{\theta}$ (mean, median or variance) for each Bootstrap sample.

  i.e., In this step for this study, quantification is obtained for each resample and the statistic of interest, $\hat{\theta}$, is computed for the parameter of interest, eigenvalue in particular.
- Step 3: Finally, the Bootstrap distribution can be used for giving insight about the original sample.

  i.e, The confidence intervals may be computed for our result (distribution) obtained in step 2.

Various confidence intervals can be computed for the statistic of interest using Bootstrap. Bootstrap can be used to estimate the confidence interval of the complicated parameters without any complex mathematical effort. Confidence interval estimation can be done using the normal approximation or the percentile method. The normal approximation method is used to transform the parameters into a normal distribution and standard normal distribution values may be used. For example, $\hat{\theta}$ can be any parameter of interest and the corresponding confidence interval can be defined as $(\hat{\theta} \pm Z_{\alpha} SE_{boot})$, whereas the percentile method trims the data into two extremes [α / 2 , 1- α /2], which is (0.25, 0.975) for 95 % CI, α is the significance level, that is, the probability of rejecting the null hypothesis when it is true (Hesterberg et al., 2005; Engel et al., 2012; Haukoos and Lewis, 2005). More information on how confidence intervals work is discussed in (Hazra, 2017).

The following steps are necessary for constructing confidence intervals for a statistic of interest:

1. Step 1: Draw samples with replacement from the underlying dataset B times.
2. Step 2: Compute the statistic of interest, $\hat{\theta}$ (mean, variance etc). Estimate the distribution of Bootstrap samples.
3. Step 3: Resample and re-estimate the Bootstrap distribution of the statistic B times. This is important since we assume that the Bootstrap distribution is closer to the sampling distribution.
4. Step 4: Compute the confidence intervals of the statistic using the Bootstrap distribution and hence construct CI for the percentiles.

Bootstrap is convenient and it can be applied to any statistic, furthermore, it can be used to estimate most complicated confidence intervals such as that of a median. Hence, Bootstrap is more efficient for non-parametric studies when we can not find parametric methods analytically for parameters. Bootstrap is often recommended because of its flexibility to estimate how the parameters may vary across different samples that are obtained in a similar way.

There are some recommendations on the use of Bootstrap. Markus (1994) recommends that at least 1000 Bootstrap samples must be used to get an insight into the characteristics of the data being studied with at least 200 observations in each sample may be used to assess the variation between the samples (Linting et al., 2007). Rather than directly

looking at the confidence intervals, emphasis may be placed on the interval coverage statistics which may be computed simultaneously with the confidence intervals.

2.7.2 Relationship between Bootstrap and Jackknife

As opposed to Bootstrap, Jackknife is the process of taking a subset from a sample that is taken directly from the underlying sample being studied. Jackknife can be used when the sampling strategy is complex, such as stratified random sampling. This technique estimates the properties of a dataset by repeating the analysis on the same dataset. Every time the analysis is carried out, one observation is deleted, that is, estimate the quantification results by leaving k elements of the data on each trial. This is often termed as Leave-Out method. Bootstrap is likely to give different results when repeated on the same data while Jackknife gives the same results each time, this is because of the random sampling(Hesterberg et al., 2005). Jackknife requires less computational time than Bootstrap does because it does not require estimating the distribution.

Jackknife is often referred to as an approximation of Bootstrap, in that they are all used as the means to estimate the distribution of complex datasets. The difference is that Bootstrap draws samples with replacement, whereas Jackknife draws samples without replacement. It is said that both these methods can be used for inferential purposes. Both Jackknife and Bootstrap can be used to estimate the unknown distribution of the dataset. Jackknife can be seen as a special case of Bootstrap in that all the estimations are based on the subset of the original dataset, hence Jackknife tends to be faster in smaller datasets compared to Bootstrap. Jackknife is sometimes not preferred as Bootstrap takes full advantage of the dataset by using all the information.

Although Bootstrap and Jackknife generally have common goals of estimating the sampling distribution, they usually achieve this using different sampling techniques (Fan and Wang, 1996). Jackknife usually has a finite number of resamples since the subsequent resamples are dependent on the sample size of the data being studied. Jackknife has $n-1$ samples, i.e., if n is the number of observations in a sample, then the maximum number of samples that can be attained is $n-1$, with each unequal sample sizes across the resamples. On the other hand, Bootstrap has infinite samples with equal sizes.

2.7.3 Permutation Tests Resampling

A permutation test is another type of resampling which does resampling by shuffling all possible arrangements of the dataset.

Permutation test provides an alternative resampling by shuffling the data at hand and comparing the difference between particular statistics such as the mean and standard deviations. A permutation test is very useful and important in deriving the p-values for quantifying the difference between the original sample and the resampled sample. These p-values are computed from the empirical distribution, i.e, from the observed shuffled samples.

A permutation test is the best alternative for hypothesis testing procedures such as t-tests and are more accurate than the results from traditional statistical tests. Thus, a permutation test can also be used to test the stability of quantification results by reshuffling and mixing the rows and columns of multiple-choice data. Specifically, the quantification results such as eigenvalues are computed for the original sample, then the samples are shuffled and the distribution is computed for the difference between the original sampled and the derived samples.

The following steps are necessary for the process of permutation tests:

1. Consider a categorical dataset with, rows and columns. The quantification is applied in this sample, that is, the eigenvalues, row and column weights are obtained for this sample.

2. The data table is shuffled **P** times without replacement, **P** is the number of replications or number of resamples, for an example, 200, 500 or 1000 times.

   As the dataset is shuffled, the difference between the original and resampled eigenvalues is computed, this can be quantified as the difference between the mean of eigenvalues for the original sample and resamples.

3. The empirical distribution of the differences between the mean of eigenvalues is computed from the resamples.

4. Finally the p-values are computed based on the success of the resamples in estimating the desired statistic and the point estimate such as the mean.
   Given the means of the eigenvalues for the resamples, in this step the means closer to the mean of the eigenvalues in the original sample are added and divided by the number of replications, say 200. More specifically, if 100 of 200 resamples shows that the mean is close to the original, this implies that p-value is 100/200, i.e, 0.5. the higher the p-value the closer the resampled data is to the original data. Thus the results are said to be stable.

2.8 Related Research

Quantification has a long history and among the contributions, we can point out the classification of categorical variables. One of the earliest techniques that were used to classify categorical data is correspondence analysis which was proposed by (Fisher, 1940), however, this method could only be used for classifying two categorical variables at most.

This technique is used as an exploratory method that can be used to explore frequency-based associations between two categorical variables. The frequencies of the categories are transformed to Euclidean distances using the SVD decomposition, thus CA can be successfully be used to visualize these associations. The relationships between these variables can be plotted on a lower-dimensional space, hence this method can also be described as a dimensionality reduction for categorical datasets (Nishisato, 1993; Costantini et al., 2010). The row coordinates or column coordinates can be plotted separately on different axes, this is also known as one-mode plots (Nishisato and Clavel, 2009). These coordinates can also be simultaneously plotted in one axis, this type of graph is referred to as bi-plot (Bakk and le Roux, 2017). In marketing bi-plots may also be seen as perceptual maps, that is, a graphical display used to explore the perceptions about different brands, where the proximity between brands and perceptions can be compared to gain information on how different brands are positioned against each other and against the perceptions.

The extension of CA to more than two categorical variables is called MCA, as described in Section 2.5.2, which is used to analyze multiple-choice data. Another method equivalent to CA is Dual Scaling (DS) which can handle both contingency tables and multiple-choice data (Nishisato, 2014). These methods are conceptually similar for contingency tables (Hayashi et al., 2013) and multiple-choice data (Ayele et al., 2014). DS applies to a wide range of data-types, including, rank-order, sorting data (Nishisato, 2014).

The quantification methods have been classified as descriptive methods and have been used only for Exploratory Data Analysis (EDA). It is difficult to use the quantification results for inferential purpose because of the lack of knowledge as to whether the results produced by these methods are stable (or occur by chance) (Michailidis and de Leeuw, 1998; Greenacre, 2006; Greenacre and Blasius, 2006).

Stability of results can be defined in various ways: one can consider the replication, stability under selection, statistical, numerical, analytic and algebraic stability. The

aforementioned types of stability address different aspects of stability and have been a central concept in Gifi system [3].

Unstable interpretation of results do not only disprove the results but the method used. Gifi (1990) defines stability by stating that "small unimportant changes in the data should result in small unimportant changes in the results" (Gifi, 1990), which is probably the most used definition in the quantification literature. Stability according to Greenacre (2007) is how the interpretation of CA is affected by mixing some rows and columns (Greenacre, 2007). In other words, we wish to know whether the patterns generated graphically by CA will change dramatically and hence change the interpretation.

As a way to investigate stability researchers resort to resampling methods such as Bootstrapping, Jackknife and Permutation tests. These methods can be used to ascertain the results given by the estimates of the quantification methods, that is, these methods can be used to assess how estimates such as eigenvalues changes due to changes in the data. The change in this estimate can be captured by both variance and bias.

Several stability studies have been done on multiple-choice and contingency tables and very little attention has been given to sorting and rank-order data. Takane and Hwang (2006) proposed a regularized MCA method that would be used as an inferential method. This method uses Bootstrap and Jackknife to obtain the optimal regularizing parameter, however, very little has been done to investigate under what sampling conditions and under what properties of the data quantification methods provide stable results. Furthermore, the stability analysis has been done mostly in the context of visualization and there is little attention given to the actual properties of the data and sampling conditions.

Lebart (2006) suggested Bootstrap methods for validating the quality of quantification results in the unsupervised context of MCA. In this study Lebart (2006) focused on the aspect of the interpretation of principal axes, that is, what happens to the lower-dimensional plots as the data changes by mixing rows, does the interpretation remain the same?. The study was based on the quality of visualization. This is very similar to Greenacre's study, in that the interpretation of the principal axes becomes important (Greenacre, 2006).

The symmetric map (Greenacre, 2006) or bi-plot displays both the row and column weights on the same axes however it tends to be very difficult to interpret the distance between the rows and columns since rows and column weights do not span the same space

[3] A group of scholars in Holland which is led by Jan De Leeuw under a pseudonym called Albert Gifi, involved in the development of CA by exploring more uses of MCA, this development is called homogeneity analysis and it fits in a wide range of multivariate methods. This development is thus popularly known as "Gifi System". (Greenacre, 2017; Michailidis and de Leeuw, 1998)

(Nishisato, 1995). Thus in this study, we will use one-mode plots, where we plot rows and columns in separate axes, respectively. Attention will be given to the interpretation of columns, i.e., categories, since a large number of rows in multiple-choice data often result to what is called *tall data sets*, where the number of rows is often greater than the number of questions (Markos et al., 2009). In addition, the distance and similarity measures will be used in this study, specifically cosine similarity (Kalhori et al., 2018) and Euclidean distance (Singh, 2017). Thus, stability in these studies has only been assessed for multiple-choice data and contingency table, specifically in the context of visualization.

Other approaches are undertaken in different studies to assess stability. Most of them are based on resampling methods (Balbi, 1992; Lombardo et al., 2012; Lebart, 2006; Michailidis and de Leeuw, 1998). Balbi (1992) approaches the stability analysis with a different view, where resampling is used slightly differently in that the resampling is more stratified than just random sampling without replacement, that is, stratified in this case means that the rows in the resamples have the same distribution as the original data, except that the size may be different. In other words, if the distribution of the total females in the data is 30 % of the data, then it remains 30 % even on the resamples. This is an important aspect especially if the objective is to generalize the results: it is useful to know that the resamples are representative of the population especially when the population is known. However, in most analysis, there might not be any information about the population, and in this case, we only rely on the sampled dataset at hand. The resampling methods in (Balbi, 1992) are compared based on vector spaces using the average of replications.

Another approach pursued in assessing the stability of quantification methods for contingency tables involves investigating statistical variability by formally testing the "principal inertias" for significance. This is done through a multivariate normal approximation to the multinomial distribution of eigenvalues (Greenacre, 2006). Greenacre (1984) investigated the stability of CA, in which he classified stability into two categories, the internal and external stability. This has influenced many approaches to stability analysis. Internal stability is only concerned with the data at hand and has no intention to generalize the results to a population while external deals with generalizing the results. Greenacre (2006) argues that the internal stability can still give insight even about the population, and it is suitable for the data acquired by convenience sampling (Greenacre, 2006; Michailidis and de Leeuw, 1998).

Bootstrap has been used in different settings with intent to assess the stability of ordination techniques (Knox and Peet, 1989) such as Detrended Correspondence Analysis (DCA) and other quantification techniques which are not pursued in this study such as

Non-Linear Principal Component Analysis (NLPCA) (Linting et al., 2007), and homogeneity analysis (Leeuw, 1973).

Commonly, bootstrap confidence intervals and confidence regions are assessed in different studies, where the aim of bootstrapping is to estimate the sampling distribution of the parameters of interest. Confidence regions are constructed for the parameters of quantification (eigenvalues, inertias, and optimal weights). Recent studies such as (Greenacre and Blasius, 2006; Lombardo et al., 2012; Linting et al., 2007; Greenacre, 2007; Lockyear, 2013) examine the ellipses or convex hull(Lombardo et al., 2012; Lockyear, 2013), and the interest is given to the coverage percentage. This approach is indeed appealing but it can be improved by including some aspects of stability.

This approach can be cumbersome when the aim is to assess the stability of the quantification solutions but the corresponding findings are based on the number of bootstrap replications. Hence this study does not only focus on the resampling specifications but to look at different sampling strategies. The issues related to outliers have been discussed in the following studies (Linting et al., 2007; Michailidis and de Leeuw, 1998) but the problem of stability as discussed in the Gifi system (Michailidis and de Leeuw, 1998) is deeper than just merging the categories (Nishisato, 2014). The question an analyst should ask is based on sampling, "What would happen if the data is sampled uniformly? Do the results change as the sampling strategies change ?". One way of evaluating this claim, subsampling, is not pursued in these studies (Greenacre, 2007; Linting et al., 2007; Lombardo et al., 2012). Subsampling can be very helpful especially with handling large datasets since it allows the data to be represented in a small amount of space without significant loss in accuracy (Aggarwal and Sathe, 2017).

2.9 Conclusion

In this section, we presented quantification methods and the background of these methods. These are descriptive methods for multivariate data and they can be used for EDA. These methods can visualize the relationships between categorical variables, by representing them in lower-dimensional space. Since quantification methods are classified as descriptive methods it may be difficult to know whether the visualized patterns are real or they occur by chance. One approach in assessing the stability of quantification results is through the use of resampling methods such as Bootstrap, Jackknife and Permutation Tests. These methods have been successfully used to construct confidence intervals and convex hulls. These methods are based on sampling several times from the parent sample to apply the quantification method on the resulting data. Interpreting the results from resampling methods sometimes become cumbersome when an analysis

shift focus and dwell on the number of replications. These methods are often applied to contingency tables, however, various interesting categorical datasets are often neglected or given less attention, specifically multiple-choice data. Quantification methods may be able to visualize the associations between categorical variables. The visualization of row and columns is not clear since the rows and columns do not span the same space. Another problem is that stability has been studied in the context of visualization, and whilst visualization is not bad, there are still many other properties that analysts may need to look at. Properties such as the effect of different sample sizes and the effect of sampling conditions. Thus the aim of this research aims to look at these sampling strategies and these properties on the data.

Chapter 3

Research Methodology

3.1 Introduction

In this chapter, possible ways of assessing the stability of quantification results are considered. The purpose of this chapter is to outline the methodology followed to come to methodologically sound conclusions. Figure 3.1 illustrates in detail all the steps followed in this research. Furthermore, Figure 3.1 provides a breakdown of the subsequent sections and explains each piece of the methodology.

Figure 3.1 shows how the generated multiple-choice data (see Section 3.2) was subsampled (see Section 3.4) using various subsampling methods. Furthermore, it shows exactly how the parent data is resampled in different ways (see Section 3.5), using Bootstrap. In each of the subsamples, resamples and the parent sample the figure illustrates how the quantification, in particular, DS is applied to each dataset (see Section 3.3). To further evaluate the stability of DS, finally, classification is performed on each of the datasets, using the parent sample as the ground truth (see Section 3.6).

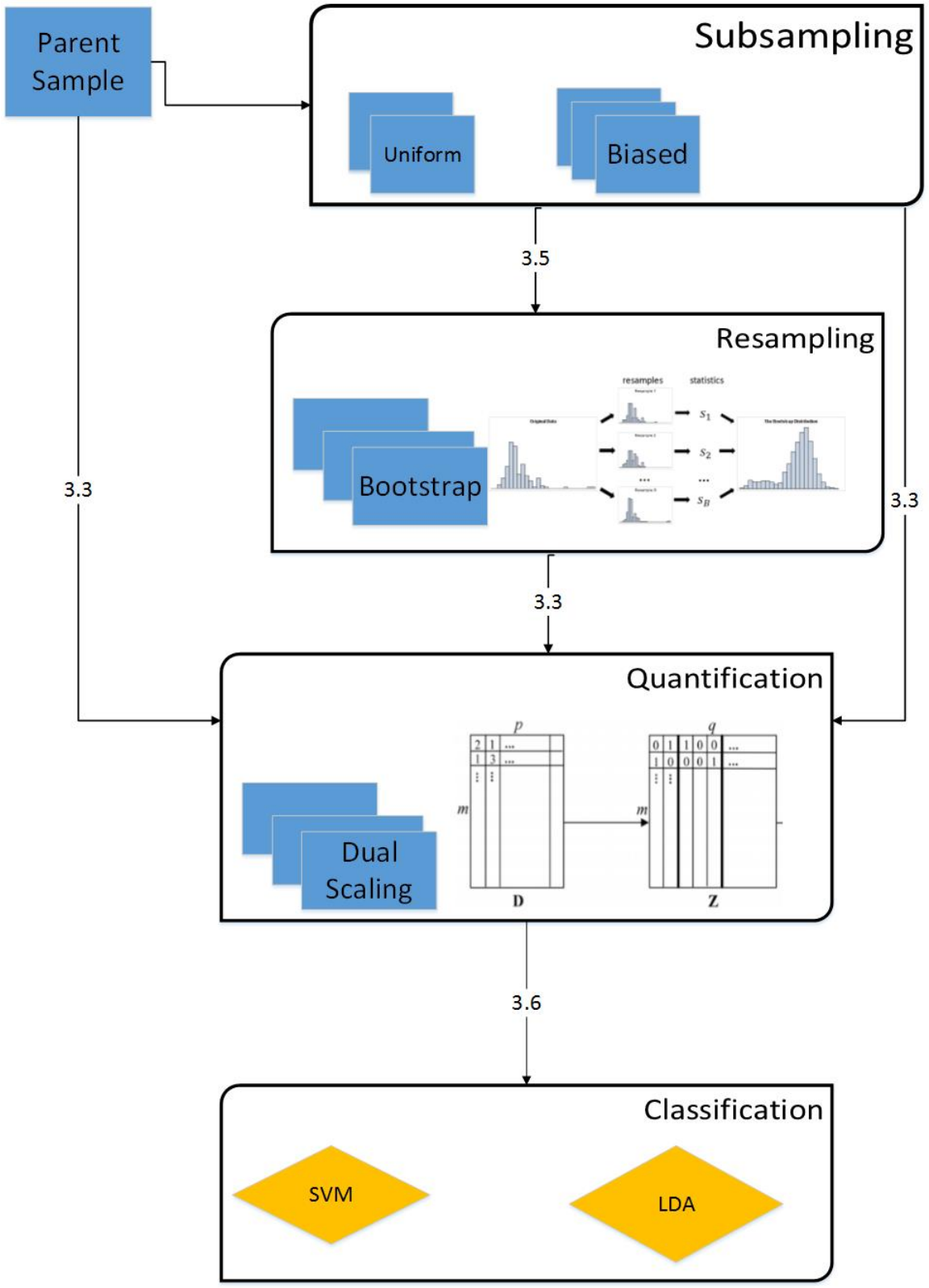


FIGURE 3.1: The procedure to be followed in this research. The numbers that appear on top of the arrows in the diagram refers to the section numbers and they will be discussed later in detail (see Section 3.3 - Section 3.5)

3.1.1 Chapter Organization

The rest of this chapter is organized as follows: firstly a discussion on the dataset used for the experiments is presented in Section 3.2, following that in Section 3.3 a quantification method that is used in this study is discussed. Section 3.4 presents subsampling methods used in this research, while Section 3.5 describes how the dataset was resampled to assess stability. The classification using DS axes or dimensions is described in Section 3.6 and finally, Section 3.7 presents experiments that were undertaken to test the hypothesis statements. In this section, the brief discussion on how the results are interpreted is also discussed.

3.2 Data

In this section, the data that is used in this study and the technique that was used to generate it is discussed. Section 3.2.1 addresses the generation of data while Section 3.2.2 further describes the data itself, namely, the structure of the data table, the sample size and the categories that are used in the entire data.

3.2.1 Data Generation

In this study, a multiple-choice dataset which was generated using a probabilistic graphical model was used. The motive behind using a synthetic dataset was so that it is possible to have control over the properties of a dataset that is under investigation in this study. This approach may have its disadvantages in generalizing the results but the arguments that are made about this dataset can be applied in various datasets which are similar to this one, including census, questionnaires, and surveys datasets.

The following section describes how the probabilistic graphical models work, specifically for generating datasets.

3.2.1.1 Probabilistic Graphical Models (PGMs)

A Probabilistic Graphical Model (PGM) is a model that employs a graph-based representation of complex distributions on a very high-dimensional space (Koller and Friedman, 2009). The nodes of these directed graphs correspond to the variables and the edges contain the information about their interactions and constitute what is known as joint probability distributions. PGMs employ Bayesian Networks (BN) to understand these distributions. BNs allows us to understand the dependencies of the variables by breaking

down the probabilities of each node, the technical details of BNs are discussed in Scutari et al. (2018). With the same notion of understanding complex distributions, the PGMs enable us to generate datasets with given prior probabilities.

Once the graph is well defined the sampling techniques can be used to generate data of any size (**N**), where the columns of these data consist of the nodes (variables) and the rows are the data points gathered from the sampling. To illustrate further, the process undertaken in this study to generate our synthetic data is hereby described.

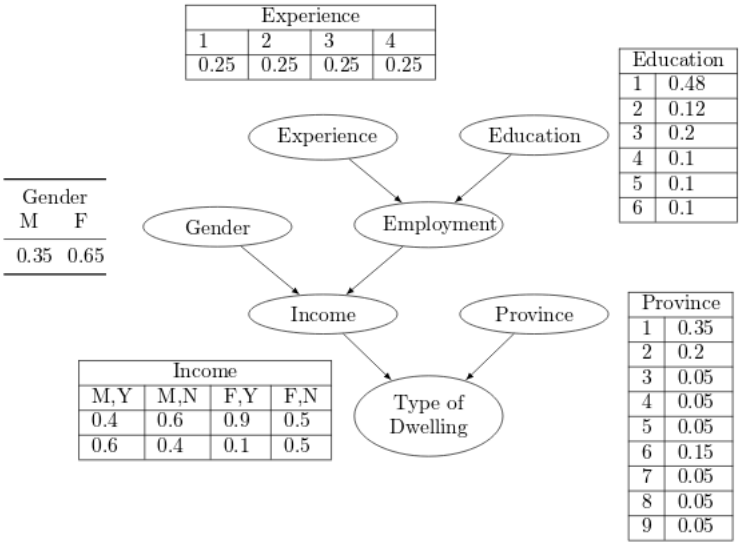


FIGURE 3.2: Bayesian Network graph for the multiple-choice dataset used in our study, nodes represent variables, while the edges show the interactions between these variables. The tables next to these nodes show the distribution of these variables. Two tables are missing in the diagram, one is for "Employment" and the other is for the last node, "type of dwelling" this is due to the length of the tables.

The first step that was needed in this study was to identify the nodes (variables). The graph as indicated in Figure 3.2 has 7 nodes, which are common variables used in the census data. Thus the data have 7 categorical variables each with a varying number of options/ categories. Of these nodes, there are several parent nodes, in particular, 4 nodes (Gender, Experience, Education and Province). The remaining 3 nodes have dependencies, meaning they have edges leading to them.

Considering the way South African population is distributed, as of the 2011 census data, the probability distribution tables for each of the nodes was created. The tables vary for each node as they appear in the diagram: some tables correspond to the marginal probabilities while others correspond to joint probabilities. For instance, it can be seen that *P (Gender)*, *P (Experience)*, *P (Province)*, *P (Education)* are marginal distributions and the remaining nodes follow joint distributions since they have dependencies.

Furthermore, although the variables are associated with common variables on census data, the probabilities indicated in the graph are completely arbitrary.

Finally, after defining the distributions of interest, a sampling mechanism is employed, specifically, Gibbs sampling (see Algorithm 1), which samples data points from a prior probability, where the new values to be generated are based on the previous values. Thus the basic idea of Gibbs sampling is that it makes a probabilistic choice for each of the variables to be sampled as opposed to picking values based on the next variable. This is a desirable property since sampling can be done from the predefined probabilities (Venugopal and Gogate, 2012).

Algorithm 1 Gibbs Sampling

procedure GIBBS SAMPLING
  $x^0 = x_1^0, ..., x_k^0$
  for t=1 to **N do**
    for each integer i in k **do**
      $x_i^{n+1} \sim P(X_i|x_1^{n+1}, ..., x_{i-1}^{n+1}, x_{i+1}^n, ..., x_k^n)$
    end for
  end for
end procedure

Consider algorithm 1, x^0 here refers to a random initial point, and

$$P(X_i|x_1^{n+1}, ..., x_{i-1}^{n+1}, x_{i+1}^n, ..., x_k^n) = \frac{P(X_i|x_1^{n+1}, ..., x_{i-1}^{n+1}, x_{i+1}^n, ..., x_k^n)}{\sum_{X_i} P(X_i|x_1^{n+1}, ..., x_{i-1}^{n+1}, x_{i+1}^n, ..., x_k^n)}. \tag{3.1}$$

Equation (3.1) implies that there is some blocking that is done when sampling from a BN, that is, sampling from conditional dependencies, for instance, to sample Gender the following probability expression can be used,

$$P(Gender|Experience, education, province, ..., Income) \tag{3.2}$$

that is, the probability of Gender given the remaining variables.

3.2.2 Data Description

In this section, a detailed description of the parent dataset is given. Table 3.1 shows a summary of description for our data. The original data which was generated consists of 7 categorical variables and 120000 rows of responses. The variables differ in the number of categories. Finally, only 12 % of the full data was used for analyses and this data also resembles the distributions of the full dataset.

TABLE 3.1: The description of data

Population	**Parent sample**	**Features**
120000 rows 7 features with varying categories	12 % of the population, 7 features, With varying categories Total sample size = 14400 rows	Gender [1,2] Employment [1,2] Education [1-6] Experience [1-42] Type of Dwelling [1,2] Province [1-9] Income [1,2] 7 Questions with the total of 27 categories.

Further description of the categories as they appear in the parent sample is given in Table 3.2. The frequencies of each of the categories in Table 3.2 are further reported in Figure 3.3 with the exception of the Gender variable which is reported in Chapter 4 in Section 4.5. It is generally expected that the rows with similar profiles are placed together in the 2D plot, that is, they have similar values of quantification (Abdi and Béra, 2017). This study focused only at the association between columns, that is, between categories. Hence, the main aim of this research was to explore the effect of frequencies or counts in the categories of categorical data against the quantification methods.

TABLE 3.2: A table showing labels of different categories in our parent data

Feature (category)	Label
Gender_1	Male
Gender_2	Female
Employment_1	Unemployed
Employment_2	Employed
Education_1	No formal schooling
Education_2	Grade 12
Education_3	Bachelors
Education_4	Honours
Education_5	Masters
Education_6	PhD
Experience_1	No experience
Experience_2	1-2 Years
Experience_3	3-4 Years
Experience_4	5 or more
Type_of_dwelling_1	Informal
Type_of_dwelling_2	Formal
Province_1	Gauteng
Province_2	North West
Province_3	Eastern Cape
Province_4	Western Cape
Province_5	Northern Cape
Province_6	Limpopo
Province_7	Mpumalanga
Province_8	Kwazulu Natal
Province_9	Free State
Income_1	Low
Income_2	High

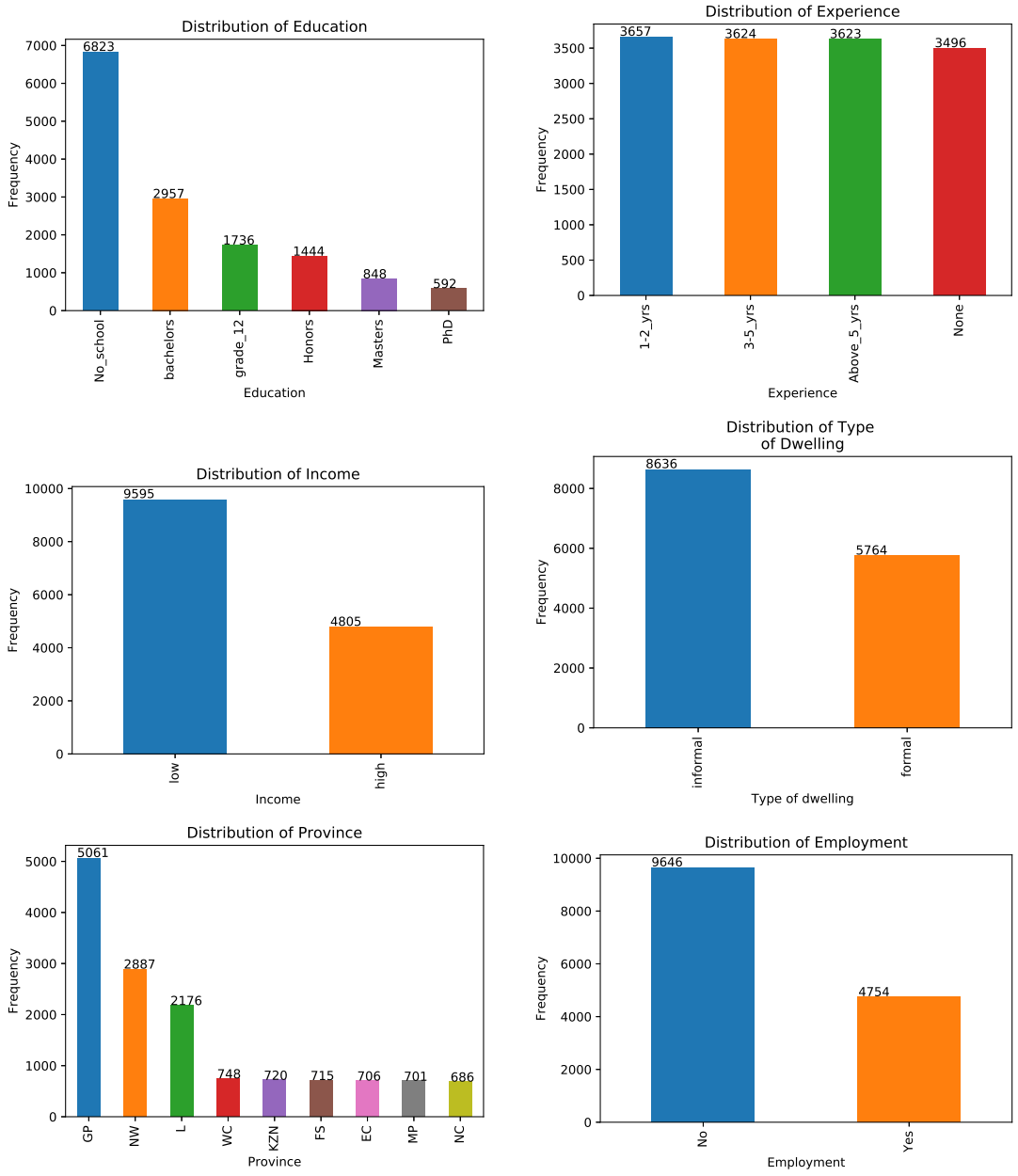


FIGURE 3.3: Data Description: Frequencies for each category

3.3 Quantification

Although there are several quantification methods mentioned in Chapter 2 in this study only Dual Scaling (DS) is considered since these different methods are closely related to each other and give similar results when decomposing an indicator matrix (Nishisato, 2014). Quantification was applied, in particular, DS, first to the parent dataset, then to the subsampling, and finally the resampled data. Therefore a quantification of the parent data was used as a point of reference or ground truth for the entire study. Following this, the dataset was subsampled in different ways, specifically, uniform subsampling

which is random sampling was used and finally biased subsampling which is convenience subsampling was also applied to the parent data. Following this DS was then applied on each subsample. Finally, DS was applied to the resampled datasets and the average results were compared with the DS results of the original data.

3.4 Subsampling

The parent sample was subsampled in different ways. The reason for this is was to know the effect of subsampling on the quantification results. In other words, this procedure aimed to assess the stability of results in different cases since stability is defined as the degree of sensitivity of quantification methods to the changes in the data. Uniform subsampling in which all the elements of the parent sample stand a chance to be selected, i.e., random sampling was considered in this study. Furthermore, the experiments were repeated 50 times to remove bias from the results. Another type of subsampling considered in this study was biased sampling, in which some elements of the parent sample have a higher chance of being selected than others.

In this type of subsampling, the extreme case where there is strictly no stratification was considered, that is, only the rows that contain a certain category were sampled. Consider a case where our target variable for biasing the data is Gender (any variable still leads to a similar conclusion), this variable is dichotomous with two categories being either male or female. In extreme cases only the rows that contain females (which in our data constitute 65 % of the data) were sampled, then this was also done for the male category. Finally, subsampling was done by considering both categories but place more emphasis on one category, for instance, 10 % of males and 90 % of females are randomly selected and investigation of the effect of biased subsampling in the DS results was conducted.

3.5 Resampling

Several samples were drawn from the parent set and subsampled datasets. This was carried out using resampling methods, specifically, Bootstrap in this study. The idea is that the samples are drawn with replacement from the parent set and subsampled data. The recommended number of resamples is likely to vary depending on the size and properties of the data (Linting et al., 2007; Markus, 1994). Pattengale et al. (2010) discussed in detail the different criteria regarding the number of replications in Bootstrapping, one of which is called a frequency criterion. In this criterion the researcher resamples with

a multiple of 50 (depending on the size of the machine) until convergence is reached, that is when the variance or standard error is no longer changing. Depending on the sample size of the data, some studies used as little as 100 resamples (Xu, 1999), thus in this study, different replications to determine the effect of resampling on multiple-choice data were explored. The quantification was then applied across the resamples of Bootstrap. As the resamples are obtained, quantification as then applied across them. The reason for this is that was to observe the effect of resampling and subsampling in the quantification results: do they remain the same or they change with the change in the dataset?

This study considered 3 experiments: firstly, the resampling of the parent sample, secondly, the resampling of the uniform subsamples and the resampling of the biased subsampled data was also considered. Finally, to strengthen results Permutation test p-values were used (see 2.7.3 in Chapter 2) . The results of the permutation test p-values accompany all the experiments and give precise significance about the variation (differences) in the data.

3.6 Classification

Our data has 7 categorical variables, of which 6 of them are used as predictors while one dichotomous variable is used as the target variable. Our target variable is the "Type of dwelling", which happens to be the bottom node in figure 3.2.

Dual scaling was also applied on the parent data biased using other target variables which are dichotomous for binary classification. Dual scaling results that are returned include row and column weights, both normed and projected weights. Our interest is given to the projected row weights, ρy, that is, in this case, the row weights are used to predict the type of dwelling. The projected weights are used since they contain more information about the original data and they are thus capable of reconstructing the original data with which the results come from. After obtaining the predictors, ρy, the classification methods (Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM)) are applied to these weights several times until convergence is attained, beginning with 10 repetitions. The classification was applied to the uniform, biased subsamples and finally to the resampled datasets.

Consider the confusion matrix in Table 3.3. A confusion matrix is a performance evaluation metric for classification problems where the true values are known, that is, it is a performance evaluation measure for supervised learning models (Ohsaki et al., 2017). The entries of this matrix are defined as follows:

- True Positives (TP): The probability of classifying a class positive when it is positive.
- True Negatives (TN): The probability of correctly classifying negatives.
- False Positive (FP): The probability of incorrectly classifying negatives as positives.
- False Negative (FN): The probability of incorrectly classifying positives as negatives.

TABLE 3.3: Confusion Matrix

	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

Accuracy, recall and precision are used to assess the variation between the parent sample, the subsample and resampled dataset. Accuracy is defined as the proportion of correct classifications (that is True Negative (TN) and True Positive (TP)) from an overall number of cases, and thus accuracy can be computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.3}$$

The recall is defined as the proportion of correct positive classification from the cases that are actually positive:

$$Recall = \frac{TP}{TP + TN}. \tag{3.4}$$

Finally, the precision is defined as the proportion of correct classification from the cases that are predicted as positive. Therefore precision is calculated as follows:

$$Precision = \frac{TP}{TP + FP}. \tag{3.5}$$

3.7 Experiments

Several experiments were conducted to test the hypotheses underlying our study namely: "There is no significant difference between the quantification results of the subsampled

and original parent data" and "There is no significant difference between the quantification of the resampled and the original parent data". The first comparison is on the quantification results, in particular, the eigenvalues and the optimal weights, where the eigenvalues refer to the squared correlation ratio η^2 and optimal weights refer to the quantification values of the rows and columns. Finally, to evaluate the stability of quantification results, the results of the classification were compared.

3.7.1 Uniform subsampling vs the original data

One of the experiments that were conducted was the comparison of the uniform subsampling results against those of the original data. In this experiment, the objective was to investigate the effect of uniform subsampling in the quantification results. The specific results of interest were the eigenvalues (structure of the data), projected column and weights (one-mode plots). The two results were compared in terms of two aspects, namely: visual inspection and numerical results. The structural differences between the eigenvalues and the one-mode plots of the two results were compared using an error metric (Mean Absolute Error (MAE) [1]). The Mean Absolute Error is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x - x_i| \tag{3.6}$$

Where, x represent the actual values and x_i are the estimated values. In our study, the actual values compared are the eigenvalues of the parent data and the estimated values refer to the eigenvalues of subsampled or resampled data, where possible.

Furthermore, the permutation tests were conducted to validate the results. Permutation tests were able to provide us with the actual distribution of the structural differences between the original and the subsampled datasets.

To interpret permutation test results, the p-values were used in this way, the larger the p-value indicates that the results are more similar to each other. Permutation tests are discussed in Section 2.7.3.

The error metric, MAE is also able to indicate whether the two underlying results are similar or different. The smaller the error, the more similar are the underlying results between subsampled and the parent sample.

The following steps were followed to accomplish this experiment:

[1] MAE is defined as an average absolute difference between estimated and the actual values (Nguyen et al., 2019). The estimate values refer to those of the subsampled data and the actual refers to the original parent data in the study and the classification results

- Step 1: Uniform subsampling of the original data
  Given the original data, which is also called Parent data. The data was subsampled without replacement several times (replications=minimum of 10 times) and the average results are retained. In this step, for example, 1 % of 14400 rows is sampled without replacement. The 1 % sample is later on used for training classifiers and the remaining 99% is then used for testing.

- Step 2: compare eigenvalues and the column weights
  After taking the averages, the average results of eigenvalues and the weights. The results are then compared using the MAE. The higher the MAE between original data and uniform subsampled data, the more dissimilar the two results. During this stage, the only data considered is the training data. Lastly, the permutation results were considered, particularly, the p-value of the permutation test (the higher the p-value the more similar are the results of the subsampled data to the original data).

- Step 3: Classification
  - During this stage, the remaining data is considered e.g. 99% remaining from the sampled 1%. The 99% data is then mapped to a dual scaling space. It is this data that is used for testing the classifiers, LDA and SVM.
  - The LDA and SVM were applied to the training sets.
  - The results tested on the mapped datasets, i.e, 99% for 1% dataset.
  - Finally, the comparisons are done between the uniform subsampled data and the parent data. The specific measures were accuracy, precision and recall.

3.7.2 Biased subsampling vs original data

The data was also biased in different ways. One kind of bias considered was the case where the data is extremely biased by one category, another is a more stratified bias where only a percentage of each category is taken. DS was applied to the resulting data and the results were carefully studied under the aforementioned comparisons, that is, the error metrics, the mean, variance and the classification results. To gain more insights on the results, permutation tests were applied to the results of subsampled and the parent data.

This experiment aimed to investigate the effect of biased subsampling in quantification results, specifically, dual scaling results. The results of eigenvalues were examined, the row and column weights. The following steps were followed to complete the stratified biased subsampling experiment.

- Step1: Determining the sizes of the biased subsample

  In this step, the previous subset sizes were used to determine the number of rows for each category of interest. That is, one categorical variable was chosen, specifically, Gender, the variable could be any categorical variable, thus these results can be extended to other categorical datasets. This variable consists of two categories. Firstly the percentages of the sub-sample sizes were used in the first experiment, for instance, a dataset was randomly sampled with 1% males and 99% females from the parent set of 14400 rows. The parent set consist of 5100 rows of males and 9300 rows of females. Therefore, taking 1% of the males and 99% of the females and thus form a new subsample with 51 males and 9207 females. This subsample now consists of 9258 rows in total.

- Step 2: Apply DS to each of the subsamples

  In this step, dual scaling was applied to each of the subsets. The results were then compared using MAE and the permutation test p-values.

- Step 3: Classification

  TRAINING

  The subsamples were then used the data was changed into a suitable form. That is, the "Type of dwelling" variable was singled out to make it a target variable for classification. Thus the remaining six variables were then used as the data for predictors. More specifically, dual scaling was then applied to the six variables. After that, the column weights ρy were used as predictor variables. LDA and SVM models were trained using these results from the subsets.

  TESTING

  For testing LDA and SVM models, parent data was used as an unknown data, this data was projected onto the DS spaces of each of the subset variables. Once the parent data is projected onto a new space classification models were then tested using the resulting datasets.

  - Finally, the comparisons were done between the biased subsampled data and the parent data. The specific measures were accuracy, precision and recall.

3.7.3 Uniform vs biased subsampling

The results of the two subsampling strategies were also compared to give insight into how similar or different the results are. In an event where the results are similar, that would indicate that the type of subsampling does not play a significant role.

This experiment aimed to compare the difference between the uniform and biased subsampled dataset. That is, which subsampling strategy gives more reliable results, reliable means consistent results and similar results to the parent set. Moreover, this experiment aimed to test the significance of using one subsampling strategy over the other. Thus the subsampled datasets were compared with an equal number of rows. The following steps were taken to accomplish this objective.

- Step 1: Determining the sample size of each subsampled data

  In this step, the existing biased subsampled dataset from experiment 2 was used. The percentages for each sample size were then computed. For example, the first subsampled data had 9258 rows, then the percentage of these row over the parent set was computed as (9258/14400) * 100 = 64%. Thus in comparing the biased data against uniformly subsampled data, the parent data was then subsampled uniformly using the percentages computed from the resulting subsamples of experiment 2.

- Step 2: Apply DS and compare the quantification results

  Once the percentages are determined, the subsamples were randomly sampled uniformly, with replications. Then the DS was applied to each of the subsampled data and the average results were retained for comparison. As in experiment 1, specifically, the eigenvalues were compared, the row and column weights using the MAE and permutation tests' p-values.

- Step 3: Classification

  TRAINING

  After thorough comparisons, the subsampled data was trained, that is, the biased subsampled and the uniform subsampled data. LDA and SVM were applied to these datasets. Finally, the parent dataset was projected onto DS spaces of the subsampled data.

  TESTING

  LDA and SVM models were then tested on the projected data. This implies that the experiment yielded two sets of classification results. The results were compared these using the MAE. The smaller the error the more significant the results of the subsampled data.

3.7.4 Resampling parent data vs the original data

The dataset was resampled in different ways and the results were compared with the parent sample results. One way of looking at resampling was a Bootstrap of the parent

sample with 10 resamples and DS was applied to these resamples and the average results were compared with the parent sample results.

This experiment aimed to investigate the effect of resampling parent data over the quantification results, more specifically, dual scaling results. In this experiment, we pursued a method of resampling the parent data to its original size and several replications. In this experiment, we used Bootstrapping (resampling with replacement) to examine the effect of resampling original data on dual scaling results. The following procedure was followed to accomplish our objective. The procedure may be summarized as follows:

- Step 1: Determining the parameters and fitting dual scaling

  In this step, the original data was used to resample N replications with replacement. In this step, dual scaling as also fitted to each of these replications. Finally, an average of the replications was taken. 50 replications were used for this experiment.

- Step 2: Mapping the new data onto a DS space

  In this step, dual scaling results were used from the Bootstrapping to project the original data onto the DS space. That is, from 10 replications the resulting new data from the projection were also 50. Thus this data was used for testing the LDA and SVM classifiers.

3.7.5 Resampling vs biased Subsampling

Again the resampled results were compared with the results of the biased subsamples. Given that the results for the Bootstrapped biased and the biased subsamples were available, the similarity of these results was investigated in the same way as the comparisons before resampling, that is, they were compared using the error metric (MAE) and classification. It was important to compare the Bootstrapped results and biased subsamples. For this kind of comparisons, the Bootstrap of the biased subsamples against the biased subsamples that are not resampled was considered. Another comparison was that of resampling the biased data to recover the size of the original data. This kind of comparison aimed to investigate the effect of the resampling on the biased subsamples.

This experiment aimed to investigate the effect of resampling biased data over the quantification results, more specifically, dual scaling results. In this experiment, Bootstrapping was used to examine the effect of resampling biased subsampled data on dual scaling results. The following procedure was followed to accomplish our objective. In this experiment the method of resampling up to the size of the parent data was pursued, that is, a subsample was taken randomly from experiment 1 and resampled until the rows

of a new dataset are equal to 14400, which is the sample size of the parent data. The procedure may be summarized as follows:

- Step 1: Determining parameters

  In this step, the original data was used to resample N replications with replacement. In this step, dual scaling was also fitted to each of these replications. Finally, an average of the replications was taken. The number of replications used for this experiment was 50.

- Step 2: Project the parent data onto Bootstrap DS space

  In this step, dual scaling results from the Bootstrapping were used to project the original data onto the DS space. That is, from 50 replications the resulting new data from the projection were also 50. This data was used for testing the LDA and SVM classifiers.

3.7.6 Resampling vs uniform subsampling

Resampling was also performed on the uniformly subsampled data and the results were compared against the uniform subsampled data. At this point, a subsample that is a percentage of the data, say 1 % of 14400 (144 rows) was considered. The results for these subsamples already exist from the subsampling section, but in this section, those subsamples were then resampled up to the size of the original data.

The detailed procedure followed in this experiment is as follows:

- Step 1: Subsampling uniformly In this step, the data was subsampled uniformly and divided into training and test set. To illustrate this procedure 1 % of the data (144 rows) was subsampled as an example, this is our train set. Then the remaining 99 % of the data becomes our test set. It is worth noting that this was done several times to ascertain reliability.

- Step 2: Fit dual scaling to the train set In this step, dual scaling was fitted into the subsampled data and the results were retained for comparisons.

- step 3: Resampling up to the size of the parent data

  In this step the data obtained in step 1 was resampled, that is the subsamples. The subsampled data was resampled with replacement, the resamples have the size of the parent data, 14400.

- step 4: fit dual scaling on this data and compare with parent data: Dual scaling was the fitted to the resamples and results were retained for comparisons. The results were then compared with the parent data using MAE and permutation test p-values.

- Step 5: Classification of the column weights

  In this step, LDA and SVM were fitted on the train set that was obtained by resampling. Thereafter, the test set obtained in step 1 was projected onto the DS space of the resampled data. Finally, the LDA and SVM were applied on the test set, the accuracy, precision and recall were retained.

Most of these experiments involve some repetition to ensure that the results are reliable. The graphs that are produced on repeated experiments are reported with error bars, it is in this error bars confidence interval (CI) were used to illustrate possible variation. In particular, a 95 % confidence interval was used since this is a widely accepted CI. A confidence interval is calculated as follows:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \tag{3.7}$$

Where $\bar{x}$ is the average score, σ is a standard deviation, z is a standard normal distribution score (z-score) and n is a total number of score that are averaged. Thus in this study, $\bar{x}$ represent the average of either, MAE, p-values or the classification evaluation measures. The z-score is 1.96 for 95 % confidence interval, thus the equation becomes

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} \tag{3.8}$$

3.7.7 Similarity and Distance Measures

One of the key aspects of the experiments in this study is the visual inspection of the data, where the one-mode plot described in Section 2.8 of Chapter 2. The interpretation of the one-mode plots is made easier in this study by the use of similarity and distance measures. Cosine similarity and Euclidean distance were used. Cosine similarity is a measure or metric used to assess the degree of similarity between two vectors (Zhu et al., 2011; Kalhori et al., 2018). The vectors referred to in this study are the column weights, which are the categories in essence. Cosine similarity can be calculated as follows:

$$cos(x, y) = \frac{< x, y >}{||x|| \times ||y||}$$

.

where $< x, y >$ is the inner product between two vectors x and y. $||x||$ represent the norm of vector x.

Euclidean distance is a metric used to measure dissimilarity between two vectors. The calculation of euclidean distance is Pythagorean and is obtained as follows:

$$distance(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.9}$$

. where n is the number of elements in a vector and $(x_i - y_i)^2$ represent the sum of squared differences between the two vectors.

3.7.8 Hierarchical clustering

As an extension of similarity measures and giving a clear and concise grouping of categories, we also applied Hierarchical clustering to our column weights (px). Hierarchical clustering is one of the unsupervised clustering methods in which the categories or variables are clustered iteratively based on the distances (Govender and Sivakumar, 2020). This type of clustering can occur in two ways: the first clustering method begins with one data point as a single cluster and then with the assistance of distance measures the other clusters are then formed by combining similar data points sequentially in a hierarchy, this is called **agglomerative**. The second method is based on starting with one big cluster and subsequently divide the cluster in into smaller clusters. This is called divisive *hierarchical* clustering. In this study we used agglomerative hierarchical. The results of hierarchical clustering are often presented in a concise graph called "a dendrogram". This is a graph that shows how different data points cluster with each other to form a new cluster and this is recursive until there are no more clusters to join. The dendrogram is formed based on some metric, in this study Euclidean distance was employed. There are various ways to form the clusters:

1. **Singlelinkage** - This method combines clusters based on the minimum or shortest distance.
2. **Completelinkage** - The clusters are formed based on the maximum or furthest distance.

There is also an *averagelinkage*, *centroids* methods that may be employed alternatively. The first two methods are generally used simply because they are intuitive (Murtagh

and Contreras, 2017). In this study we used a *completelinkage* method because of its popularity. Furthermore, this study used "Euclidean distance" for the formation of clusters.

3.8 Conclusion

In this chapter the methodology underpinning this study was described: the data was briefly described and the quantification methods were described in the manner in which they are used. Since the main focus of this research is to investigate the effect of subsampling and resampling on quantification results, the analysis of the results involves the investigation through the use of classification methods, LDA and SVM. In the next chapter, the results of the experiments undertaken in this study are reported. The first three experiments were used to test the first hypothesis based on subsampling and the remaining experiments were used to test the second hypothesis that is based on resampling.

Chapter 4

Experimentation

4.1 Introduction

In this section, we present the results of our experiments. Several experiments were undertaken to test our two hypothesis statements (see Section 1.3). The first hypothesis states that there is no significant difference between the quantification results of the subsamples and the original parent sample, while the second hypothesis states that there is no significant difference between the quantification results of the resampled and the original parent sample. In essence, we hypothesize that the small change in data by mixing rows or columns of the data yields smaller changes in the results of quantification.

4.1.1 Chapter Organization

This chapter is organized as follows: Section 4.2 and 4.3 presents the dual scaling (DS) analysis and classification on the full data, respectively. In addition to this, the similarity and distance measures are presented in support of the results of the one-mode plot. Section 4.4 presents the comparison of DS for parent data vs the uniform subsampled data. Section 4.5 presents the DS of biased subsample vs the parent data and Section 4.6 is concerned with the DS of the uniform subsampled vs biased subsamples, that is, in this section, we investigate which case gives similar results to the parent sample given that the sample sizes are similar. The aforementioned experiments are associated with the first hypothesis statement. To test the second hypothesis the following experiments were undertaken: Section 4.7 presents the resampling of the parent sample, while Section 4.8 is concerned with the resampling of uniform subsampled data and finally the last comparison in Section 4.9 is concerned with the resampling of the biased subsample.

Section 4.10 we present a summary of the results. Finally, in Section 4.11 we present the sub-conclusion for Chapter 4.

4.2 DS of the parent sample

The purpose of this section is to present DS analysis of the original parent sample. This experiment is divided into two parts: the first is the DS on the parent sample, then the last part deals with the classification of the row weights, ρy. As mentioned before, the parent sample consists of 7 categorical variables and 14400 rows. The total number of non-trivial solutions that we can get from the data is defined as the total number of categories minus the number of questions. Let $X_{(N\times n)}$ be a data matrix, with $n = 7$ questions/variables and $N = 14400$ and the total number of categories be $m = 27$. Then the total number of non-trivial solutions from SVD decomposition of the indicator matrix[1] of our parent data (see Section 2.4.2) is given as: $m - n = 27 - 7 = 20$.

DS analysis is focused on two sets of results, firstly the eigenvalues and secondly the optimal weights. The eigenvalues, in this case, provide information about the structure of the data, and the optimal weights, in this case, refers to the row and column weights, where row weights refer to the quantification of respondents and column weights refers to the quantification of categories in our categorical data.

Figure 4.1a shows the distribution of eigenvalues, and this plot is referred to as a Scree plot (Patil et al., 2008). The eigenvalues that result from the SVD of the indicator matrix obtained from the parent data are arranged in descending order where the first eigenvalue shows the maximum information accounted for by the first component. Figure 4.1a shows a plateau-like distribution of eigenvalues; this structure guides us on what to look for on the subsamples. The sum of all eigenvalues is called the *trace*. If we take each of the eigenvalues and divide by the *trace*, then this is called Percentage of Variance Accounted For (PVAF). The first few eigenvalues usually account for larger variance, hence our interest is focused only on the first few eigenvalues. In our data $trace = 2.857$ and the first two eigenvalues are, 0.228 and 0.217, respectively. Thus the first PVAF is $(\frac{0.228}{2.857}) * 100 = 7.986$ as shown in Table 4.1. The third row in Table 4.1 represents the cumulative PVAF for our eigenvalues which means that the first two eigenvalues account for 15.579 % of the total variance.

Once the eigenvalues are computed it is generally easy to compute the row and column weights corresponding to those eigenvalues. Row weights, in this case, refers to the weights of the respondents that is, numerical values corresponding to the rows while

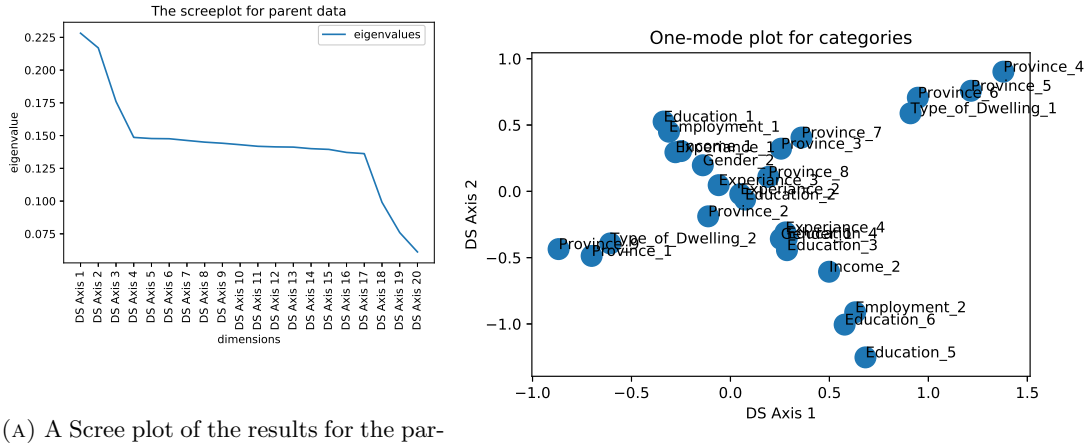
[1] (Gower et al., 2011)

TABLE 4.1: Summary of DS results for the parent data

	DS Axis 1	DS Axis 2
eigenvalues	0.228177	0.216944
PVAF	7.986200	7.593057
Cum PVAF	7.986200	15.579257

column weights are the numerical values for the categories. These categories are visualized in a lower-dimensional space, to explore associations between the categories. Hence, figure 4.1b shows the associations between the 27 categories in a lower 2D space, that is dimensions associated with the first two components. This graph is called a one-mode plot. The closer the categories are to each other the more similar or related are the two categories. The one-mode plot is also referred to as a non-symmetric plot where the axes show the first and the second component of either the columns or rows. In this case, the one-mode plot of interest is for categories, which correspond to the columns. If the categories are close to each other then there is a strong association between these categories. The association in one-mode plots is measured by Euclidean distance (Nishisato and Clavel, 2009). Categories with similar relative frequency are strongly associated with each other.

Consider figure 4.1b which shows a one-mode plot for the categories of the original parent sample. The one-mode plot of the original parent sample resembles an X-like structure of categories. The X-like structure is not geometrically important but gives an idea of how the categories are associated with each other. In our data, Type of dwelling 2, that is informal settlement is associated with Gauteng province and Free State, while Type of dwelling 1, i.e. formal settlement is associated with province Western Cape, Limpopo and Northern Cape. A category for an employed individual is associated with an individual with a Masters or PhD. In essence, what these associations mean is that according to our parent data a person with PhD or Masters is likely to be employed and likely to have many years of work experience in their field while an individual with only grade 12 or no formal schooling is likely to have lesser work experience and also may reside in an informal settlement. These associations are further summarised in Table 4.2



(A) A Scree plot of the results for the parent sample

(B) One-mode plot

FIGURE 4.1: A scree plot and one-mode plot for the parent sample

TABLE 4.2: Notable similarities in the X-like structure

Group 1	**Group 2**	**Group 3**	**Group 4**
Education_No			
Employment_No			Experience_above_5
Province_MP			Education_Honors
Province_EC	Province_WC	Type_of_dwelling_2	Education_bachelors
Province_KZN	Province_NC	Province_FS	Income_high
Education_grade_12	Type_of_dwelling_1	Province_GP	Employment_Yes
Province_NW	Province_L		Education_phD
Experience_None			Education_Masters
Experience_1-2yrs			Gender_male
Gender_female			

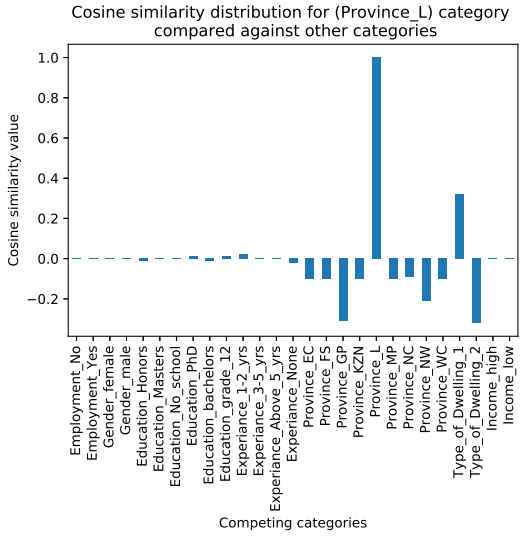
4.2.1 Similarity and Distance Measures

In this section, further exploration of the association between categories was performed on the column weights. The similarity measures considered include Cosine similarity measure. Furthermore, the distance measure used in this study was Euclidean distance.

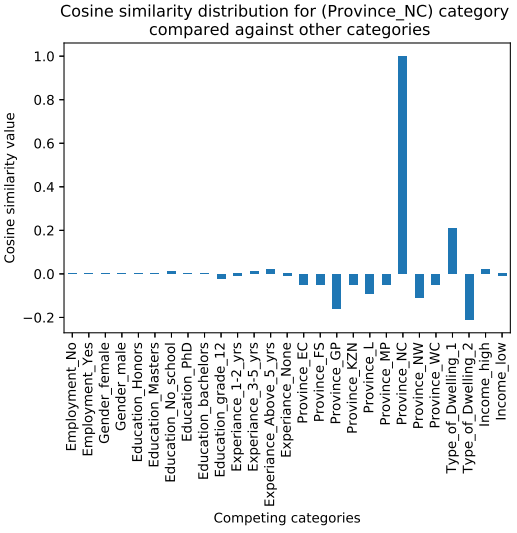
The results for the cosine similarity measure are reported in Figure 4.2. The results should be interpreted with respect to the groupings in Table 4.2, specifically, with Group 2 categories. The range of cosine values for this data is [-1,1], this is because the column weight vectors contained negative numbers. In this case, "0" means that two categories are dissimilar, "1" very similar and "-1" means that two categories are diametrically opposed, that is, the exact opposite.

Consider Figure 4.2a, "Province_L" and "Type_of_dwelling_1" have the cosine similarity value of about 0.3 and this means these categories are 30 % similar, whereas

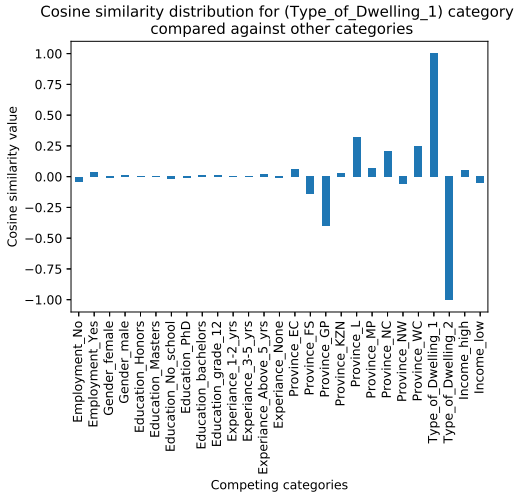
when compared with "Province_WC" the "Province_L" is exactly opposite. To further explain these association, consider Figure 4.2b, "Province_NC" is more similar to "Type_of_dwelling_1". In Figure 4.2c, "Type_of_dwelling_1" is strongly associated to "Province_L", followed by "Province_WC" and "Province_NC" and this clearly depicted in the one-mode plot in Figure 4.1b [2].



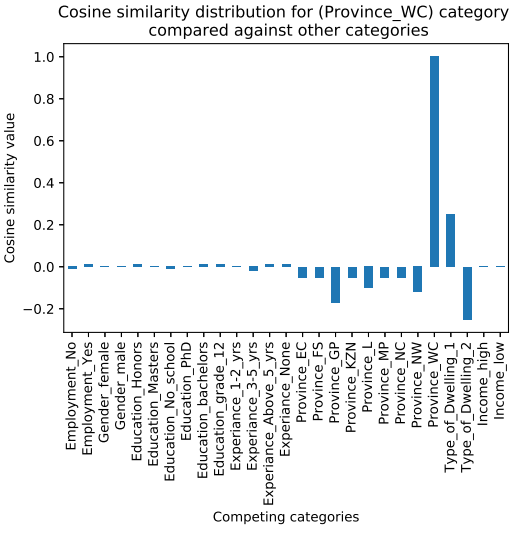
(A) Limpopo province cosine similarity



(B) Northern Cape province cosine similarity



(C) Type of dwelling 1 cosine similarity



(D) Western Cape province cosine similarity

FIGURE 4.2: Cosine similarity values of some categories, the categories in consideration are from Group 2 (see Table 4.2)

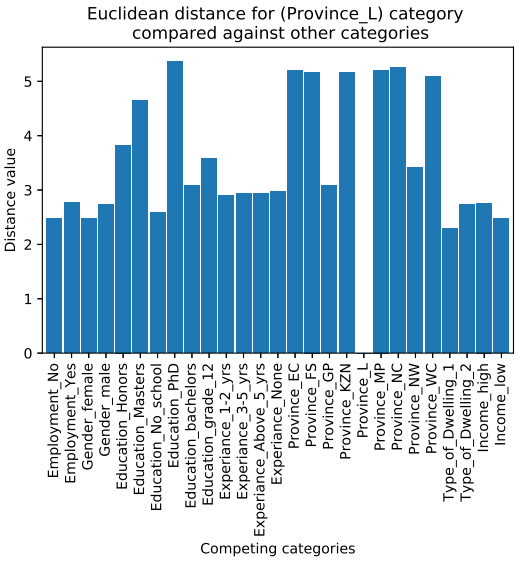
Euclidean distance measure was also used in the column weights and the pairwise Euclidean distances are represented in the graph reported in Figure 4.3. The higher the

[2]The rest of the results of cosine similarity for parent data are reported in Appendix C

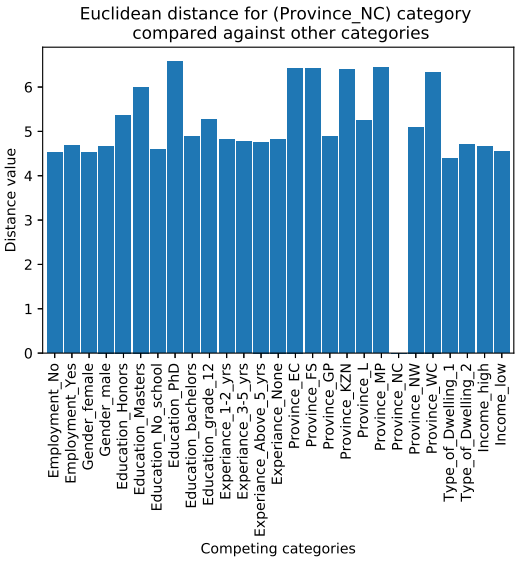
euclidean value, the more dissimilar the vectors. The similar categories as presented in Figure 4.3 are close to each other.

Considering Figure 4.3a, "Province_L" is closer to "Type_of_dwelling_1" with a euclidean distance of 2.29, followed by "Income_low" and "Employment_no" at 2.47. Similarly, Figure 4.3b shows that "Province_NC" is closer to "Type_of_Dwelling_1" with a euclidean distance of 4.39, followed by "Employment_no" and "Gender_female" at 4.53. It is worth noting that these relations are maintained in these different diagrams, the three provinces (Figure 4.3a to Figure 4.3c) are all closer to "Type_of_Dwelling_1" [3].

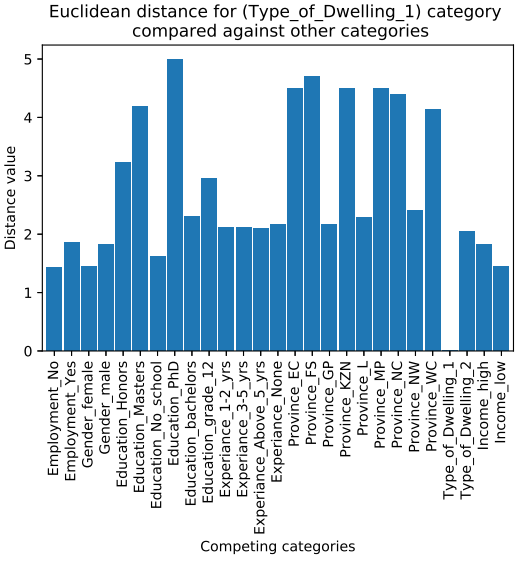
[3]The rest of euclidean distances are reported in Appendix C



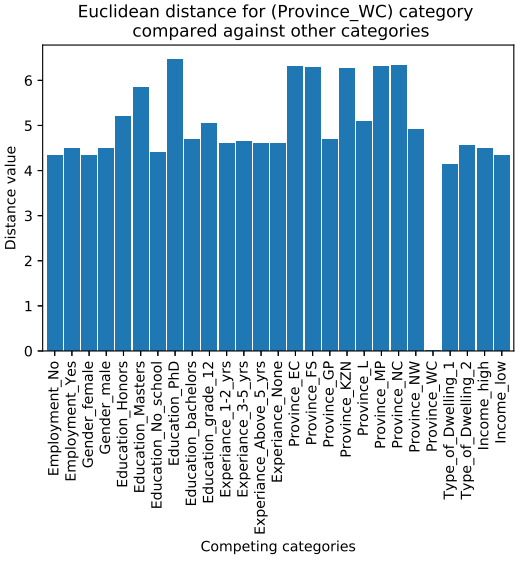
(A) Euclidean distance for Limpopo province



(B) Euclidean distance for Northern Cape province



(C) Euclidean distance for Type of dwelling 1



(D) Euclidean distance for Western Cape province

FIGURE 4.3: Euclidean distance values of some categories, the categories in consideration are from Group 2 (see Table 4.2)

In order to simplify interpretation on similarity or distance measures, hierarchical clustering was performed on the column (px) weights. The results are reported in Figure 4.4. The results show evidence of 3 clusters as opposed to our initial 4 groupings that are reported in Table 4.2. In Figure 4.4 the first cluster is formed by the categories that were initially grouped to group 2 in Table 4.2, the second cluster is formed by the combination of higher education qualifications, male gender, experience above 5 years category and high income and the remaining categories form a third cluster which is much bigger than the first two clusters. As the chief aim of this study was to assess the

stability of results, these clusters should not be largely affected by minor changes in the data.

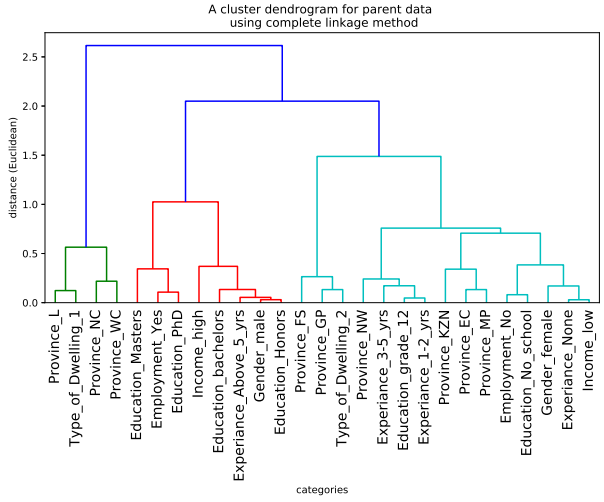


FIGURE 4.4: A dendrogram showing the number of clusters formed from the column weights, px

4.3 Classification of the parent data

To establish the ground truth for the entire study, we performed LDA and SVM classification on the original data where the training data consisted of the 6 predictor variables (i.e. All variables excluding "type of dwelling") and target class was "type of dwelling". In other words, we trained SVM and LDA on 14400 rows and then tested the classification models on the same data. In this experiment, we retained the accuracy, precision and recall classification measures as described in Chapter 3 (see Section 3.6). The results are shown in Table 4.3.

The LDA accuracy lie between 73 % and 79 % while SVM accuracy lie between 71 % and 79 % at 95 % confidence interval. The precision for LDA lies between 68 % and 76 % while it lies between 71 % and 75 % for SVM. Finally, the recall evaluation measure for LDA lies between 62 % and 66 % while SVM recall is between 58 % and 66 %.

TABLE 4.3: Classification results of the parent data

	lda_accuracy	lda_prec	lda_recall	svm_accuracy	svm_prec	svm_recall
results	0.76 (+/- 0.03)	0.72 (+/- 0.04)	0.64 (+/- 0.02)	0.75 (+/- 0.04)	0.73 (+/- 0.02)	0.62 (+/- 0.04)

4.4 Experiment 1: Uniform subsampling vs the original data

The purpose of this section is to investigate the effect of subsampling the parent data uniformly on the quantification results, in particular, DS. Important aspects to look for in this section are the effect of different sample sizes on the DS results. This section together with section 4.5 and section 4.6 allows us to test the first hypothesis statement which states that "there is no significant difference between the quantification results of the original parent sample and the subsampled data". The two major results of quantification include eigenvalues and the optimal weights. The first comparison in this experiment is based on eigenvalues and the visual inspection of column weights, i.e., quantification results of categories. The second part is about the classification of row weights, ρy.

Several subsamples were considered, from the smallest subsample to the largest subsample. The data was subsampled into 1 %, 2 %, 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40%, 45 %, 50 %, 55 %, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 % and 99 % of the full data, where the size of the full data is 14400. The results of eigenvalues and optimal weights were compared using the error metric (MAE) and the classification results. The MAE in this experiment measures the difference between the eigenvalues of the original parent sample and the ith subsample.

To minimize bias in our uniform subsampled data, the experiments were repeated several times, then dual scaling was applied in each of the subsamples. The average of the results was computed. A summary of the scree plots for the different subsamples is reported in Figure 4.5.

4.4.1 Visual inspection

In this section, we explore the differences between the DS results of the uniform subsampled data and the original parent sample. The Scree plots and the one-mode plots are used to visually compare the DS results. Figure 4.5 shows the Scree plots of different uniform subsampled data. The first two subsamples (1 % and 2 %) deviate largely from the plateau-like structure obtained from the original parent sample in Figure 4.1a. This is reasonable since 1 % and 2 % are 144 and 288 rows of the full data, respectively. Considering the last two subsamples in our data, 95 % and 99 %, that is, 13680 and 14256 rows, respectively. The scree-plots of the two subsamples in Figure 4.5 are very similar to the plateau-like structure of the original parent sample in Figure 4.1a hence

for this, we claim that small unimportant change in the data results in a small unimportant change in the DS results. This means that when we subsample small subsamples then this is a hugely important change. We included 1 % and 2 %, all up to 99 % simply because we were interested in testing the stability of smaller subsamples. The magnitude of this small change will be investigated later in this chapter.

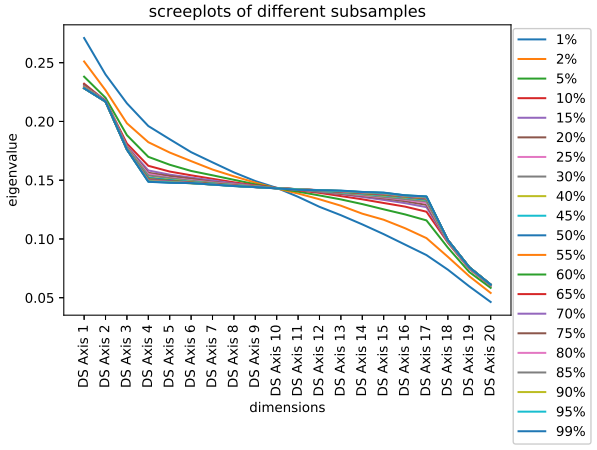


FIGURE 4.5: Scree plots of different uniform subsamples

Following the results of eigenvalues, one-mode plots for each of these subsamples were investigated and the results support the same claim that 1 % and 2 % of subsample are just too small as subsamples. To elaborate on this point, we first briefly describe the structure of the one-mode plots for the parent data:

- Consider Figure 4.1b, the parent sample shows an X-like structure of the categories. Related categories are closer to each other in the one-mode plots, however, this does not tell us about the strength of association.
- Further breakdown on this structure, Table 4.2 shows that the categories that are associated with each other can be summarised in the groups that appear here.

On the contrary, the one-mode plot for the 1 % and 2 % subsamples show a very different structure. The one-mode plots in Figure 4.6a and Figure 4.6b are very different from each other. We also considered the last two subsamples (95 % and 99 %) as these are kinds of subsamples that result from small changes in the parent data. Consider Figure 4.7a and Figure 4.7b, respectively. These figures both show a similar pattern to the pattern obtained from the parent sample in Figure 4.1b since 95 % and 99 % subsamples mimic the parent data which means that in our data 95 % and 99 % is enough to give quantification results similar to the original parent sample.

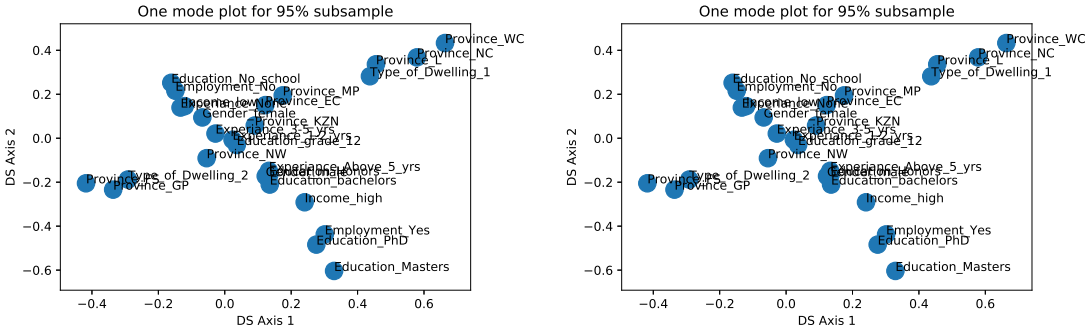
To show to what extent are the two sets of subsamples different from the results of the parent sample, we used the Mean Absolute Error (MAE) metric, permutation tests

p-values and classification of the row weights. Section 4.4.2 further breaks down the details of the comparisons. The rest of one-mode plots for other uniform subsamples are reported in Appendix A.



(A) One-mode plots for 1% subsample (B) One-mode plots for 2 % subsample

FIGURE 4.6: One-mode plots for 1% and 2 % uniformly subsamples



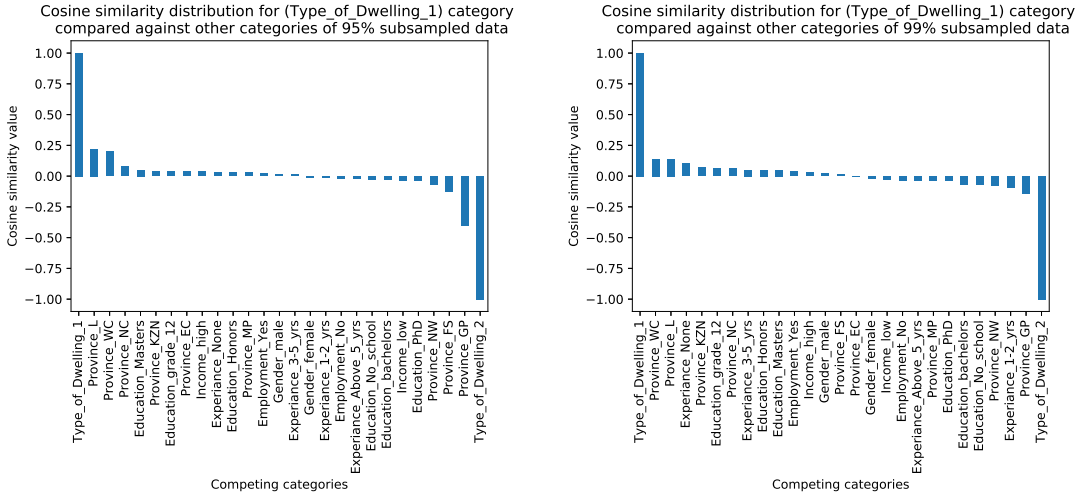
(A) One-mode plots for 95% subsample (B) One-mode plots for 99 % subsample

FIGURE 4.7: One-mode plots for 95% and 99 % uniformly subsamples

We also used cosine similarity and Euclidean distance to investigate whether there is any major change of the quantified subsampled data when compared to the quantified parent data. The results of cosine similarity and Euclidean distance are presented in Figure 4.8a and Figure 4.8b. The results shown in Figure 4.8a and Figure 4.8b are associated with 99 % subsampled data, this is to emphasize the chief aim of this study, assessing the impact of small change in the data to the quantified data.

The category considered here is "Type_of_dwelling_1", hence Figure 4.8a and Figure 4.8b are compared closely with Figure 4.2c and Figure 4.3c in Section 4.2.1.In Section 4.2.1, the cosine similarity suggested that the most similar categories to the "Type_of_dwelling_1" category was "Province_L" followed by "Province_WC" and "Province_NC", respectively. However the Euclidean distance in Figure 4.3c suggested that the closest categories to "Type_of_Dwelling_1" were "Employment_No" followed by "Gender_Female" and "Income_low".

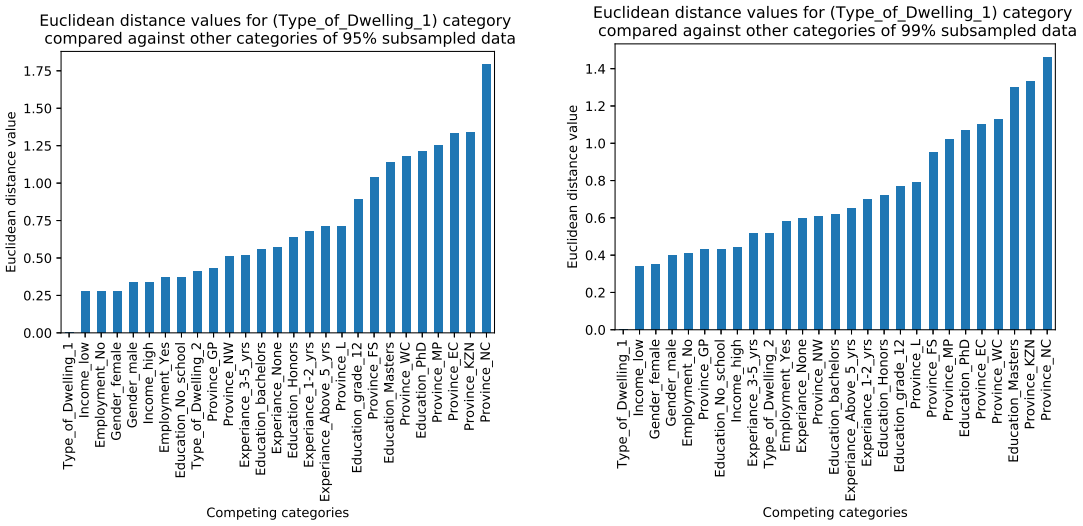
Considering the results of the subsampled data, Figure 4.8a, the categories that are similar to "Type_of_Dwelling_1" are "Province_L" followed by "Province_WC" and "Province_NC".



(A) Cosine similarity for 95 % subsampled data (B) Cosine similarity for 99 % subsampled data

FIGURE 4.8: Cosine similarity for 95% and 99 % uniformly subsamples

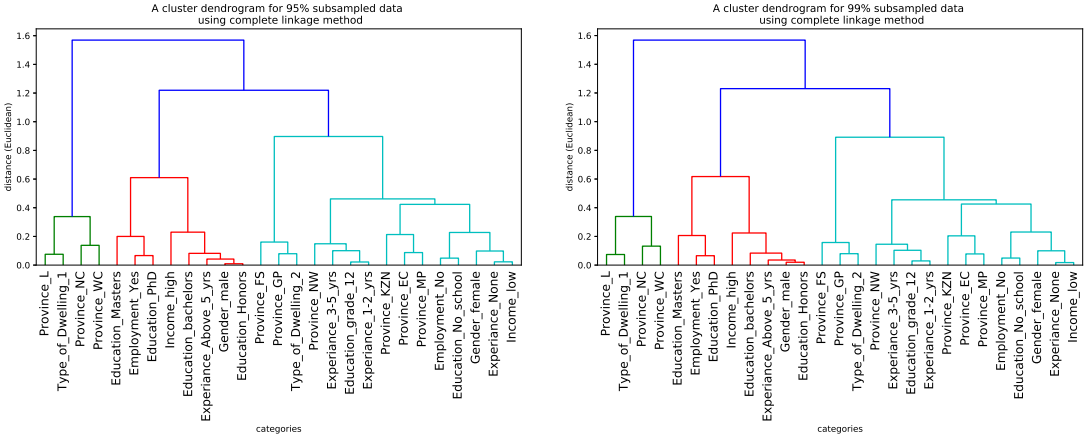
The Euclidean distances of the 95 % and 99 % subsampled data were also considered and results are reported in Figure 4.9a and Figure 4.9b. Although these Figure 4.9a and Figure 4.9b both show that "Type_of_Dwelling_1" is strongly associated with "income_low", "Employment_No", "Gender_female" and "Gender_male" in that order, one of the differences between Euclidean distance at 95 % and 95 % is that Figure 4.9a suggest that "Type_of_Dwelling_1" is more closer to "Income_high" than to "Employment_Yes". Figure 4.9b suggest that "Type_of_Dwelling_1" is rather closer to "Employment_Yes" than is to "Income_high". This indicates that these two subsamples still retain the interpretation of Euclidean distance that is similar to the one associated with the original parent data in Section 4.2.1.



(A) Euclidean distance for 95 % subsampled data (B) Euclidean distance for 99 % subsampled data

FIGURE 4.9: Euclidean distances for 95% and 99 % uniformly subsamples

The hierarchical clustering was also used to extend the application of similarity/distance measures on quantified data. The results for clustering are reported in Figure 4.10a and Figure 4.10b and these results are then compared with the results obtained for the parent data in Section 4.2.1. These two figures correspond to the 95 % and 99 % subsampled data, respectively. These two figures resulted in 3 clusters and the categories that are in each of these clusters are exactly the same as those in Figure 4.4 and this implies that the 5 % and 1 % change in the data does not affect results negatively or change the interpretation of the results.



(A) Hierarchical clustering results for 95 % subsampled data (B) Hierarchical clustering results for 99 % subsampled data

FIGURE 4.10: Hierarchical clustering results for 95 % and 99 % subsampled data

4.4.2 Comparisons

The primary hypothesis we wished to test is that the quantification results, in particular, the eigenvalues and the optimal weights, stay the same as the data is subsampled. Without loss of generality, throughout this study the stability of the quantification has been described as "the small and unimportant changes in the data should result in small and unimportant changes in the results" and with this definition, in mind, we explore the different sample sizes in the data from larger (1 %) to smaller (99 %) subsample. We hypothesize that the larger the sample size, the closer the results should be to the parent data results. These claims will be evaluated using both the MAE and classification results.

Figure 4.11 shows the graph of permutation test p-values. The graph is thus in agreement with the hypothesis that claims that the smaller the sample size of the subsampled data, the more the results deviate from the ground truth. The p-values are directly proportional to the size of the sample, that is, the smaller the subsample size, the smaller the p-values of the permutation test.

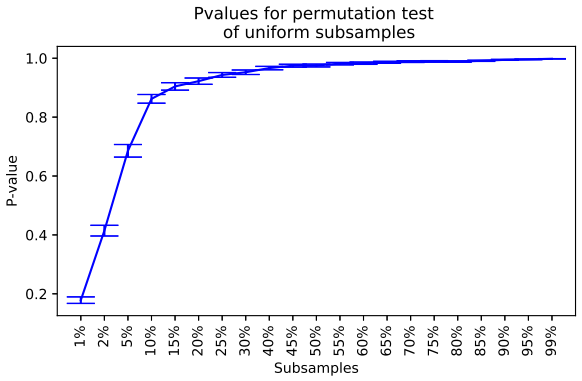


FIGURE 4.11: Permutation test p-values

The second part of our results pertains to the MAEs between the parent and subsample. The results are reported in Figure 4.12.

We compared the eigenvalues of the parent data against the eigenvalues of the resulting subsamples using MAE. What these tell us is how different or similar are the two sets of the eigenvalues. The higher the error the more dissimilar are the pairs being compared. Admittedly the magnitude of the differences between the two sets for any subsample is very small, however, they are comparable. To start with, 1 % and 2 % subsamples are the smallest subsamples in our data, and thus show very large differences compared to the subsequent subsamples.

The MAEs for different subsamples decreases as we increase to larger subsamples. Adding to that, this is in agreement with the reported p-values of the permutation tests. Thus our hypothesis is proven to be true under the MAEs and the permutation tests.

It is also evident that at 20 % subsample the results stay the same or are quite similar to the original data. As the MAEs reported in Figure 4.12 show a constant pattern they suggest that a subsample of 20 % and above is convenient and better to approximate the consistent pattern since these MAEs flatten here. However, it is larger subsamples that are more similar to the original data.

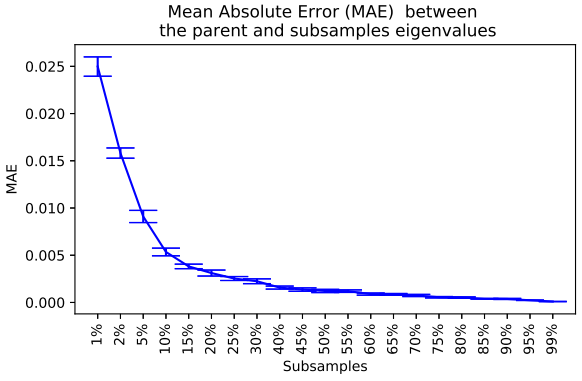


FIGURE 4.12: MAE between uniform and parent data

The results from the subsamples were also subjected to classification. To facilitate this we used two classifiers, LDA and SVM, on the row weights of the resulting quantification results and we computed accuracy, recall and precision for both classifiers. The results of accuracy for both classifiers are reported in Figure 4.13 and Figure 4.14, respectively. The results are reported with error bars as there were some fluctuations in our results caused by subsampling.

The results are also in agreement with our MAEs and permutation test p-values, that is to say, they show consistency in classifying the row weights. Figure 4.13 and Figure 4.14 shows that accuracy of LDA and that of SVM in all subsamples lie between 68 % and 77.8 %, this is in agreement with the ground truth in section 4.3, where the accuracy of LDA was reported to lie between 73 % and 79 % and the accuracy of SVM lie between 71 % and 79 % at 95 % confidence interval. However, classifiers do not perform well on smaller subsamples such as the 1 % and 2 % and this is thus expected since the data at that level has less information. The results of classification improve on subsequent subsamples.

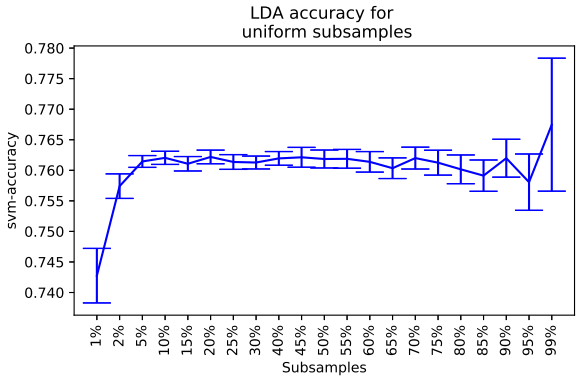


FIGURE 4.13: LDA accuracy results for uniform subsampled data

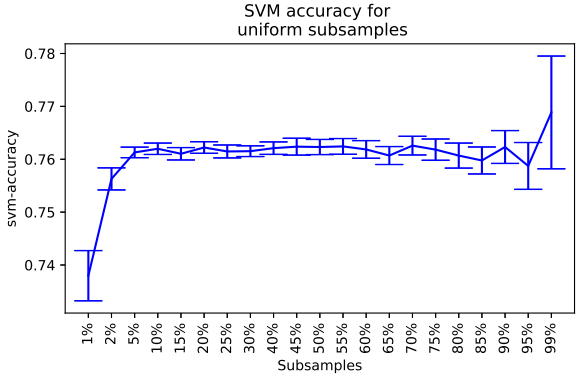


FIGURE 4.14: SVM accuracy results for uniform subsampled data

4.5 Experiment 2: Biased subsampling vs original data

The purpose of this section is to further test the hypothesis statement which states that *"There is no significant difference between the results of the parent and the results of the subsampled data"*. So for this particular experiment by subsampled data we are referring to the biased subsampled data. In addition to this explanation, this experiment seeks to investigate whether the exclusion or inclusion of some categories affects the overall results and interpretation of dual scaling results.

The data was subsampled with some bias. Several types of bias were considered including extremely biased subsampling and some stratified biased subsampling where we only sampled data with a certain distribution of data points, i.e, with a certain percentage of each category. Section 4.5.1 contains the reports on the visual inspection of the results for both stratified and extremely biased subsamples, and finally, section 4.5.2 is primarily focused on the comparisons of results against the ground truth and the evaluation of stability using classification.

In this experiment, we considered "Gender" as the variable of interest, as mentioned earlier on, the rest of the analysis looks at some stratification which is another form of biased subsampling. Furthermore, in the case of using other dichotomous target variables, the results would still lead to the same conclusions [4].

A brief description of the samples being compared:

- The two extremes are as a result of sampling only data from the first or the second category, where the first category consists of 35 % of rows and the second consists of 65 % rows. The first category, corresponds to the **males** while the second corresponds to **females**.
- The subsequent samples correspond to the weighted average of two categories, for instance, 1 % males against 99 % females, as we increase the percentage of one category we decrease the other category and the overall sample size shrinks. For instance, considering Figure 4.15, we notice the exact number of rows for each category, where males constitute of 6231 rows and females make up 9300 rows. This would mean that 1 % of 6231 is combined with 99 % of 9300 rows. Furthermore, to remove bias this experiment was repeated several times since the rows are sampled randomly.

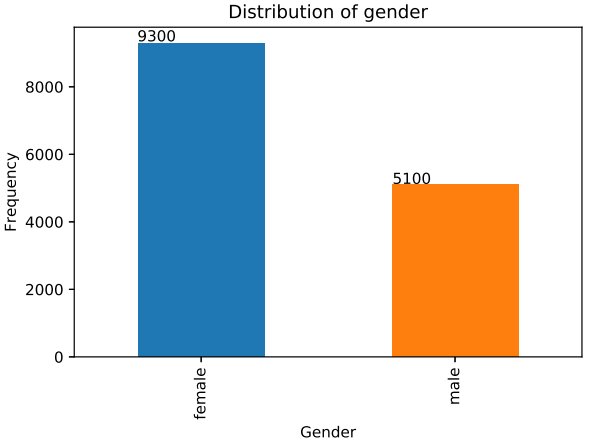


FIGURE 4.15: The distribution of the rows for each categories

4.5.1 Visual Inspection

Figure 4.16 shows the distribution of eigenvalues for the different types of bias introduced in our subsampling. The Scree plots change with each type of bias in the data. To draw insights about these changes, we first look closely at the extremely biased datasets in Figure 4.17 compared with the parent sample.

[4] See the investigation carried on "Employment" and "Income" variables in Appendix A

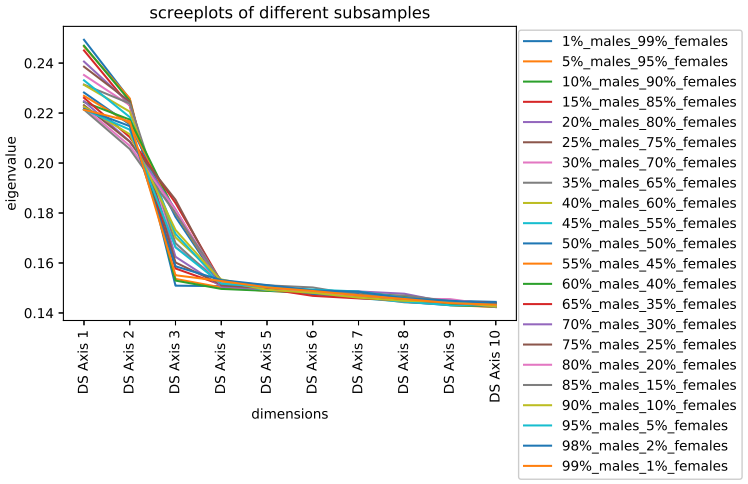


FIGURE 4.16: Scree plots of biased subsamples

Considering Figure 4.17, where we compared the scree plots of extreme cases, we notice that there are differences between the scree plots of eigenvalues for the parent sample and the extreme cases. The data biased by the second category seems to be more different from the original data compared to the case of the first category. However the difference is only visible in terms of a structure of the one-mode plot, but the combinations thereof of categories does not vary much significantly. This is also a positive aspect in terms of the stability of results.

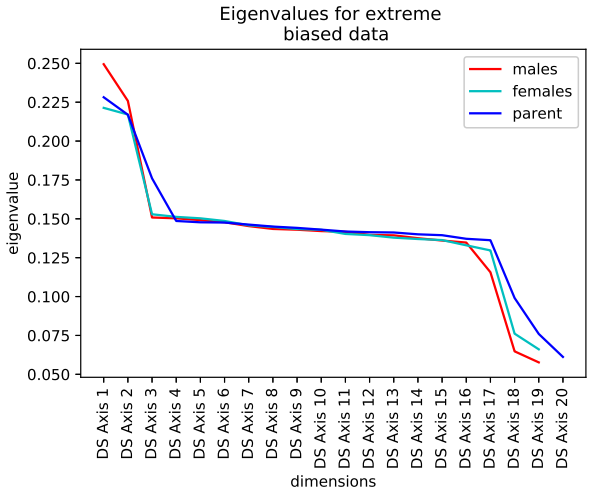
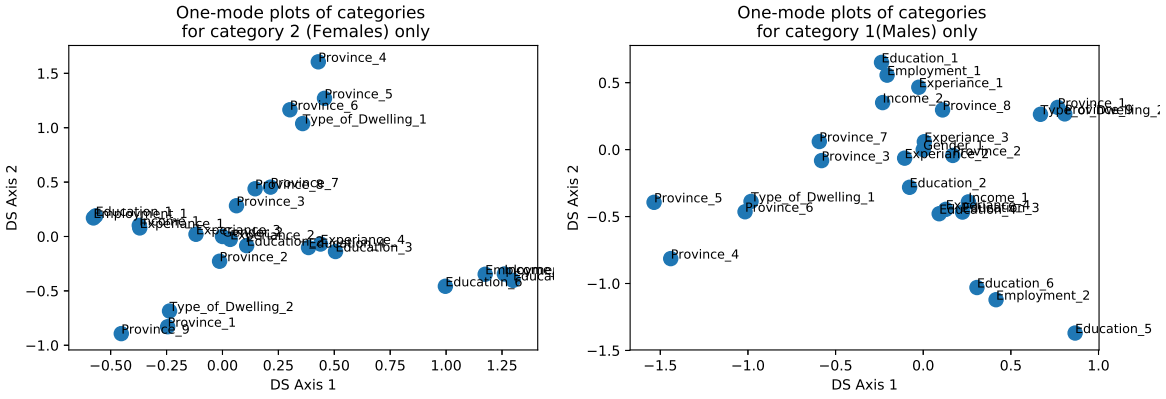


FIGURE 4.17: Scree plots of extreme biased cases vs parent

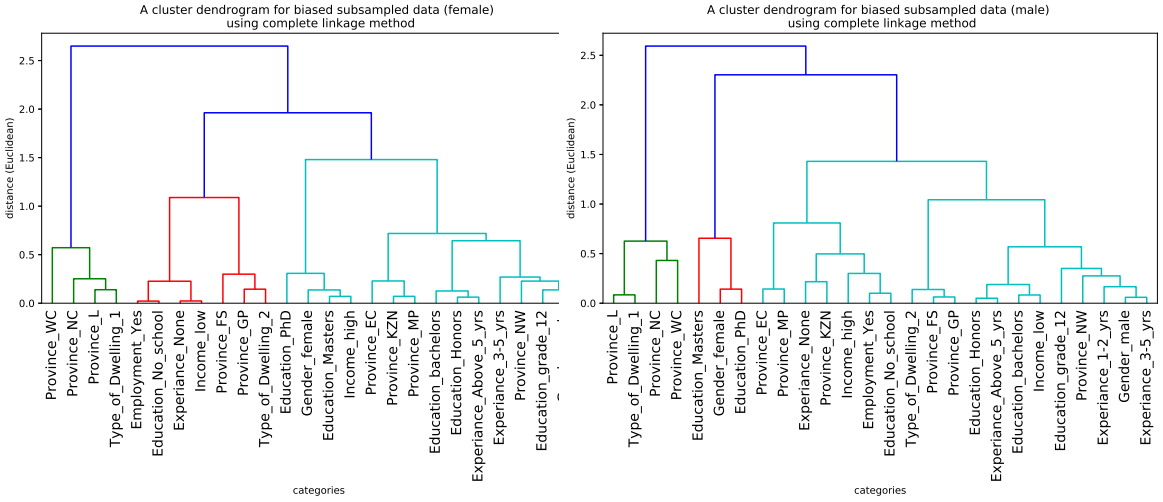
An example of a combination seen in Figure 4.18a, is that of province 4, 5 and 6 which are strongly associated with the second type of dwelling, that is "formal". Furthermore, this pattern is still maintained in Figure 4.18b which is associated with the "male" category. Take into consideration that "males" makes up 35 % of the data while "females" makes up 65 % of the full data. This is investigated later on to determine whether the sample size is the key role player in the stability of results or the stable results are attained by other properties of the data such as a strong correlation between the categories.



(A) One-mode plot for females (B) One-mode plots for males

FIGURE 4.18: One-mode plots for the categories

To further investigate the effect of biased subsampling in the results of quantification the Hierarchical clustering was used to cluster the categories of the biased subsampled data. The results are reported in Figure 4.19a and Figure 4.19b. Both Figure 4.19a and Figure 4.19b have 3 clusters, however, the dendrogram associated with quantified data of females only shows a great similarity to that of the original parent data in Figure 4.4. The second cluster for male category dendrogram consists of fewer categories.



(A) Hierarchical clustering for females (B) Hierarchical clustering for males

FIGURE 4.19: Hierarchical clustering for male and female categories

4.5.2 Comparisons

In this experiment, we were testing the hypothesis which states that *"the quantification results stay the same as the data is biased, that is, excluding a certain category of a variable affects the results"*. The are various issues one can look at in this scenario. One of them would be whether the variable being studied is an influential variable but that is a separate issue altogether. On one of the experiments as far as biased subsampling is concerned we considered "Gender" as the variable of interest, as mentioned earlier. The rest of the analysis looks at some stratification which is another form of bias.

We first look at the permutation tests p-values for both cases. Figure 4.20 shows that the p-value for the female category is higher than that of the male category which suggests that when we perform biased subsampling on the data the category with many rows is likely to produce results that are similar to the parent data. In other words, the sample size plays a pivotal role in biased subsampling and when this kind is applied in practice the bias must be kept minimal.

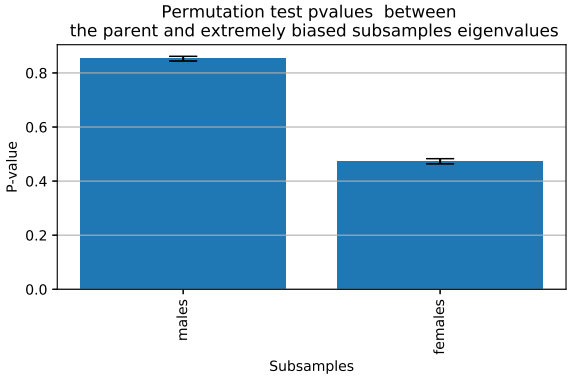


FIGURE 4.20: Permutation test p-values for extreme biased subsampled data

Secondly, considering Figure 4.21, we notice that the p-values are much higher when the data preserves the properties of the parent data. What we mean is that when the data mimic the same distribution of the parent data, then the results also mimics the results of the parent data. We get the evidence of this claim in Figure 4.21, wherein subsamples have 50 % females and 50 % males. This is then just halving of the data in the distribution of the parent data is still maintained. The highest p-value presented in Figure 4.21 is associated with 50 % females and 50 % of males. This subsample hereafter will be referred to as a "50/50" subsample.

We then looked at the Mean Absolute Errors (MAE) between the extremely biased datasets and the parent dataset. The interpretation of MAE is that, the smaller the MAE the closer the biased subsampling results to the results of the parent data. For the extremely biased subsampling data with male category has a lower MAE compared with

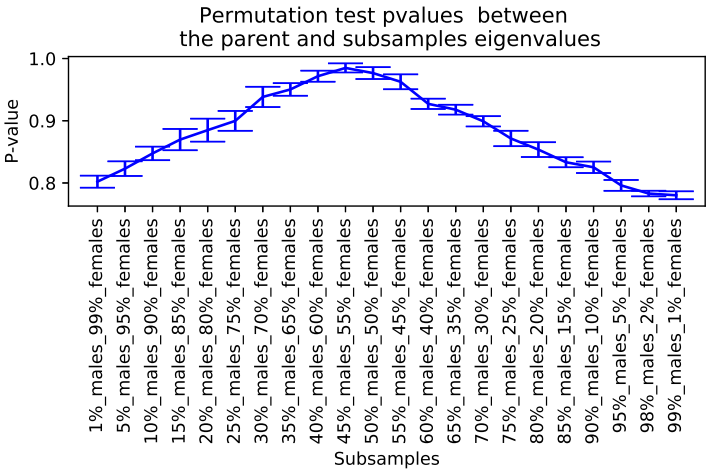


FIGURE 4.21: Permutation test p-values for biased subsampled data

the female category as reported in Figure 4.22. This simply means that the MAE, in this case, agrees with the permutation p-values. The agreement suggests that the category with fewer rows is more likely to yield results similar to the results of the parent data. Contrary to what we would suggest with uniform subsampling, the smaller rows the better.

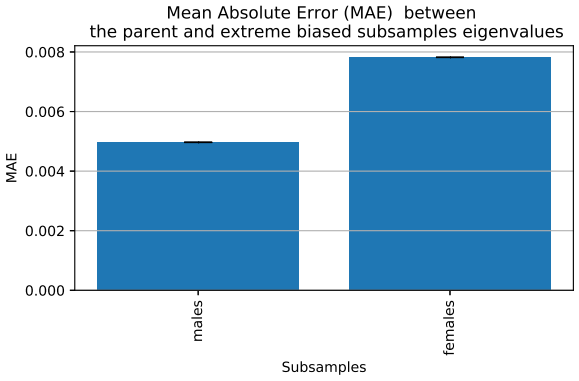


FIGURE 4.22: MAE for extreme biased subsampled data

Finally, we investigated the stability of DS on stratified biased subsamples through the use of LDA and SVM. The results of both LDA and SVM accuracy are reported in Figure 4.24 and Figure 4.25, respectively. These results show that the imbalance in the categories affects the classification accuracy results which are much higher when there is an equal distribution of categories and the results decline immediately when there is an imbalance or imbalance due to sampling. However, the classification results still show some consistency in that the accuracy lies between 75.8 % and 77.5 % on average throughout biased subsampling. This means that we can still make some predictions and it does not matter whether the categories are imbalanced or not. Also, this means

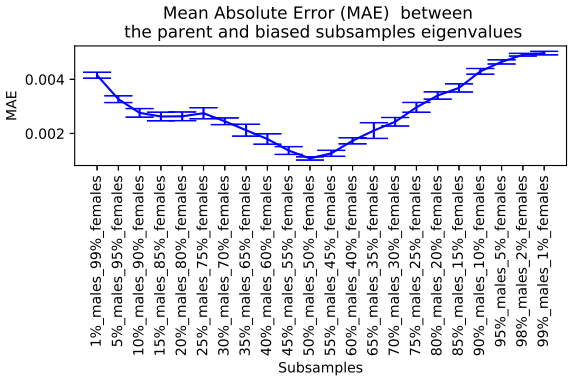


FIGURE 4.23: MAE for biased subsampled data

that the classification of DS is not easily influenced by a particular category, which is a good thing.

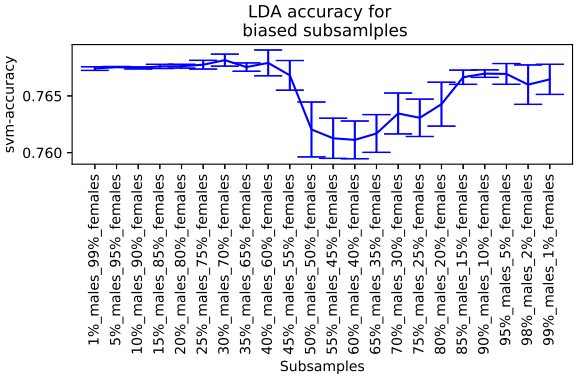


FIGURE 4.24: LDA accuracy of biased datasets

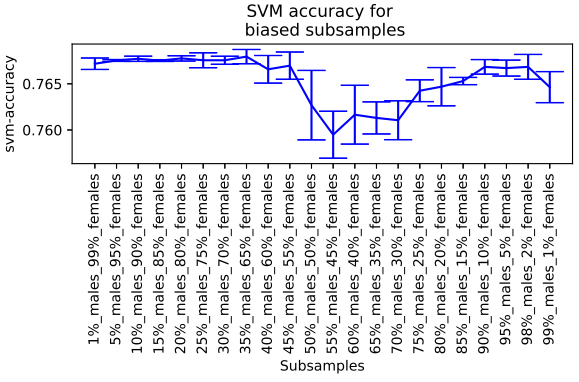


FIGURE 4.25: SVM accuracy of biased datasets

Similar experiments were undertaken using another dichotomous variable on the same dataset and the results are reported in Appendix B. The results of using the other dichotomous variable also lead to the same conclusions which shall be discussed later in this section. By dichotomous variable, we refer to a variable with exactly two categories

such as that of Gender. Thus the experiments were tested on "Income" and "Employment" variables. In addition to these supplementary experiments, we also tested the hypothesis using the 2017 General Household Survey (GHS) data and the results are reported in Appendix B. The results also suggest that the 50/50 subsample yields better results and there is more stability on this data due to its distribution.

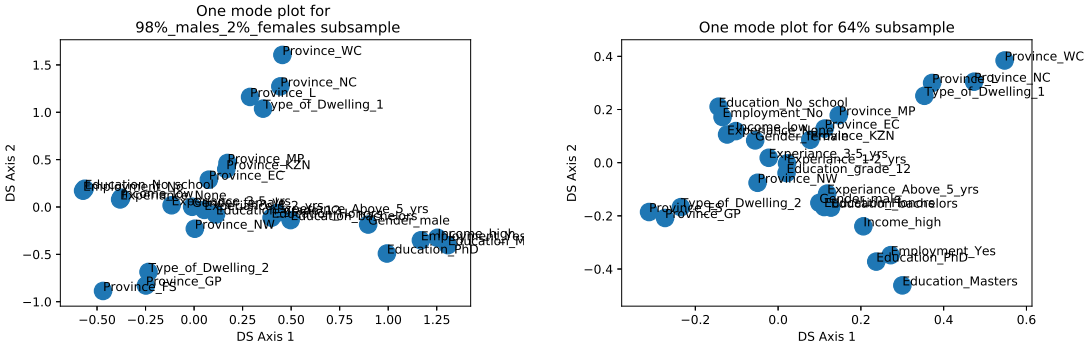
4.6 Experiment 3: Uniform subsampling vs biased subsampling

To investigate the effect of subsampling uniformly over biased subsampling, we compared the results of the two subsampling strategies, biased and uniform subsampling, respectively under similar conditions. To get the resulting difference between the two strategies we compared the conditions that yield the data with same sample size, for example, the extreme bias with the first category yields 35 % of the original data. We then compared this directly against when subsampling 35 % of the data uniformly. The results are reported in section 4.6.1 and section 4.6.2.

4.6.1 Visual Inspection

We first investigate if there are similarities between the structure of the one-mode plot for data subsampled uniformly and that which is subsampled with bias. The two scenarios that are reported in Figure 4.26a and Figure 4.26b and have one thing in common, that is, sample sizes and the only difference is the type of sampling. To investigate the effect of sample size and sampling strategy on the dual scaling results we randomly selected some cases with similar sample size and compared them critically.

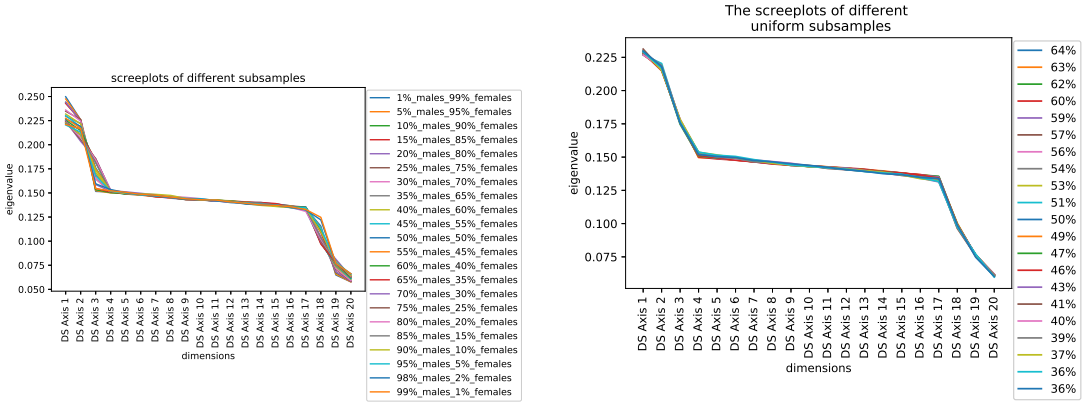
Considering Figure 4.26 it is quite clear that, although structurally, the one-mode plots for both biased and uniform subsample are different from the X-like structure in Figure 4.1b, they still show similar groupings as far as the categories are associated. This means that even if geometrically the categories are placed in different positions, the relationship between the categories is still maintained and therefore does not change the interpretation of DS results (Nishisato and Clavel, 2009). Greenacre and Hastie (1987) affirmed that when interpreting the plots of Correspondence Analysis (CA), which is analogous to DS, the groupings (within distance) are more valuable than the between distance, meaning, even if the coordinates are rotated what is important is the groupings (Nishisato and Clavel, 2009).



(A) One-mode plots for 64 % biased subsample (B) One-mode plots for 64 % uniform subsample

FIGURE 4.26: One-mode plots for different sampling strategies

Considering Figure 4.27, it may seem that the distribution of the eigenvalues of biased subsampled deviates from the original structure of the parent set but the difference does not affect the interpretation of results widely. This implies that uniform subsampling retains more important information than the biased subsampling strategy, thus with biased subsamples there exist more discrepancy, especially for influential [5] categories in the data. Further analyses are considered in Section 4.6.2.



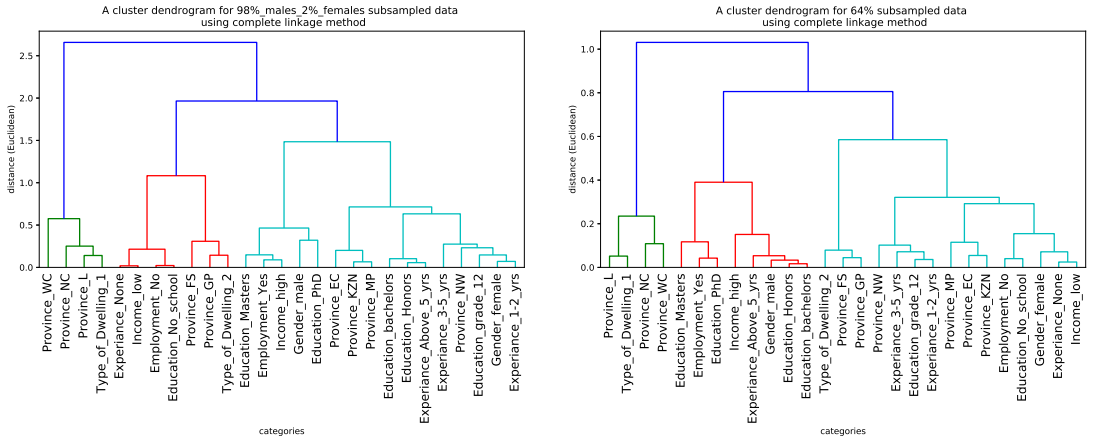
(A) Scree plot of eigenvalues for biased subsampled data (B) Scree plot of eigenvalues for uniform subsampled data

FIGURE 4.27: Scree plots for different sampling strategies

Further investigations on these two subsampling strategies were undertaken by utilization of hierarchical clustering algorithm on the quantified data of both uniform and biased subsampling. The results are reported in Figure 4.28a and Figure 4.28b. The results show that the hierarchical clustering of the uniform subsampled data in Figure 4.28b retains the number of clusters and also, the 3 clusters are all made up of categories

[5] An influential observation or variable, is one in which its inclusion or deletion from the dataset would cause a significant change in the results (Jolliffe, 2002).

that appear in the clustering of a parent data in Figure 4.4. However, Figure 4.28a deviates from the ground truth in that, although the number of clusters formed is 3 and the first cluster contains all the categories similar to the results in Figure 4.4, the remaining two clusters differs slightly and therefore in this case the uniform subsampling seems to be reliable to reproduce the results of the parent data.



(A) Clustering results for biased subsampled data (B) Clustering results for uniform subsampled data

FIGURE 4.28: Clustering results for different sampling strategies

4.6.2 Comparisons

The dataset was subsampled in different ways with a bias, and this leads to subsamples of unequal sample sizes. It is with this regard that we investigated the effect of biased subsampling on dual scaling results as opposed to the uniform subsampling. Consider Figure 4.29 and Figure 4.30. In these figures, we compared the biased subsampling against the uniform subsampling and the metric used here is MAE. Consider the MAE for biased subsampling in Figure 4.29. The highest MAE is associated with subsamples that have imbalanced categories and in which case, the data that has more females than males deviates largely from the parent set. More females here implies that the data has more rows of females either due to subsampling or it may be just the distribution of the data. On the other hand, in Figure 4.30, the highest MAE is associated with the smallest uniform subsample. This shows that one may desire the uniform subsampling over biased subsampling in this case. The highest subsample in Figure 4.30 has the smallest MAE which brings us to the conclusion that the more data points we have if subsampled uniformly then better results are guaranteed.

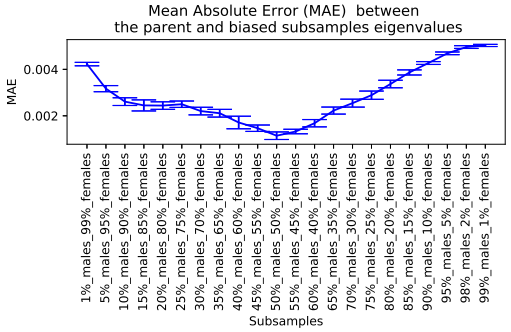


FIGURE 4.29: MAEs for biased subsampled data

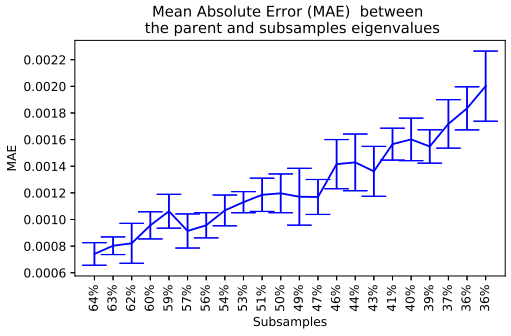


FIGURE 4.30: MAEs for uniform subsampled data

We also looked at permutation test p-values and the results are reported in Figure 4.31 and Figure 4.32. The p-values show much variation when the data is skewed to one category and when some extent of balance is attained the p-values immediately increase showing that the subsampled data yields similar results to the parent data. In this case, the 50/50 sample still yields better results than other subsamples.

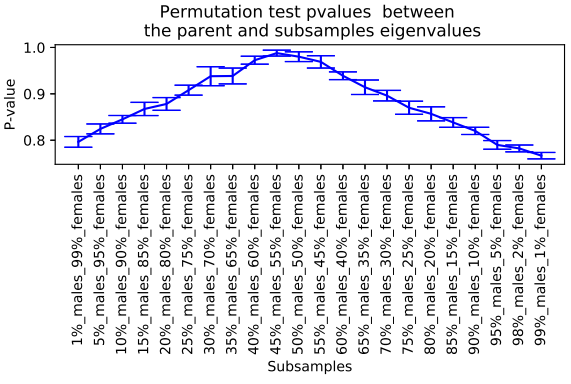


FIGURE 4.31: Permutation test p-values for biased subsamples

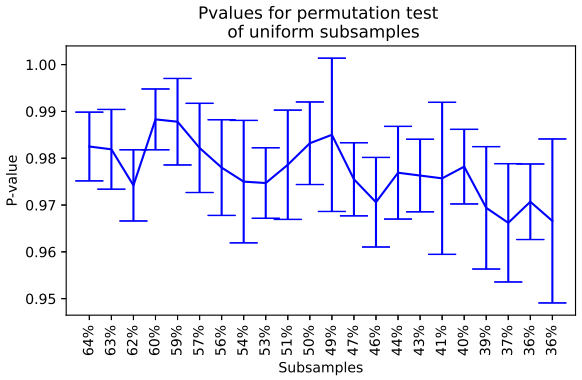


FIGURE 4.32: Permutation tests p-values for uniform subsampled data

Finally, we considered the classification of row weights using two classifiers: SVM and LDA. The experiments were conducted for both biased and uniform subsampled data and the results for biased data are reported in Figure 4.33 and Figure 4.34, while the results for uniform subsampled data are reported in Figure 4.35 and Figure 4.36.

The LDA and SVM accuracy results for uniform subsampled data in Figure 4.35 and Figure 4.36 shows that accuracy increases with an increase in the sample size but then the results between 75 % and 77 %. These are in agreement with the results presented in section 4.3

On the other hand, the results for biased subsampled data as reported in Figure 4.33 and Figure 4.34 shows that the larger sample size does not necessarily yield better accuracy. Furthermore, the results of accuracy for the biased subsamples lie between 75 % and 79 % for both LDA and SVM, with the highest accuracy being that of the 50/50 subsample. This interval is in agreement with the ground truth presented in section 4.3.

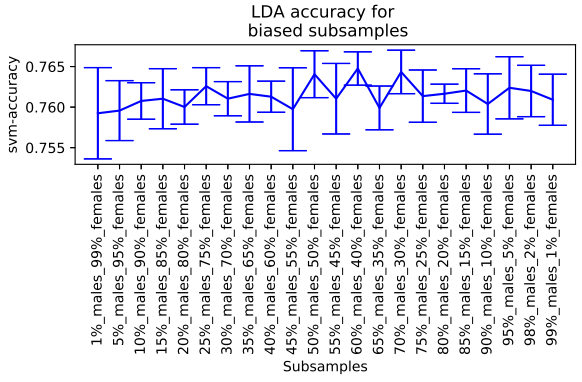


FIGURE 4.33: LDA results for biased subsampled data

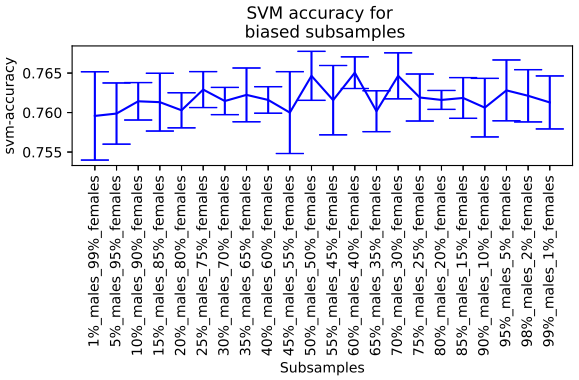


FIGURE 4.34: SVM results for biased subsampled data

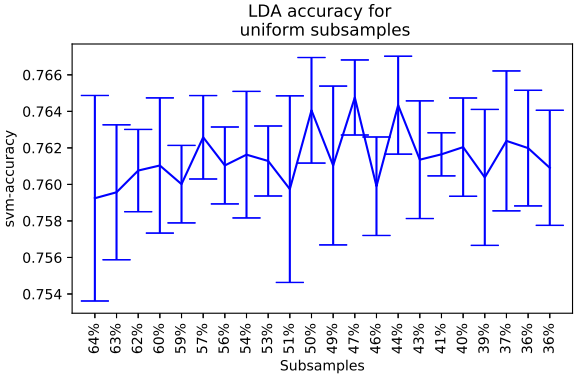


FIGURE 4.35: LDA results for uniform subsampled

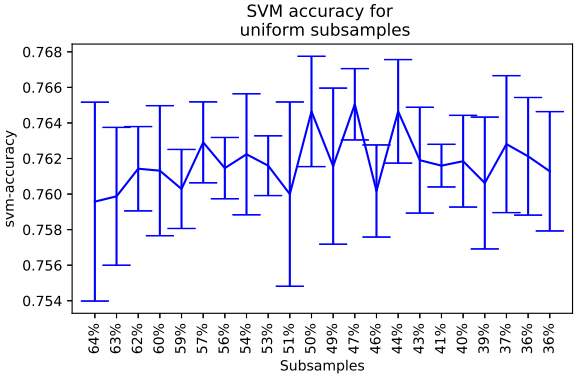


FIGURE 4.36: SVM results for uniform subsampled

Altogether, what we learn is that the data that is distributed in the same way as the parent data yields the same results that the parent data would yield when not subsampled. In other words, stratified random sampling[6] is rather better than extreme subsampling.

[6]Stratified random sampling is a type of sampling which ensures that the sampled data is in proportion to the population, that is, the data is divided into subgroups called stratum in an original percentage as in the population, but with lesser sample size (Mondal and Mandal, 2020)

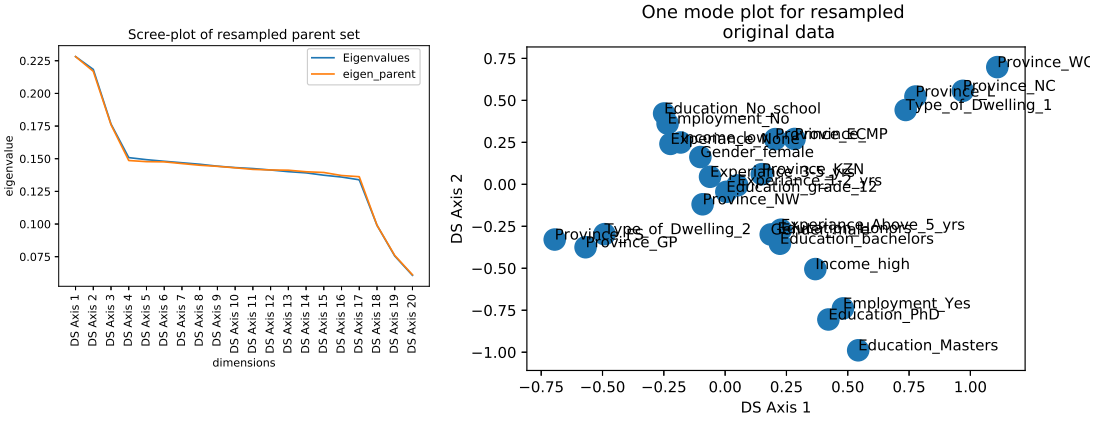
Subsampling, in general, leads to loss of information, but extreme may suggest a greater loss. Finally, we also learn that given a variable with two categories of unequal rows, the one with fewer rows is more likely to yield similar results to the parent data.

4.7 Experiment 4: Resampling vs parent sample

In this experiment, we tested the hypothesis that states that *"The quantification results stays the same as we resample the parent data"*. Resampling was performed on the original data and the average quantification results were retained for all the resamples. The results are reported in Section 4.7.1 and Section 4.7.2.

4.7.1 Visual Inspection

Consider Figure 4.37a where we roughly looked at the structure of eigenvalues for the parent data and the resampled parent data. There the structure of the eigenvalues for resampled data is similar to the structure of the parent sample although there are small discrepancies which affected by the number of replications in each resampling experiment. The higher the replication size the more similar the results to the parent sample, where higher can be any size from 10 to over 200. To ascertain our findings we also looked at the one-mode plot of the resampled parent data in Figure 4.37b. We noticed that the X-like structure of column weights that were obtained in the parent set in Figure 4.1b is reconstructed and is similar to the structure of the parent data. This is a very significant finding since this show that there is a consistency in dual scaling, such that the assignment of numerical values to the categories leads to the same interpretation.



(A) Scree plot of the resamples (B) One-mode plots for the average column weights

FIGURE 4.37: A scree plot and one-mode plot for the resampled parent sample

The resampled parent data was also subjected to hierarchical clustering in order to further investigate the stability of results under resampling. The results of clustering are reported in Figure 4.38. The dendrogram has 3 clusters and these clusters are exactly similar to the clusters formed by the original parent data in Figure 4.4. This shows that the mixing of rows does not necessarily affect quantification results negatively.

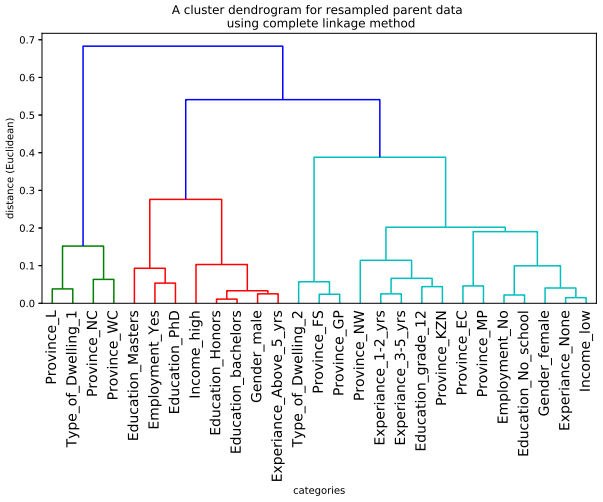


FIGURE 4.38: Clustering results for resampled parent data

4.7.2 Comparisons

We then used our quantified data to test whether the classification models may arrive at a similar conclusion and also to check whether these results can be used for inference. We trained the resampled datasets and tested on the parent data as the unseen data.

Table 4.4 gives a summary of results for experiment 4. Considering the permutation test p-value for resampled uniform data, the value of 0.9999 is high enough to depict that the resampled parent data yields similar results to the parent data. Furthermore, the MAE value (0.00001) in Table 4.4 is very small and close to zero, thus also confirming the level of similarity between the *resampled parent data* and parent data.

TABLE 4.4: Summary of results for resampled parent data

	P-value	MAE	LDA Accuracy	SVM accuracy
Resampled Parent Data	0.999	0.000001	0.766204	0.766574

Finally, the accuracy for LDA was about 76 % which is the same for SVM as indicated in Table 4.4. This is a very good indication that resampling the data does not lead to

a large loss of information indicating that resampling methods can be trusted as a tool for ascertaining the stability of results. Thus the dual scaling results can be trusted.

The same experiment was tested on the 2017 GHS data and the results are reported in Table B.5 of Appendix B. Although the classification accuracy for GHS data may have high values of accuracy, these values are maintained from the parent data to the resampled parent data.

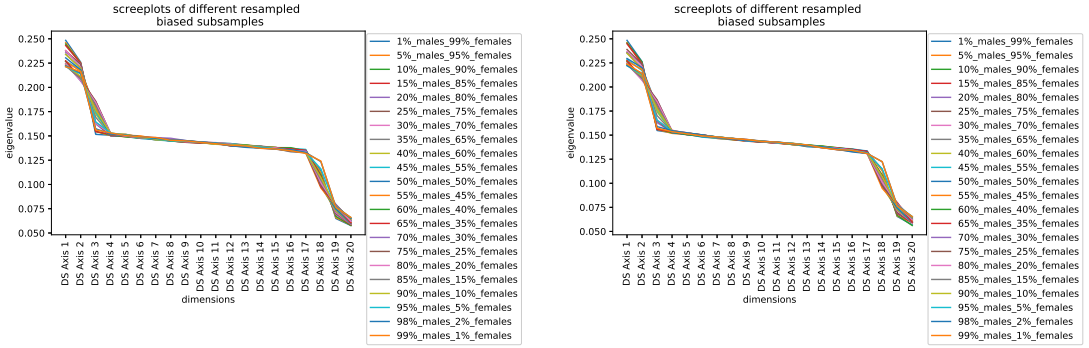
4.8 Experiment 5: Resampling vs Biased Subsampling

As opposed to resampling the original data, we further investigated the effect of resampling the subsamples of the data. In this experiment, we considered resampling with replacement and thus recovering the sample size of the parent, which was 14400 rows. In this experiment we looked at sampling the biased subsamples with replacement and recovered the sample size of the parent sample. For instance, we had several biased subsamples such as 50/50 subsample, where the rows of males to females were evenly distributed and represented half of the parent data. We then expanded this data by bootstrapping (sampling with replacement) and the resampled datasets were larger than the root data where we resampled from. By doing this, the aim was to investigate whether the results of quantification will remain the same or change with an increase in the sample size. The results are reported in Section 4.8.1 and Section 4.8.2. [7]

4.8.1 Visual Inspection

Consider Figure 4.39a and Figure 4.39b which present the Scree plots of the biased subsamples and the resampled biased subsamples. The results of the eigenvalues are not different from that of resampled biased subsamples and thus we conclude that resampling the biased subsamples with resamples of equal size to the 14400 rows of the parent sample do not reconstruct the quantification results of the parent samples but the results of the biased subsamples are reconstructed rather. In other words, it does not matter how much the data was resampled and to which sample size was it resampled, the results remain similar to the general results of that subsample or more specifically the raw data before. Looking at the Scree plots for the biased subsamples and resampled biased subsamples in Figure 4.39a and Figure 4.39b, the results confirm that resampling the biased subsample recovers the structure of the parent sample.

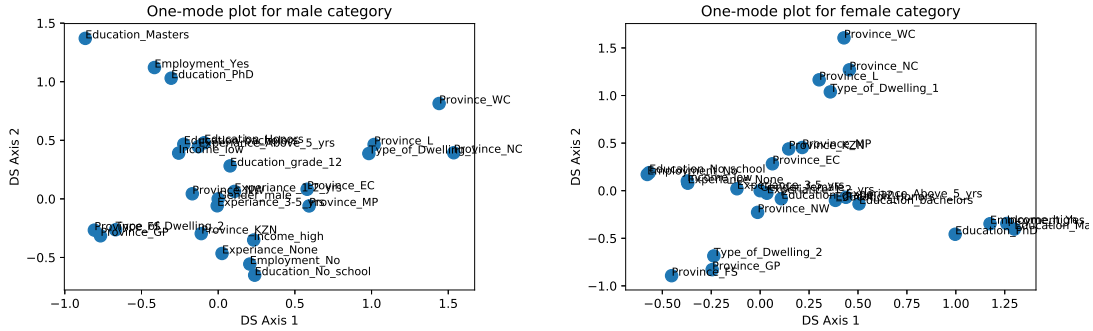
[7] A similar experiment was performed on 2017 GHS data and the findings suggest that the results still hold even on a real dataset. The results are reported in Appendix B



(A) Scree plots of biased samples (B) Scree plots of resampled biased samples

FIGURE 4.39: Scree plots of eigenvalues for the biased subsamples and the resampled biased subsamples

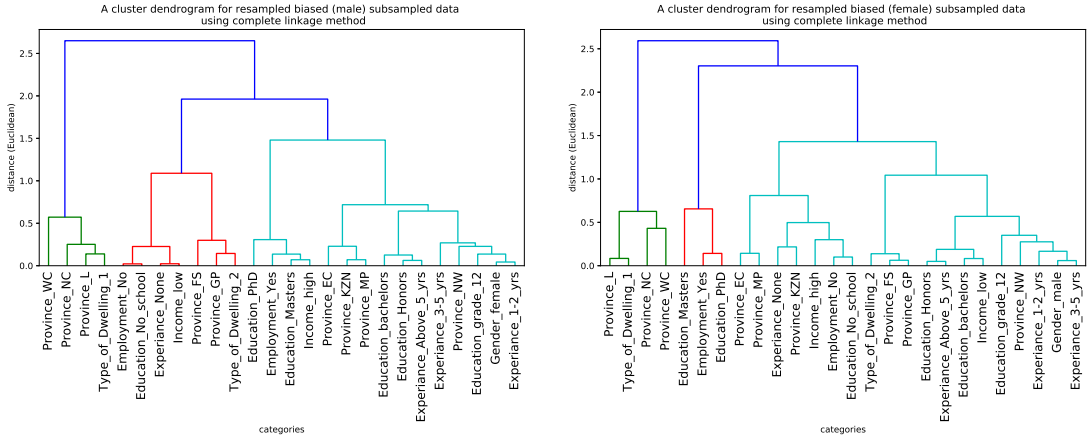
We also looked at the one-mode plots for the two extreme cases in Figure 4.40a and Figure 4.40b and the results confirm that the resampled biased subsamples reconstruct the structure of the original data.



(A) One-mode plots for resampled biased subsamples for males category (B) One-mode plots for resampled biased subsamples for females category

FIGURE 4.40: One-mode plots for resampled extreme categories

Hierarchical clustering was also applied on the resampled biased subsampled data and the results are reported in Figure 4.41a and Figure 4.41b. In both case of Figure 4.41a and Figure 4.41b, the hierarchical clustering shows that there are 3 clusters formed by the categories. However, the categories formed through resampling are more similar to the results obtained for the normal (without resampling) biased subsampling in Figure 4.19a and Figure 4.19b.



(A) Clustering results for resampled biased subsamples for males category

(B) Clustering results for resampled biased subsamples for females category

FIGURE 4.41: Hierarchical clustering results for resampled extreme categories

4.8.2 Comparisons

The biased subsamples were then resampled several times. This type of resampling followed the same procedure as that of resampling the uniform subsamples in that the sizes of the resamples were equal to the size of the parent sample. This is reported in Section 4.9. Dual scaling was then applied to these resampled datasets and the average DS results were computed. To evaluate the effect of resampling, the resampled biased subsamples results were compared with the parent sample.

The first comparison was based on the MAE between the eigenvalues of the resampled biased subsamples and the parent sample. Consider Figure 4.42 where the MAEs for the subsamples containing evenly distributed in terms of categories shows smaller MAE as compared to the skewed datasets, that is when rows of categories are not equal. Furthermore, the subsamples which contain more females than males had a slightly larger MAEs as compared to those that had fewer females than males. This is in agreement with the results shown in Section 4.6.

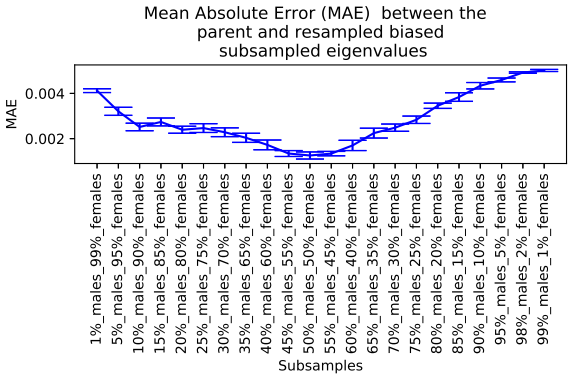


FIGURE 4.42: MAEs for eigenvalues of resampled biased subsamples

We also looked closely at the resampled extremely biased subsamples and the results are reported in Figure 4.43. The subsample with rows containing only males has a smaller MAE as compared to that of subsample containing only females. This result confirms that the male category has more properties similar to the parent data than the female-biased dataset, also this signifies that the overall larger sample sizes when the data is subsampled using biased subsampling do not necessarily yield better results.

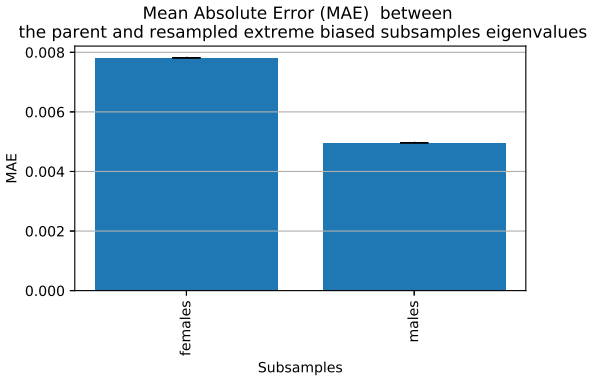


FIGURE 4.43: MAEs for eigenvalues of extremely biased data

The permutation tests were also considered for the resampled biased subsamples and the results are reported in Figure 4.44 and Figure 4.45. Considering Figure 4.44 the p-values are much higher for evenly distributed data and decline as the data is biased towards one category. Moreover, in Figure 4.45 we learn that the p-value is higher for the male category. This is in agreement with the results of the MAE.

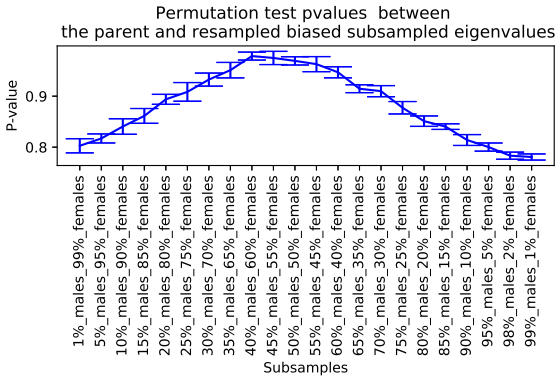


FIGURE 4.44: Permutation test p-values for resampled biased subsampling

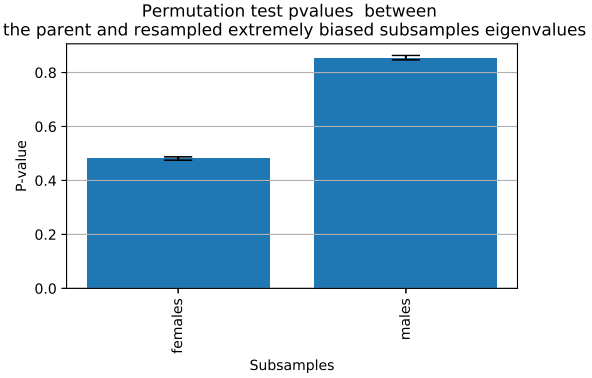


FIGURE 4.45: Permutation test p-values for resampled extremely biased subsampling

Finally, we considered the classification using row weights where the target variable was "Type of Dwelling" variable. The results of two classification methods (LDA and SVM) are reported in Figure 4.46 and Figure 4.47.

Considering the classification results for resampled biased data, we notice that the LDA and SVM results are similar to the results obtained in experiment 2 in Section 4.5, that is, the accuracy is higher for subsamples containing more males than that which has more females. Furthermore, considering Table 4.5 we notice that the accuracy and precision values are much higher for the data containing only the male category. Finally, the classification results for stratified biased subsampling show stability of DS in that accuracy still ranges between 75 % and 77 %. Lastly, the classification of extremely biased data is not meaningful, but looking at Table 4.5, the evaluation measures for both LDA and SVM are outside the bounds that we have gotten in experiment 4.4.

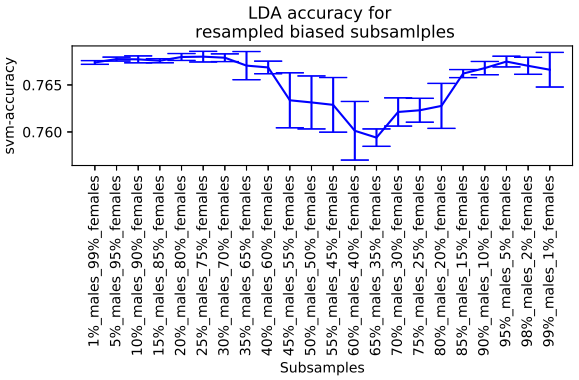


FIGURE 4.46: LDA accuracy for resampled biased data

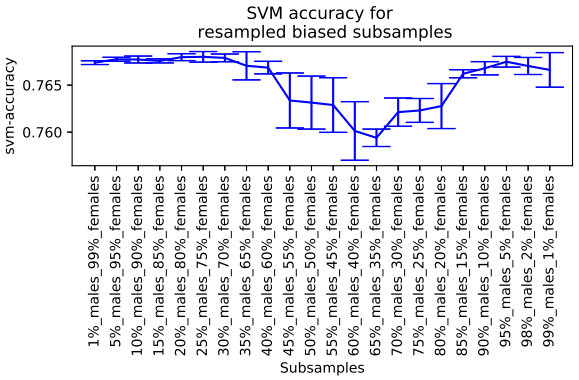


FIGURE 4.47: SVM accuracy for resampled biased data

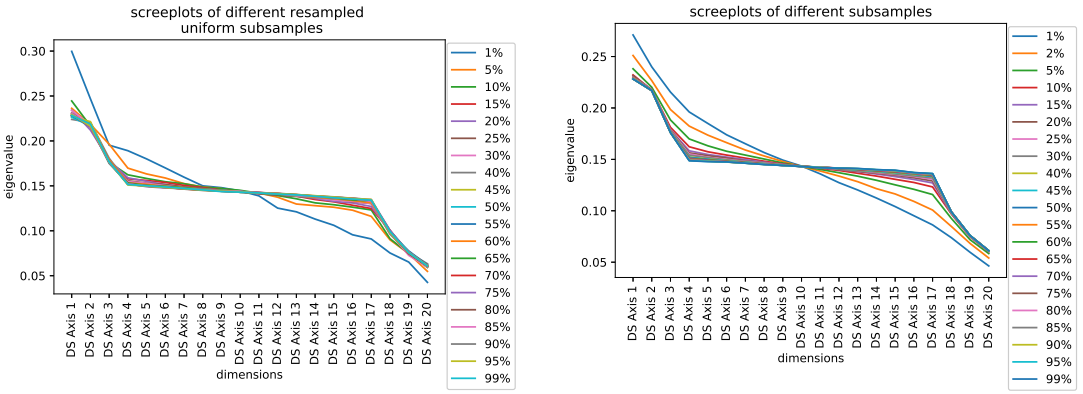
TABLE 4.5: Classification results for resampled extreme biased data

	lda_accuracy	lda_precision	lda_recall	svm_accuracy	svm_precision	svm_recall
males	0.664306	0.575906	0.612075	0.664028	0.575654	0.611207
females	0.493403	0.371754	0.384976	0.492153	0.369854	0.381853

4.9 Experiment 6: Resampling vs uniform Subsampling

We considered resampling uniform subsamples. In this experiment we sampled from the different subsamples used previously, that is 1%, 2 %,..., 99 %. Given a subsample, we sampled from this sample with replacement in a way where the size of the parent subsample is reconstructed as 14400 rows. The results are reported in Section 4.9.1 and Section 4.9.2. The results distribution of eigenvalues reported in Figure 4.48a are similar to the ones in Figure 4.48b.

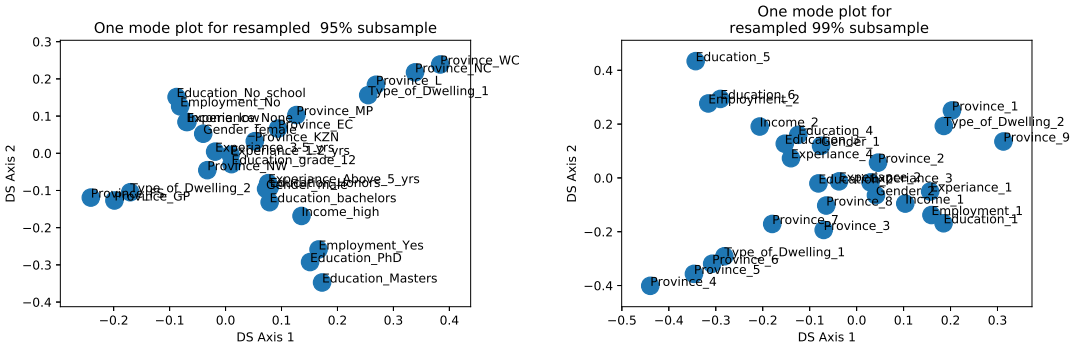
4.9.1 Visual Inspection



(A) Eigenvalues for bootstrapped uniform subsamples (B) Uniform subsampled eigenvalues (non-resampled)

FIGURE 4.48: Scree plot uniform subsamples and resampled uniform subsamples

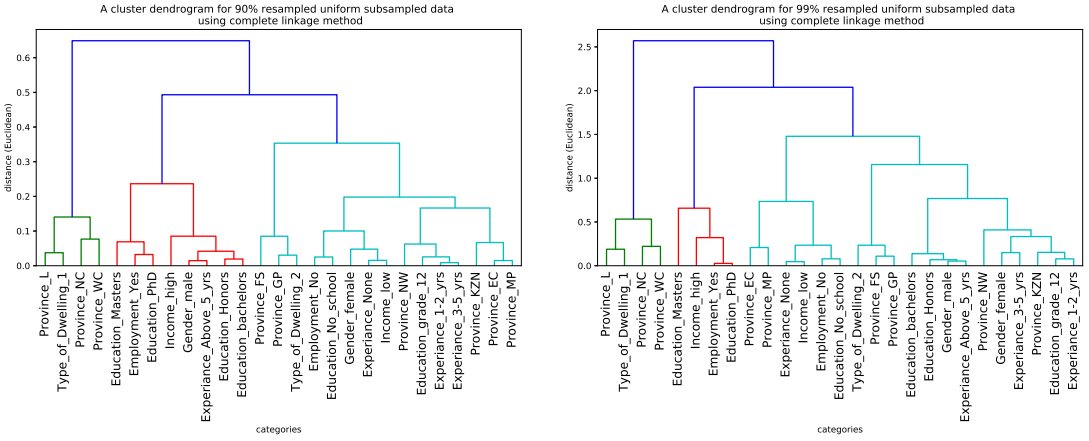
The one-mode plots for the resampled uniform data still gives the same interpretation as that of experiment 4.4 in Section 4.4.1. One-mode plots for 95 % and 99 % uniform subsamples are reported in Figure 4.49a and Figure 4.49b.



(A) One-mode plots for 95% subsample (B) One-mode plots for resampled 99 % subsample

FIGURE 4.49: One-mode plots for resampled 95% and 99 % uniform subsampled data

The hierarchical clustering was also applied to the resampled data to further explore the stability of quantification results under resampling and the results are reported in Figure 4.50a and Figure 4.50b. The 90 % and 99 % resampled uniform subsampled data both grouped data into 3 clusters. These clusters are similar to the results obtained in the uniform subsampled data in Figure 4.10a and Figure 4.10b.



(A) Hierarchical clustering results for resampled 90% subsample (B) Hierarchical clustering results for resampled 99 % subsample

FIGURE 4.50: Hierarchical clustering results for 90% and 99 % uniform subsampled data

4.9.2 Comparisons

The resamples are based on resampling the already existing subsamples from uniform subsampling. This experiment aimed to investigate whether resampling the subsampled data would yield or reconstruct the results of the parent sample. In simpler terms, this experiment seeks to investigate whether we can rely on resampling in cases where we have a small sample size. Will the results of resampled data give better results than subsamples which were not resampled?

We used the MAE to investigate how similar is the structure of eigenvalues for subsampled data as compared with a resampled uniform subsample. The results of the MAE, are reported in Figure 4.51. The results show a direct proportionality, that is, as we increase the sample sizes of subsampled data, the similarity of results between the parent data and subsampled data also increases. Moreover, the results are more similar to the results yielded in Section 4.4.2 than to the parent data results. Furthermore, the results were also subjected to permutation tests for further investigation. The results of the permutation test presented in Figure 4.52, show some consistency with results obtained in Section 4.4.2. The p-values are higher for larger subsamples and smaller for smaller subsamples and this a clear indication that resampling techniques preserve the properties of the original data, this shall be discussed in Chapter 5.

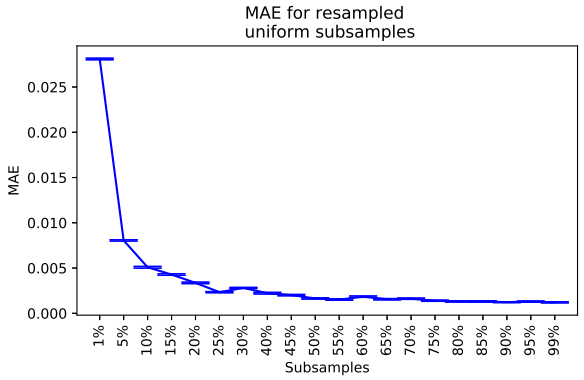


FIGURE 4.51: MAEs for eigenvalues of resampled uniform data

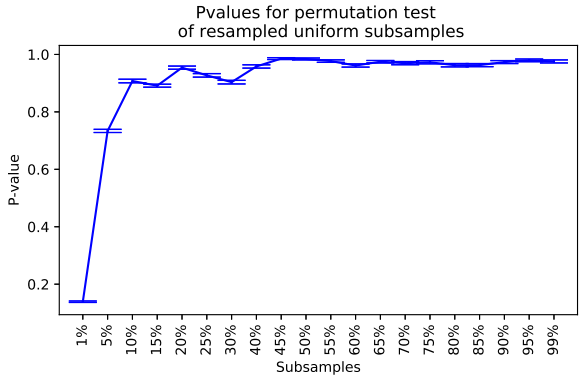


FIGURE 4.52: Permutation test p-values for eigenvalues of resampled uniform data

Finally, we considered the classification of row weights with resampled row weights and the results for LDA and SVM accuracy are reported in Figure 4.53 and Figure 4.54, respectively. The results show many similarities to the results obtained in Section 4.4.2. Although there are some fluctuations in the values of both LDA and SVM accuracy, the results show that the accuracy results are directly proportional to the sample sizes and this is the same observation we noticed with uniform subsampling in Section 4.4.2. Also, the results suggest that the accuracy for both LDA and SVM ranges between 75 % and 79 %.

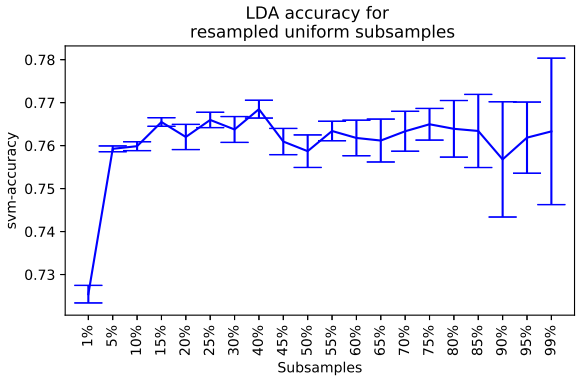


FIGURE 4.53: LDA accuracy for resampled uniform data

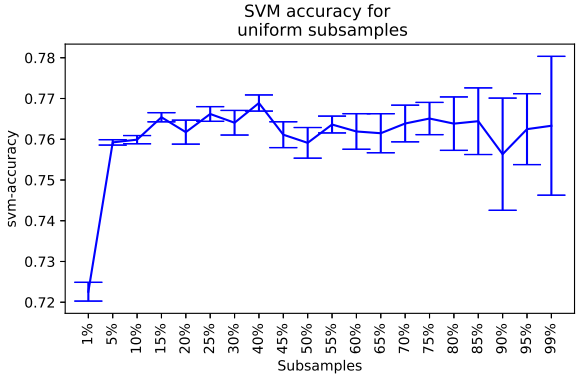


FIGURE 4.54: SVM accuracy for resampled uniform data

4.10 Summary of results

Table 4.6 gives a summary of experiments, their intended hypotheses and the key findings. If used well these findings may guide an interested researcher on when to use subsampling and which subsampling should be used given the objectives of the project or study. For instance, experiment 6, resampling of uniform subsampled data may refer to just a sample that is publicly available for the researcher. Now if the given sample violates the sample size assumptions, then bootstrapping may be applied to such data, however, this does not guarantee that the results will be the same as those of the population data. More discussions on the application of the results can be found in Chapter 5.

TABLE 4.6: Summary of Experimentation results

Experiment	Description	Findings
Experiment 1	Test whether there is a significant difference between the uniform subsampled data and parent data	There is a difference which is determined by the sample size: larger subsamples minimize the difference
Experiment 2	Test whether there is a significant difference between the biased subsampled data and parent data	-There is a difference which is slight bigger than that of the uniform subsampled data. - Evenly distributed (50/50) subsamples resemble the results of the parent data - Extremely biased datasets resemble the results of the parent data better on categories with fewer number of rows
Experiment 3	- Test whether there is a significant difference between the uniform and biased subsampled data. -Which one of these give similar results to those of the parent data	- There is a significant difference which differs in the types of sampling - The results for uniform subsamples are dependent on the sample size - The results of uniform subsampling give a better estimation of the parent data - The results of biased subsampling are independent of the sample size, but work better on stratification (proportionate subgroups)
Experiment 4	Test whether there is a significant difference between the resampled parent data and parent data	- There is a difference, though not significant. - The results give better results that are similar to the parent data - It replaces the efforts of going to the field collecting more data
Experiment 5	Test whether there is a significant difference between the resampled biased subsamples and biased subsamples	- There is a difference, but not significant. - The resampled data mimics the biased set rather than the parent set
Experiment 6	Test whether there is a significant difference between the uniform subsampled data and the resampled uniform data - What is the effect of sample size ?	-There is a difference, although not a significant one - The increase of sample size does not substitute the collection of more data - The resampled data mimics the input data rather than the population data

4.11 Conclusion

In this chapter, we tested two hypotheses on the stability of quantification results. The first hypothesis is based on subsampling and the second is based on resampling. The first three experiments correspond to the first hypothesis which states that there is a significant difference between the quantification results of the subsampled and the original parent data, while the last three experiments are used to test the second hypothesis which states that there is a significant difference between the quantification results of the resampled and the original parent data. Indeed there is a difference between the quantification results of subsampled and the original parent data. The difference is much higher between the results of biased subsampled data and the results of the parent data. However, although there is a difference, it is not significant since the classification results are not significantly different from the ground truth. Thus, for this reason, the quantification method, specifically DS is stable concerning subsampling and resampling. The rest of the implications of the results are discussed in Chapter 5.

Chapter 5

Summary and Discussion

5.1 Introduction

The main focus of this study was to investigate the stability of quantification results concerning subsampling and resampling. In this chapter, we further discuss the results and summarize the conclusions on the hypotheses tested. Although the background of other types of categorical data has been described in the background section, the main attention of this research was based on multiple-choice data.

5.1.1 Chapter Organization

The remainder of this chapter is organised as follows: Section 5.1.2 discusses the effect of subsampling on dual scaling, Section 5.1.3 deals with the results and effect of resampling in quantification, while Section 5.2 is devoted to the discussion of other scholars' findings concerning stability.

5.1.2 Subsampling

The first three experiments in Chapter 4 are associated with subsampling the parent data in different ways, namely: uniform (see Section 4.4) and biased subsampling (see Section 4.5), respectively.

In experiment 1 (see Section 4.4) of Chapter 1, uniform subsampling introduces an appealing perspective in the investigation of the stability of quantification results.

Considering the results of MAEs in Figure 4.12, the values of the MAE decrease with an increase in the size of the subsample, while Figure 4.11 shows that the permutation

test p-values increase with an increase in the subsample size. Secondly, the MAE values are very small and this is an indication that subsampling, in general, has a very small effect on the results of quantification. The MAEs for uniformly subsampled data flatten after the 20 % subsample and this indicates that any subsample less than 20 % in our data poses a danger of misrepresenting the parent data and thus it is not encouraged. These subsamples were included for the sake of experiments but should not be employed in practice.

In addition to these findings, the classification results in Figure 4.13 and Figure 4.14, give us accuracy results of LDA and SVM, respectively. Accuracy values lie between 68 % and 79 %. The good thing about classifying the uniform subsampled data is that even these smaller subsamples such as the 20 % subsample with 2888 rows still yield classification accuracy within the bounds of the ground truth.

Experiment 2 (see Section 4.5) on the other hand deals with biased subsampling of the parent data. In this experiment, we used two types of biased subsampling, stratified and extremely biased subsampling (see Section 4.5.2). These two subsampling techniques of biased subsampling uncovers some new insights on the stability of quantification results.

Firstly, considering the MAEs (Figure 4.42) and permutation test p-values (Figure 4.21) of stratified biased subsampling in Section 4.5.2, we noticed that the highest MAE value was associated with the data that consisted of distribution similar to the parent data, that we call a 50/50 subsample. The values of the MAE are generally very low, while the p-values are very high and this suggests that there is no significant difference between the results of parent data and these biased subsamples. Considering Figure 4.46 and Figure 4.47 where the classification results are presented, we notice that LDA and SVM accuracy results fluctuate between 74 % and 79 % for all the subsamples.

The second part of the analysis was devoted to extremely biased data. This is where we subsampled data with an exact row of one category in a dichotomous variable. The main variable we looked at was "Gender", with male and female categories, and the distribution of these categories is represented in Chapter 3, Section 3.2.2. The exact distribution is described in Figure 3.3. We noticed that there are more females than males in our data. Bearing that in mind, we looked at the MAE of these subsamples and the results are reported in Figure 4.22. The MAE is higher for data containing only females and lower for data containing only males. Contrary, the data that has more females coupled with other rows from the male category gives much better results, this can be seen from the distribution of the parent data. Similarly, the permutation test p-values for data containing more females is smaller than that of subsample containing more males which further confirms that the dataset containing more males has more similar results to the parent data than the one containing only females.

Considering the results in experiment 2 (see Section 4.5) and the supplementary results in Appendix A, the distribution of "Income" variable was more "No" and less "Yes". All these results suggest that categories with fewer rows are more likely to produce results similar to the parent data. This leads us to a further investigation of one property that might have influenced the results: the **sample size**.

Henceforth in experiment 3 in Section 4.6, we then focused on comparing the results of data subsampled uniformly and that which is obtained by biased subsampling, to test whether the sample size is the main **property** in yielding similar results to the parent data (as this may be suggested by the Central Limit Theorem [1]) . This theorem suggests that the results of a sample (or subsample) tends to become more similar with an increase in the sample size. However, for biased subsampling, this may not be the case as the results suggest in Section 4.5.2. Considering Figure 4.27a and Figure 4.27b we computed the actual sample sizes of these biased subsamples (that is 1 % females and 99 males % yield a certain number of rows), then we subsampled their equivalent subsamples uniformly (with equal sample size). We discovered that the same sample size does not yield the same quantification results because of the imbalances in the categories.

The subsamples of uniformly subsampled data are more likely to yield better results than biased subsampled data with the same size as uniform subsampled data and therefore, it is wiser to consider uniform subsampling when approaching the issue of subsampling multiple-choice data for use in the DS method. Also, the more balanced the data is in terms of rows of categories the more results will resemble those of the population or parent data. Thus, it is better to use stratified random subsampling.

Finally, regarding the classification of these two differently subsampled datasets, the results in Table 4.5 do not give any meaningful insights. Thus in practice, we would not encourage classifying the row weights obtained by extremely biased subsampling.

The overall conclusion about subsampling is that it is important to consider the type of subsampling before you begin with subsampling. If you care about the sample size, it is better to use uniform subsampling and if you care about the true representation of the data you may use stratified biased subsampling with even distribution, that is the 50/50 subsample. As far as subsampling is concerned, subsampling proves DS as stable and appropriate for quantification of categorical data.

[1]The Central Limit Theorem(CLT) suggests that as the sample size increases the distribution of the mean can be approximated by a normal distribution. In other words, the mean becomes similar to the mean of the population/ parent data (De Koker et al., 2020)

5.1.3 Resampling

The last three experiments in Chapter 4 were associated with resampling of data in different ways and the following experiments were undertaken, namely: Experiment 4 (see Section 4.7) was concerned with the resampling of the parent data, Experiment 5 (see Section 4.8) is dealing with resampling of biased subsampled and Experiment 6 (see Section 4.9) deals with resampling of uniform subsampled data.

In experiment 4 in Section 4.7 the parent sample was resampled 100 times, however, the number of replications differs according to the dataset. Although some literature may suggest 200 and above, even as little as 10 replications can reproduce results and thus be used for bootstrapping (Pattengale et al., 2009). This was proven empirically by observing the results from the 10th replication that they were converging (no longer changing). Although there are some small differences in the results of the resampled data, the difference is not significant since the MAE in Table 4.4 was close to zero (0.00001) and the p-value (0.9999) was the highest for the resampled data. Therefore the results are more similar to the original parent data than any size of subsampled data from the parent data. Furthermore, the classification results as displayed in Table 4.4 are more similar to the results of the original parent data in Section 4.3. Thus resampling reassures that the DS is stable under resampling.

The last two experiments (Experiment 5 and 6) consider a resampling of subsampled data, that is, given the uniform or biased subsampled data what is the impact of resampling these subsamples up to the size of the parent data (or the known population size)? Does it improve the results? Is it as good as having the population data? Hence, in experiment 5 (see Section 4.8), these questions are evaluated and answered. Firstly, consider the MAEs between the resampled biased data and parent data, in Figure 4.42. We discovered that as contrary to what we thought that the results of resampled data will be similar to those of our parent data, the results are more similar to the results obtained in Experiment 2 (see Section 4.5). Furthermore, considering Experiment 6 (see Section 4.9), the MAE (see Figure 4.51) and permutation tests (see Figure 4.51) also lead to the same conclusion that the results are more similar to the results of the uniform subsampled data as reported in Figure 4.12 and Figure 4.11 in Section 4.4.2. This indicates that using resampling methods may not necessarily yield desired results of the population (parent) but rather yields results of the sample (subsample), that is, the size before resampling. Ultimately, it also indicates that *"the resampling methods may not be used to replace data collection of adequate sample size but to intensify what is already at hand and allow quantitative analysis methods to be applied on smaller datasets that violate some assumptions about the sample size"*.

5.2 Discussion

This dissertation aimed to investigate the stability of quantification results. Stability in our study refers to the ability of the quantification results to remain the same as we change the data. A similar definition is found in (Gifi, 1990) where stability is defined as the degree of sensitivity of the quantification method to the changes in the data which is described as "small unimportant changes in the data should lead to small unimportant changes in the quantification results". The emphasis of stability analysis was more focused on one particular kind of categorical data, i.e, contingency tables. Hence this study is more concerned with the investigation of the stability of multiple-choice datasets. This is a new challenge to assess the stability of quantification based on many facets of sampling.

Resampling methods can help us answer these questions. In addition to the use of resampling methods, this study set out to investigate the effect of subsampling data, and whether or not it affects the quantification results.

These questions have been considered for contingency tables, however, to the best of our knowledge these questions are not addressed in these studies in the context of multiple-choice data (Lombardo et al., 2012; Linting and van der Kooij, 2012; Linting et al., 2007; Lebart, 2006).

Thus experiments undertaken in this study are associated with the quantification of multiple-choice data, however, the results may also apply to other kinds of categorical data, in particular, sorting, rank-ordered and contingency tables. This is only possible if the mentioned datasets are pre-processed in a suitable format. Some additional experiments were undertaken to extend the generality of our results (see Appendix B). Although the results from our 2017 GHS data may not show fancy trends, we still arrive at the same conclusions. We are confident that these subsampling techniques and resampling techniques may be applied to other multiple-choice data similar to our synthetic data and the 2017 GHS data.

Besides resampling data where the resamples have the same size as the parent data, another type of resampling undertaken in this study involved resampling the subsampled data, where the size of the resamples was made to be equal to the size of the parent sample. The results of this experiment (see Section 4.9.2) show that the resampling method, bootstrap, in this case, does not necessarily recover the results of the original sample although it may seem that since the resamples have the size of the parent data then DS results associated with them should give results that are very similar to those of the parent data. This indeed is in agreement with what Fisher (1995) was warning against, that *"Bootstrap should not be used as a substitute for collecting more data"*. In

contrary to this warning, Yu (2003) suggested that the resampling can be used to compensate for lack of data collection, that is, drawing a larger sample size from the smaller sample as if it was a population would help in yielding better results than if we would analyze the smaller population. This is rather not advisable empirically as tested in Experiment 6 (see Section 4.9).

5.2.1 Extension of DS to other data types

Dual scaling is a quantification method for handling different types of categorical data. As seen in Chapter 2, this method yields similar results to those of Multiple Correspondence Analysis (MCA) for multiple-choice data.

This means the results that we get in this study may apply to other quantification methods[2], especially MCA. This holds mainly because quantification can be seen as a set of related methods for quantifying categorical data. Thus the weights obtained through such methods are similar for similar types of data (Maraun et al., 2005).

For a quick comparison of quantification methods on different types of data such as contingency tables and multiple-choice data, firstly consider the inertias (see Section 4.2) or the percentages of variance accounted for (VAF) by each component of our parent data in as reported in Section 4.2 in Table 4.1. The values of VAF for the first and second dimensions are very small in our data since the first two components only account for 15.579 % (that is 7.986% and 7.593 % for the first and second dimension, respectively) of the total variance. This is indeed very small as compared to the inertias (eigenvalues) of contingency tables. This is not entirely a problem since this is because the remaining variance is coupled to other components and this is often the case with real datasets, such as that of 2017 GHS data. However, if a researcher finds this problematic the suggestion is to cut some outliers or categories with small frequencies [3].

The resampling of subsampled datasets where the sample size of each replicate is equal to the sample size of the parent sample does not yield new or better results that resemble the properties of the parent sample but instead, the resampling affirmed the results for the subsamples. This property is an indication that the resampling methods, especially bootstrap should not be used as a substitute to collect more data.

Finally, the resampling methods can help uncover possible variations in the results of quantification. Although resampling methods are useful techniques they pose a major

[2]The reader is referred to the following sources of other related quantification methods, including Homogeneity Analysis (Di Franco, 2016), MCA (Greenacre and Blasius, 2006), etc

[3]The outliers are not dealt with in this study since the aim was to evaluate the stability of results, not necessarily the application of DS

challenge on memory when they are used in multivariate data. The solid theoretical foundation of these methods was focused on univariate data and thus need careful attention when applied to multivariate data.

5.2.2 Classification

Classification using the projected weights involves discarding some irrelevant components of the solution. This is normally done through the univariate tests of comparison (Greenacre and Blasius, 2006). Although using all the components may not enable precise and reliable predictions, the aim of this study was not to get accurate results or better classification results, but to test whether the results of classification seen on the parent sample remain the same as the data changes. If these results are changing significantly that shows instability in the quantification results.

The findings of our study regarding classification as shown in Section 4.4.2, Section 4.5.2, Section 4.6.2, Section 4.7.2, Section 4.8.2 and Section 4.9.2 indicate and thus confirm that the quantification results are stable given the subsampling and resampling strategies. Furthermore, although the results of quantification have slight changes over the uniform and biased subsampling strategies, classification results did not change significantly, and this confirms that dual scaling is indeed stable.

Chapter 6

Conclusion and Recommendations

6.1 Introduction

The purpose of this chapter is to review the overall conclusion given the objectives of the study and the discussions presented in Chapter 5. In Section 6.2 we discuss the overall conclusion for the study while Section 6.3 deals with possible recommendations and suggestions for future studies.

6.2 Conclusion

The are many problems underlying quantification methods and in this research, we were interested in the stability of the results. Although quantification methods may allow us to visualize the associations between the row and columns in one-mode plots, biplots and joint plots, the problem may be whether the patterns projected on these graphs are true or they occurred by mere chance. To investigate this question, subsampling and resampling methods were employed in our multiple-choice data. Several facets of the data and possible structures of the data were investigated using the error metric (MAE) (see Equation 3.6), permutation tests p-values)(see Section 2.7.3) and finally, using the classification methods (see Section 3.6).

The results show that the quantification results change with a change in the data. The extent to which the results change depends on the sampling strategy being used in the data, specifically, uniform or biased subsampling. Also, uniform subsampling seems to

be a better choice when the aim is to subsample based on sample size as compared with biased subsampling which is a better choice when we want the representation similar to the parent data for making inferences. On the other hand, resampling gives confidence in the quantification results by reconstructing the results of the input data from several similar datasets.

Finally, although the quantification results are changing due to subsampling and resampling, the change is not significant. Furthermore, the classification results do not change that much and therefore this implies that the DS is indeed stable.

Reflecting on our hypotheses we conclude that indeed there is no significant difference between the quantification results of subsampled data. This was supported by smaller MAE values which are very close to zero and permutation test p-values that are very close to 1, especially for larger subsamples. Secondly, there is no significant difference between the results of the resampled data and the parent data. This is generally supported by the MAE and the classification evaluation measure, accuracy for both LDA and SVM.

6.3 Recommendations

This research presented an approach for assessing the stability of quantification results on multiple-choice data. Although the results may be generalized to other multiple-choice datasets such as the census, survey data, questionnaire datasets and all related datasets, this study does not cover the numerical representation of variables. It would be interesting to investigate the quantification of categorical variables in addition to the quantification of categories. Quantification of multiple-choice data becomes more expensive when there are many categorical variables in the data as this spans more space and thus it would be interesting to consider a high-performance implementation of dual scaling and related methods.

Although bootstrap is a useful method for assessing the stability of quantification results, the permutation tests also prove to have the potential of obtaining even more precise results as the method investigates the significance of the observed variation using p-values (see Section 4.52). It would yield a great advantage in future to consider the high-performance implementation of bootstrap and related methods for assessing the stability of quantification results. This can be seen since we used 100 replications and the number of experiments needed to run was 20, which implies that a total of 2000 replications was run. These experiments were expensive on classification methods, in particular, the SVM as it naturally takes time to run on large datasets. Thus for future studies both bootstrap and SVM should be optimized by incorporating high-performance

computing. Although Hesterberg (2015) argues that computers have become faster, resampling may be burdensome for multivariate data.

Finally, we cannot overemphasize the importance of collecting more data since resampling methods should not be used as a substitute for collecting more data. This was subsequently shown in Experiment 5 and 6 (see Section 4.4.2 and Section 4.8.2).

Appendix A

Supplementary Results

A.1 Additional Experiment Results

To ascertain the validity and reliability of our study we also looked at some of the experiments that may support our conclusions. In section A.1.1, we look at the one-mode plots for the rest of the subsamples we have in experiment 4.4. Finally, in section A.1.2, we also looked at other dichotomous variables [1] and we only report on the most important aspects of the experiment, in particular, the error metric report (MAE), permutation test and the classification results(LDA and SVM accuracy).

A.1.1 Experiment 1 supplementary results

In the following section, we look at further analysis of the one-mode plots for the uniform subsampled dataset. The figures are presented in the order of uniform subsampling from the smallest to the highest percentage of subsampling. The results are presented in Figure A.1. For each of these one-mode plots we noticed that the grouping is not normally affected by the subsampling strategy, that is, even if the structure of a new subsample deviates with the structure of a parent data, the groupings are still maintained.

[1] A variable with two categories.

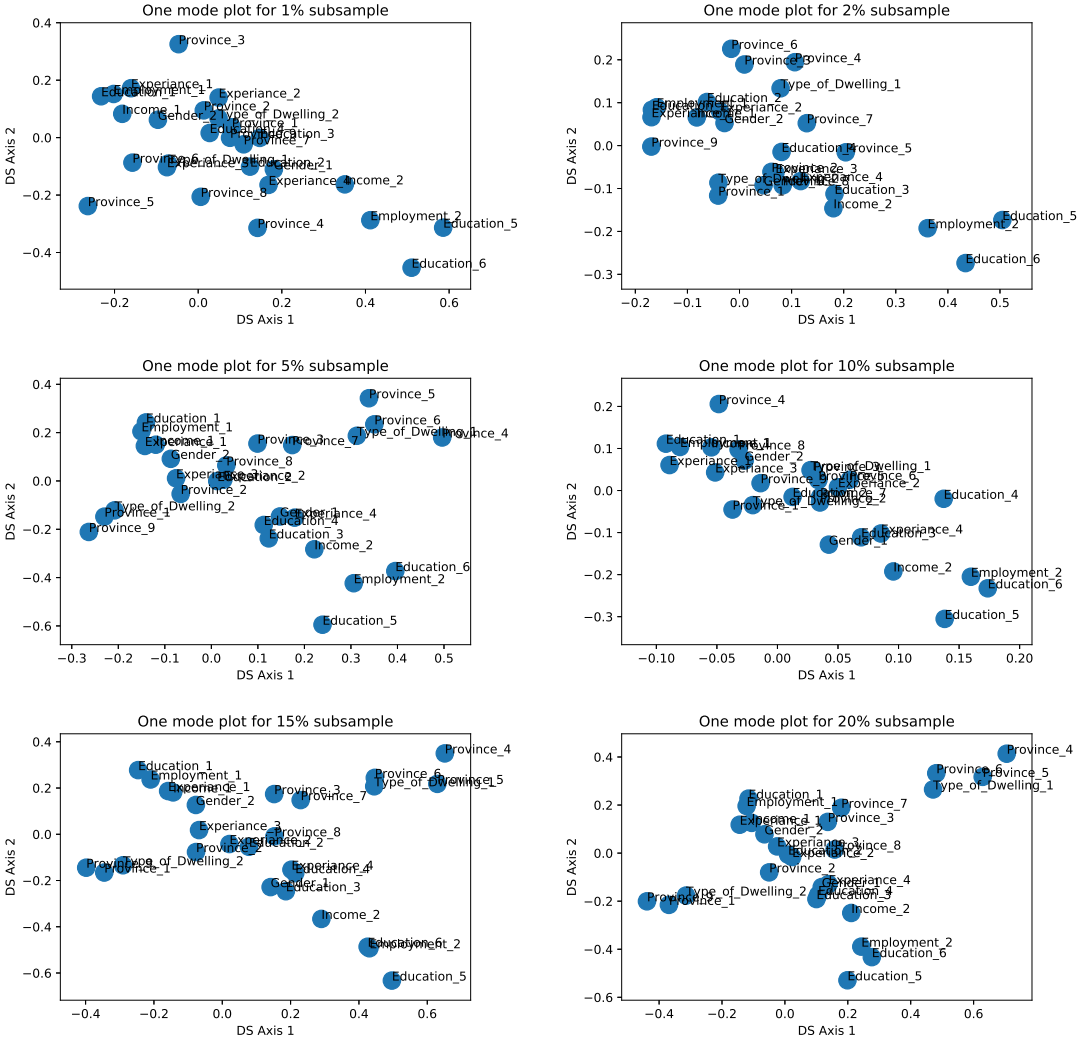


FIGURE A.1: Detailed one-mode plots for uniform subsampled data

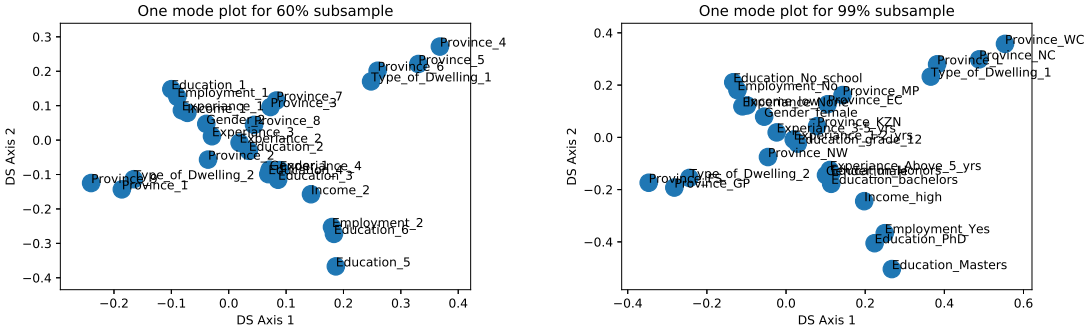


FIGURE A.3: Detailed one-mode plots for uniform subsampled data

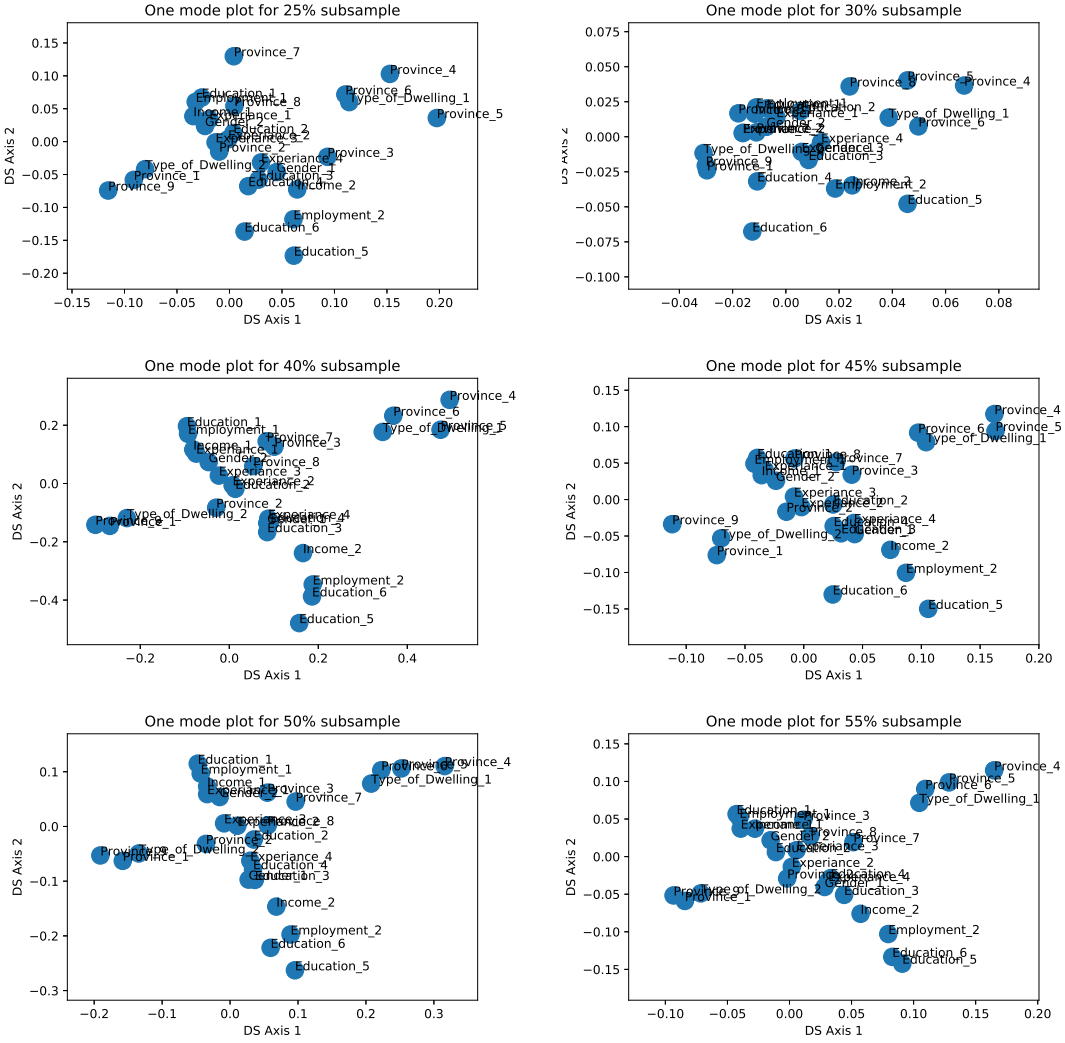


FIGURE A.2: Detailed one-mode plots for uniform subsampled data

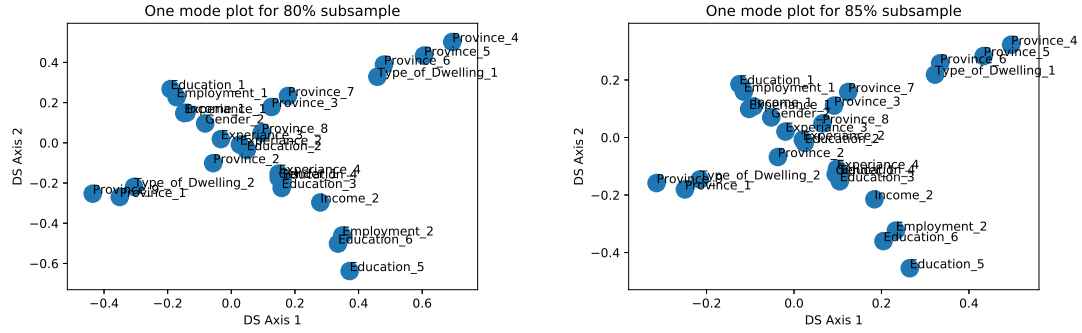


FIGURE A.4: Detailed one-mode plots for uniform subsampled data

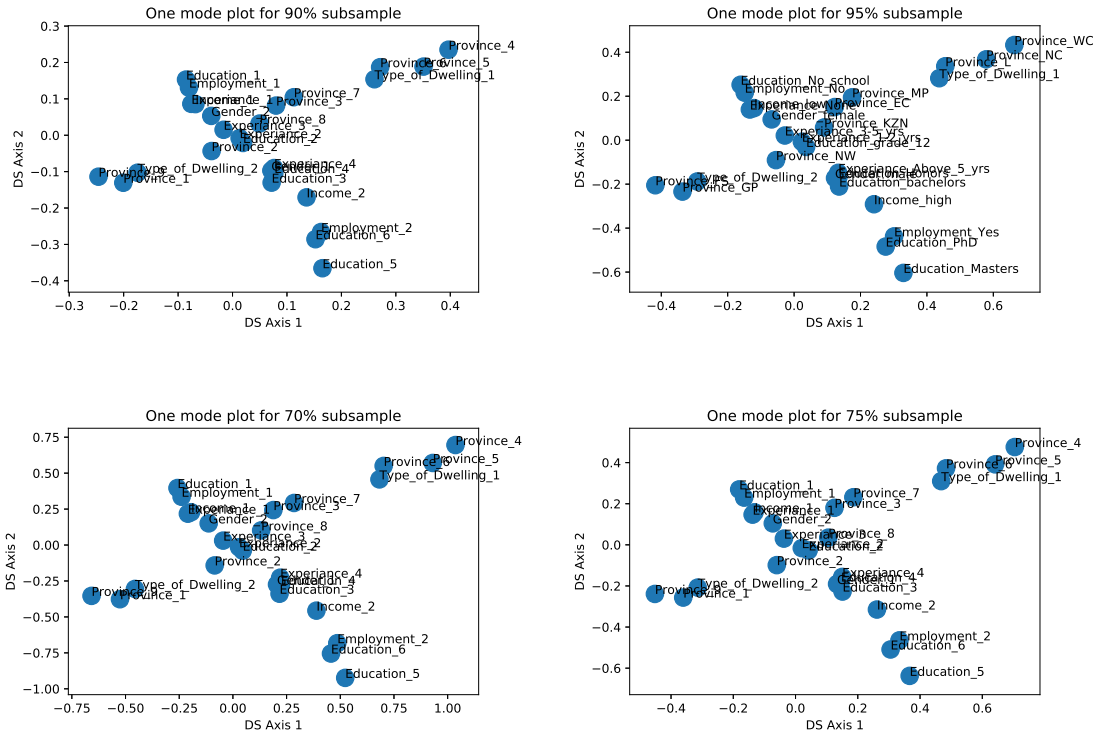


FIGURE A.5: Detailed one-mode plots for uniform subsampled data

A.1.2 Supplementary Biased Subsampling

In this section, we applied experiment 2 on two other dichotomous variables, namely, Income and Employment, respectively.

For us to understand the impact of biased subsampling in dual scaling results, we first look at the distribution of the "Income" variable as reported in section 3.2.2. There are 4805 individuals have high income and 9595 individuals who have low income. Having considered the distribution, we now present the distribution of eigenvalues corresponding to the income variable. The results of eigenvalues distribution are reported in Figure A.6.

The MAEs were computed for biased subsamples and the results are reported in Figure A.7. The MAEs are much similar to the ones obtained in experiment 2, that is, they suggest that the biased subsampling of dichotomous variables maintains the structure of results for the parent data provided that the subsampling is halved. In simpler terms, if I have a question that has options "Yes" and "No" and suppose 60 people out of 100 answered "yes" and the remaining 40 answered "no", then if I subsample 50% "Yes" again 50 % "No", this will not change the interpretation of results. For the rest of this study, the said subsample (50 % male and 50 % female) is hereafter referred to as 50/50

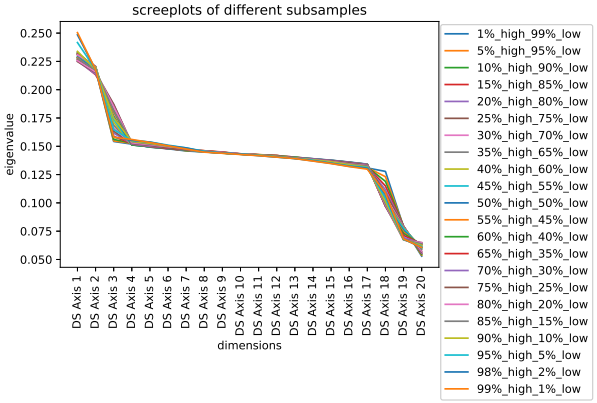


FIGURE A.6: Distribution of eigenvalues for "Income"

subsample. A moment we skew the subsampling we see a deviation on the results. The p-values of the permutation test in Figure A.8 also agree with the MAEs.

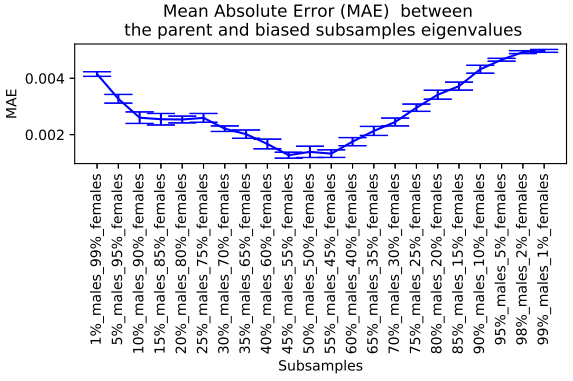


FIGURE A.7: MAE for biased subsampled data: income variable

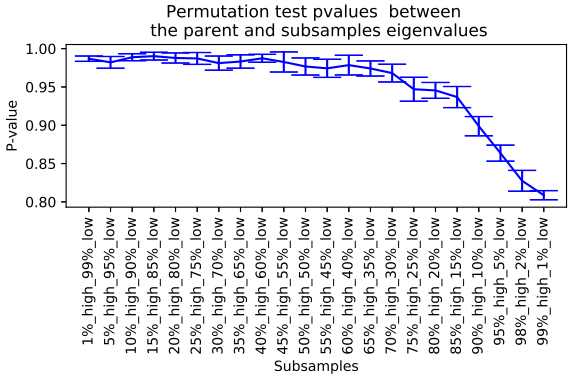


FIGURE A.8: Permutation p-values for biased subsampled data: income variable

Finally, we also looked at the classification of biased subsampled data. The results of LDA accuracy and SVM accuracy are reported in Figure A.10 and Figure A.11. Figure A.10 shows that LDA accuracy for the 50/50 subsample with 95 % confidence interval

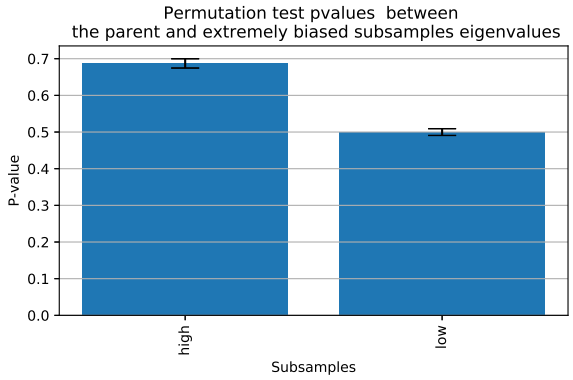


FIGURE A.9: Permutation p-values for extremely biased subsampled data: income variable

the accuracy lies between 75.75 % and 76.25 % and this interval covers exactly what we got for classification of the parent data. The SVM accuracy as reported in Figure A.11, bring us to the same conclusion that the 50/50 subsample is more similar to the parent data than any other subsample. The overall performance of LDA lies between 75.25 % and 78 %. This draws us to the same conclusions as to the ones in experiment 2.

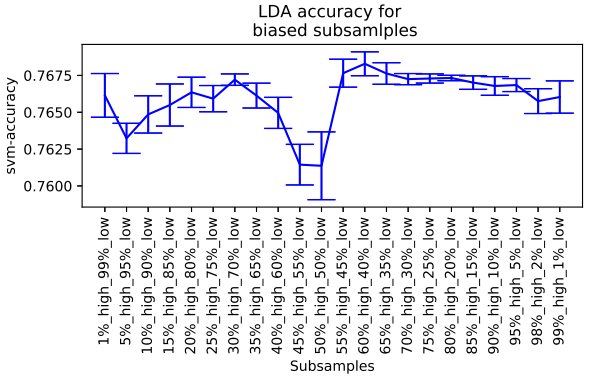


FIGURE A.10: LDA accuracy for biased subsampling: income

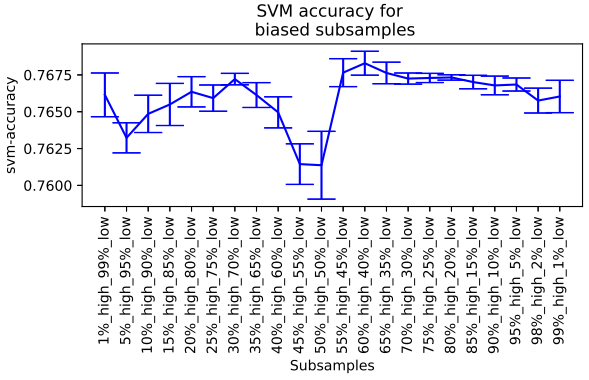


FIGURE A.11: SVM accuracy for biased subsampling: income

A.1.3 Employment

To ascertain that these results were not just a coincidence, the "Employment" variable was also subjected to test and the results of the category distribution, eigenvalues distribution, MAEs and permutation tests p-values are reported in section 3.2.2, Figure A.12, Figure A.13 and Figure A.14, respectively. The employment variable consists of two categories namely, "Yes" and "No". The distribution of categories of employment is as follows: "Yes" has 4754 rows while "No" has 9646 rows. This shows us that there are more unemployed individuals than there are employed people. The distribution of eigenvalues in Figure A.12 shows a similar distribution to the one obtained in experiment 2 (see Section 4.5).

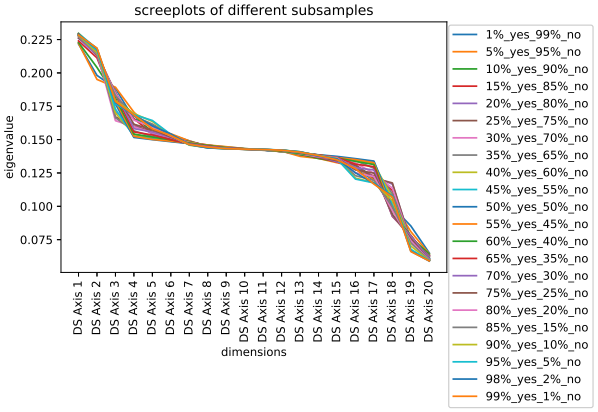


FIGURE A.12: Scree plots of different biased subsamples: employment variable

Figure A.13 shows us the MAE between the subsampled and parent data, the MAEs shows great similarity to the results obtained subsampling of "Gender" variable in experiment 4.5 in Figure A.13.

Following the MAEs we also looked at the permutation test p-values. Figure A.14 shows that we have a high p-value for a 50/50 subsample which ascertains that this subsample

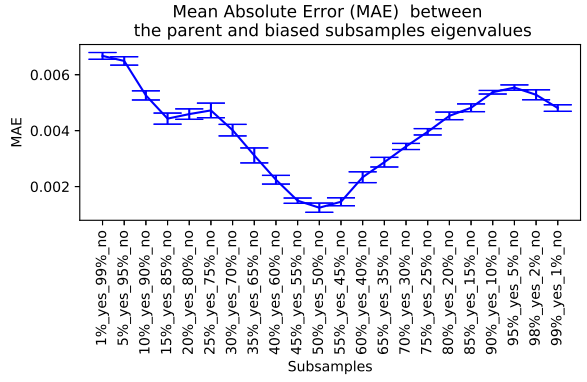


FIGURE A.13: MAE for biased subsampled data: employment variable

is very similar to the parent data, hence the results also will be similar at 95 % confidence interval. We also looked at subsampling extremely biased data, but we only present one aspect of the results in Figure A.15, that is, permutation test associated with these categories

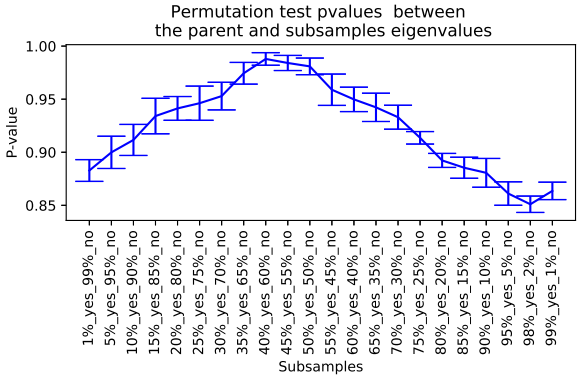


FIGURE A.14: Permutation p-values for biased subsampled data: employment variable

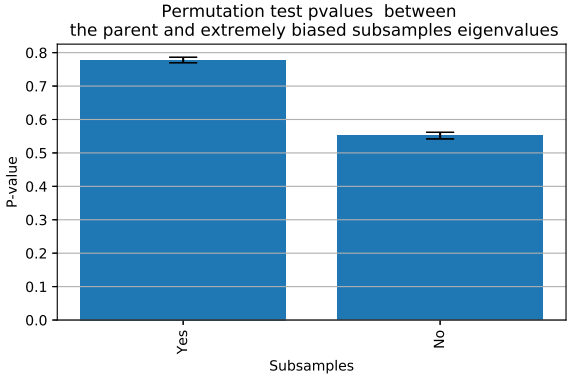


FIGURE A.15: Permutation p-values for extremely biased subsampled data: employment variable

Finally, we considered the classification results of this biased subsampled data, the results are reported in Figure A.16 and Figure A.17, respectively.

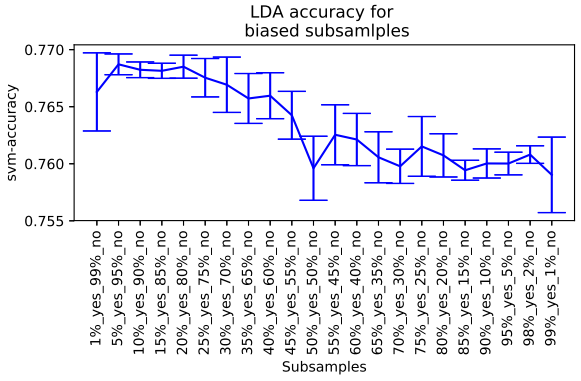


FIGURE A.16: LDA accuracy for biased subsampling: employment

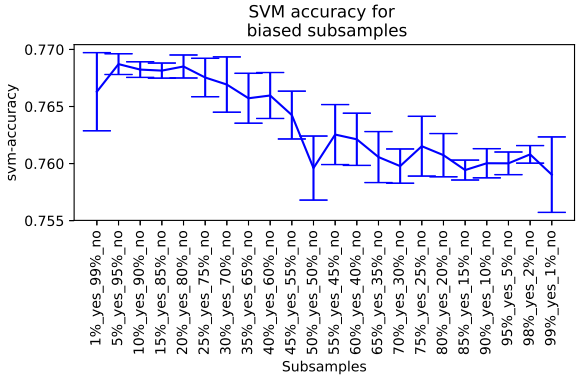


FIGURE A.17: SVM accuracy for biased subsampling: employed

Appendix B

Application to 2017 GHS data

B.1 Application on General Household Survey 2017 data

The uniform subsampling was applied to the general household survey dataset of South Africa. The results are presented in the following sections. We first present the data description in Section B.2 and then present all the analysis relating to our study, specifically, Section B.2.1 deals with the results relating to the 2017 GHS data, while Section B.2.2 reports on experiment 1, Section B.2.3 gives report on experiment 2, then experiment 3 is reported in Section B.2.4, Section B.2.5 deals with the resampling of parent GHS data, while Section B.2.6 and Section B.2.7 reports the resampling of biased subsampled data and uniform subsampled data. Finally, Section B.2.8 presents the summary of results for 2017 GHS data.

B.2 Data Description

The 2017 General household survey data of South Africa consists of 360 variables, but for the interest of our study, we only studied at most, seven categorical variables, as described in Table B.1 (StatsSA, 2017). The seven variables were chosen based on similarity to our study data, either in the number of categories of just the variable name. For an example, the variable income in our study had only two categories (yes or no), but the one chosen from this survey has several variables because it describes the different source of income, not just the question of whether you get income or not. Some categories were omitted for experimentation and illustration. In particular, in the variables: vehicle, the third category, unspecified was omitted, Dwelling, "other" category was removed and finally, in the variable Income, the "unspecified" category

was removed in the data. The data as obtained from Statistics South Africa[1] had 21226 cases and was subsampled for the purpose of the study. After removing the three categories from the data, the new rows resulted in 13553 rows with 7 categories.

TABLE B.1: Data Description of 2017 Household survey data

Category	Label
race_1	African/Black
race_2	Coloured
race_3	Indian/Asian
race_4	White
business_income_1	Yes
business_income_2	No
Gender_1	Male
Gender_2	Female
Income_1	Salaries/wages/commission
Income_2	Income from a business
Income_3	Remittances
Income_4	Pensions
Income_5	Grants (include old age grant here)
Income_6	Sales of farming products and services
Income_7	Other income sources e.g. rental income, interest
Income_8	No income
vehicle_1	Yes
vehicle_2	No
Dwelling_1	Dwelling/house or brick/concrete block structure on a separate stand or yard or farm
Dwelling_2	Traditional dwelling/hut/structure made of traditional materials
Dwelling_3	Flat or apartment in a block of flats
Dwelling_4	Cluster house in complex
Dwelling_5	Town house (semi-detached house in complex)
Dwelling_6	Semi-detached house
Dwelling_7	Dwelling/house/flat/room in backyard
Dwelling_8	Informal dwelling/shack in backyard
Dwelling_9	Informal dwelling/shack not in backyard, e.g. in an informal/squatter settlement or on farm
Prov_WC	Western Cape
prov_EC	Eastern Cape
prov_NC	Northern Cape
prov_FS	Free State
prov_KZN	Kwazulu Natal
prov_NW	North West
prov_GP	Gauteng Province
prov_MP	Mpumalanga
prov_L	Limpopo

[1]The data can be obtained at UCT Data portal https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/723/datafile/F5/?offset=200&limit=100

B.2.1 Dual scaling of the 2017 GHS data

The 2017 GHS data as described in Section B.2, has 7 variables with 13553 rows. The total number of categories for this data is 35 and the summary of the number of categories in each variable is reported in Table B.2. Thus the number of expected eigenvalues is calculated as $35 - 7 = 28$.

TABLE B.2: Summary of categories for 2017 GHS data

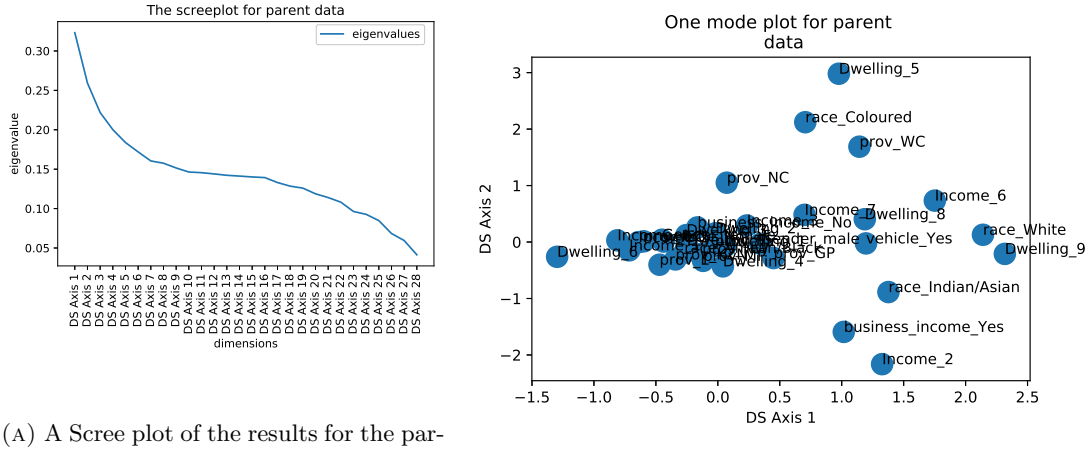
Variable	Province	Race	Income	Business income	Gender	Vehicle	Dwelling
Number of categories	9	4	7	2	2	2	9

The summary of eigenvalues are reported in Table B.3. We also looked at the distribution of eigenvalues is reported in Figure B.1a. This distribution together with a one-mode plot of column weights in Figure B.1b, was used as ground truth and was compared with the rest of the experiments.

TABLE B.3: Summary of the first 5 eigenvalues for GHS data

	DS Axis 1	DS Axis 2	DS Axis 3	DS Axis 4	DS Axis 5
Inertia	0.323155	0.259209	0.221902	0.200187	0.183810
Delta	8.078868	6.480226	5.547542	5.004665	4.595258
Cum_Delta	8.078868	14.559093	20.106635	25.111300	29.706558

In Figure B.1a we see the distribution of eigenvalues for our GHS data, it shows that eigenvalues are curvilinearly distributed. The one-mode plot of GHS data in Figure B.1b shows that the person living in Western Cape and who is a coloured person is most likely to live in a Townhouse (Dwelling_5). Another grouping that we see is that of an individual who is Asian or Indian who is likely to receive income from a business and the type of dwelling is informal and this depicts the real situation in townships. The subsequent experiments will test the hypothesis statement outlined in Section 1.3.



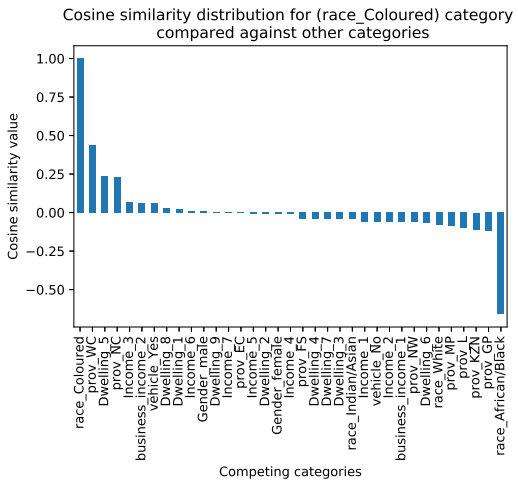
(A) A Scree plot of the results for the parent sample

(B) One-mode plots

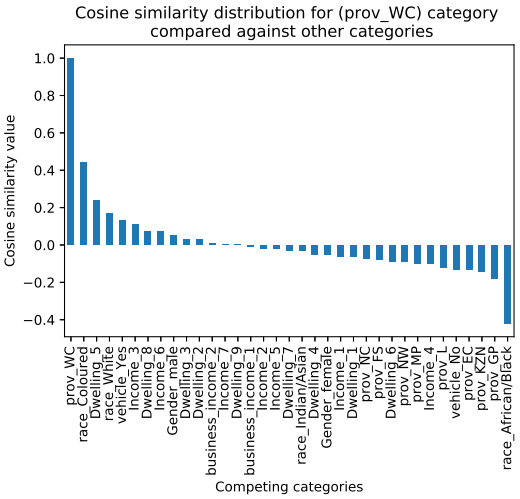
FIGURE B.1: A scree plot and one-mode plot for 2017 GHS data

In order to correctly interpret the distances or similarities between the categories, cosine similarity measure and a Euclidean distance metric were used to check the distances as depicted by the one-mode plot in Figure B.1b. Two categories were considered to make a concise interpretation. Figure B.2a and Figure B.2b correspond to the cosine similarity of "Western Cape Province" and "Coloured race". While Figure B.3a and Figure B.3b corresponds to the Euclidean distance of the same categories.

Figure B.2a shows that "Coloured race" or Coloured ethnic group people are associated with "Western Cape province", "Dwelling_5", "Province_NC" and "Income_3" category, respectively. According to Figure B.2b, the "Western Cape province" is strongly associated with "Coloureds" ethnic group, "Dwelling_5", "Whites" ethnic group and "Vehicle owners" category.



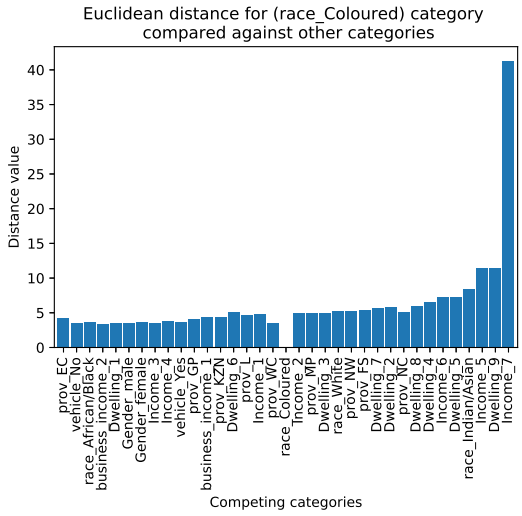
(A) Cosine similarities for Coloured race category



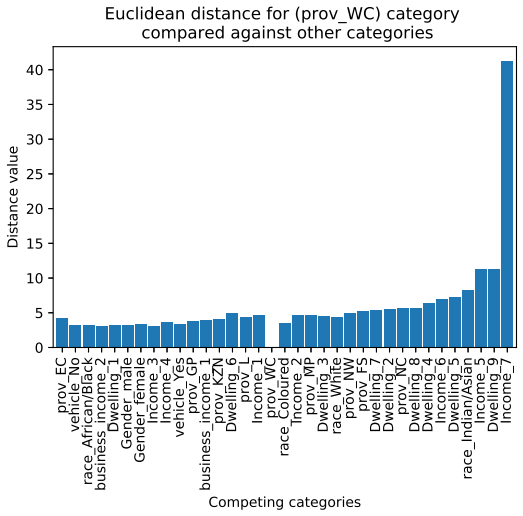
(B) Cosine similarities for Western Cape Province category

FIGURE B.2: Cosine similarities for Coloured race and Western Cape Province categories

The Euclidean distances of the Coloured race category in Figure B.3a show that this category is closer to "Business income 1", "Western Cape province" and "Male gender " category. Figure B.3b shows that the Western cape category is more closer to "Business_income_2", "Income_3" and "Gender_male" category, respectively.



(A) Euclidean distance for Coloured race category



(B) Euclidean distance for Western Cape Province category

FIGURE B.3: Euclidean distances for Coloured race and Western Cape Province categories

To explore these similarities further, the hierarchical clustering algorithm was considered and the results are reported in Figure B.4. Figure B.4 shows that the GHS data can be

divided into 3 groups or clusters, the smallest cluster contains 3 categories which are: "Western Cape province", "Coloured race" and "Dwelling_5". The third cluster consists of ""Business_income_1", "Income_2", "Income_6", "Dwelling_8", "Vehicle owners","Indian or Asian race", "white race" and "Dwelling_9". The remaining categories belong to the second cluster which has most of the categories.

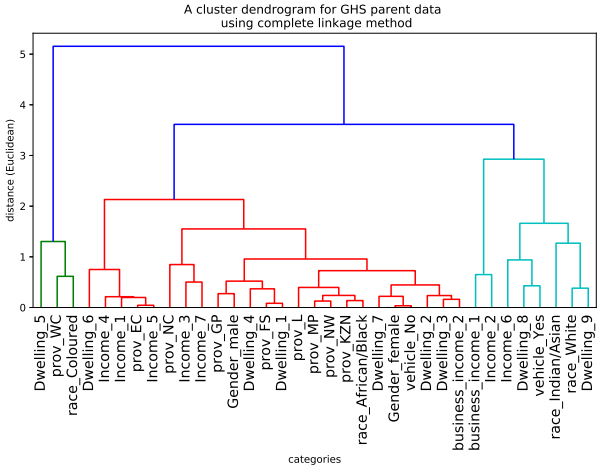


FIGURE B.4: Hierarchical clustering results for GHS parent data

The GHS data was also subject to classification, where the target variable was "Business income". We used 5 fold cross-validation and the results are reported in Table B.4. The results show that LDA and SVM accuracy lies between 93 % and 95 % at 95 % confidence interval.

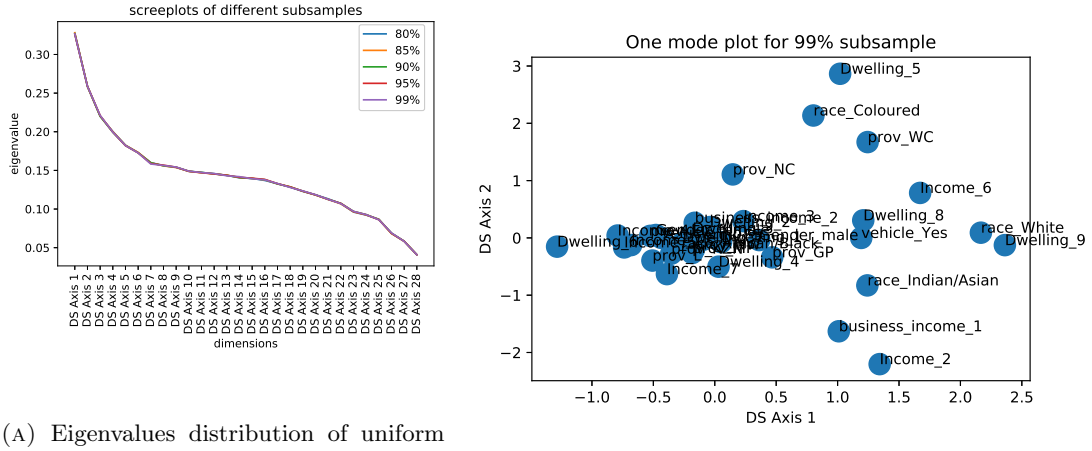
TABLE B.4: GHS classification results

	lda_accuracy	lda_precision	lda_recall	svm_accuracy	svm_precision	svm_recall
classification_parent	0.94 (+/- 0.01)	0.97 (+/- 0.03)	0.57 (+/- 0.08)	0.94 (+/- 0.01)	0.97 (+/- 0.04)	0.57 (+/- 0.03)

B.2.2 Experiment 1: Uniform subsampling vs parent data (GHS)

In this experiment, we subsampled the GHS parent data using uniform subsampling and the results retained are as follows: Scree plot of eigenvalues, Hierarchical clustering results, One mode plot, MAE, Permutation tests p-values and classification results (LDA and SVM), respectively.

Figure B.5a shows the distribution of different uniform subsamples, the distribution of eigenvalues for subsamples are similar to the scree plot of the parent data in Section B.2.1, Figure B.1a. We also looked at the one-mode plot for 99 % subsampled data in Figure B.5b and the results show that this small change in the data does result exactly in small change in the results of dual scaling results.



(A) Eigenvalues distribution of uniform subsamples

(B) One-mode plot

FIGURE B.5: A scree plot and one-mode plot for 99 % uniform subsampled data

The hierarchical clustering algorithm was also performed on the 99 % subsampled data and the results are reported in Figure B.6. The results of a 99 % subsampled data maintain the three clusters that are found in the parent data. Although there are some differences in the values of Euclidean, the grouping did not change and this proves that 1 % change in a data has an unimportant change in the results of the original parent data.

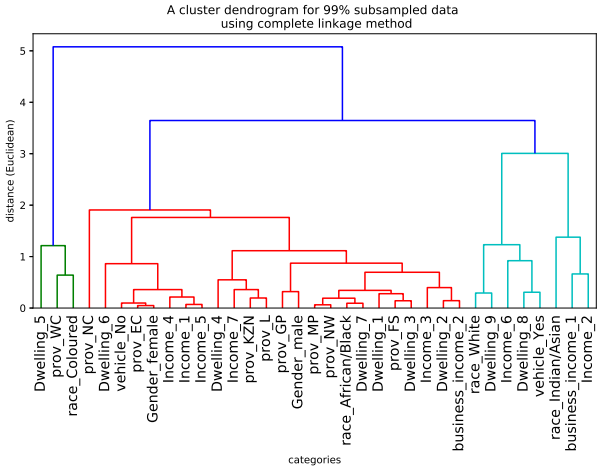


FIGURE B.6: Hierarchical clustering results for 99 % subsampled data

The MAEs as reported in Figure B.7 show that the MAEs decrease with larger subsamples. This is in agreement with what we got in experiment 4.4. On the other hand, the p-values in Figure B.8 shows an increase as we increase the sample size.

.

Finally, Figure B.9 and Figure B.10 present the classification accuracy for LDA and SVM, respectively. The results show that LDA accuracy lies between 93.32 % and 93.45.

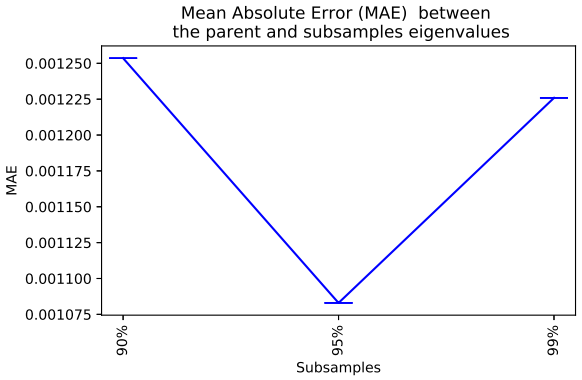


FIGURE B.7: MAE for uniform subsampled data

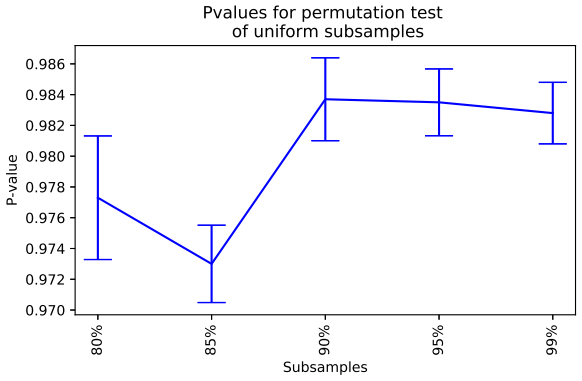


FIGURE B.8: Permutation p-values for uniform subsampled data

While SVM accuracy lies between 93.32 % and 93.48. In both cases, the accuracy for the 99 % subsample, is very low and the graphs show some convergence since we see a flat line.

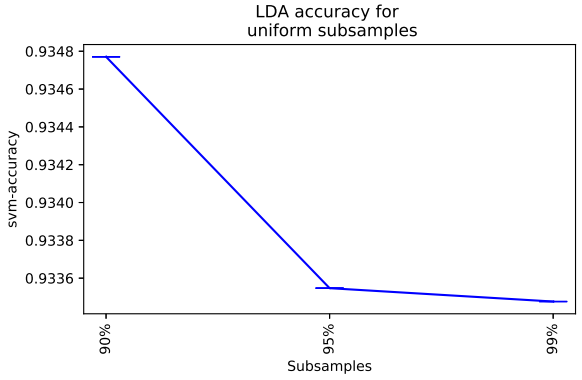


FIGURE B.9: LDA accuracy for uniform subsampled data

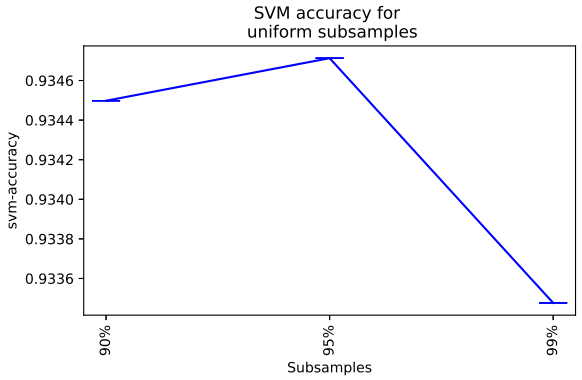
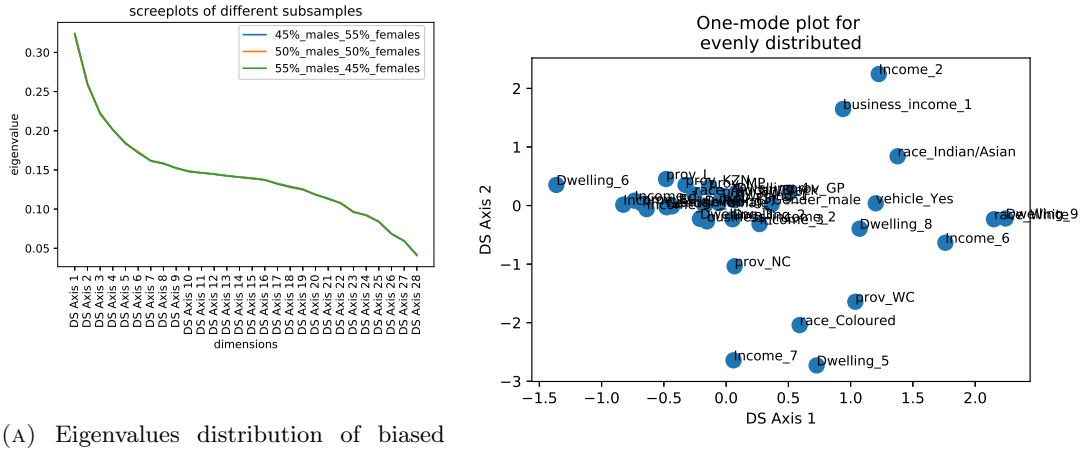


FIGURE B.10: SVM accuracy for uniform subsampled data

B.2.3 Experiment 2: Biased subsampling vs parent data (GHS)

In this experiment, we subsampled the GHS parent data using biased subsampling and the results retained are as follows: Scree plot of eigenvalues, One mode plot, MAE, Permutation tests p-values and classification results (LDA and SVM), respectively.

Figure B.11a shows the distribution of eigenvalues across different subsamples while Figure B.11b shows a one-mode plot of a 50/50 subsample.



(A) Eigenvalues distribution of biased subsamples

(B) One-mode plot

FIGURE B.11: A scree plot and one-mode plot for 50/50 biased subsampled data

The biased subsampled data was also subjected to Hierarchical clustering and the results are reported in Figure B.12. The clustering resulted in 2 clusters only as opposed to the 3 clusters in Figure B.4. It is also worth noting that the first cluster is exactly the same as the first cluster of the parent data. Thus, even though there are two clusters the similarities between the categories are not affected by the subsampling.

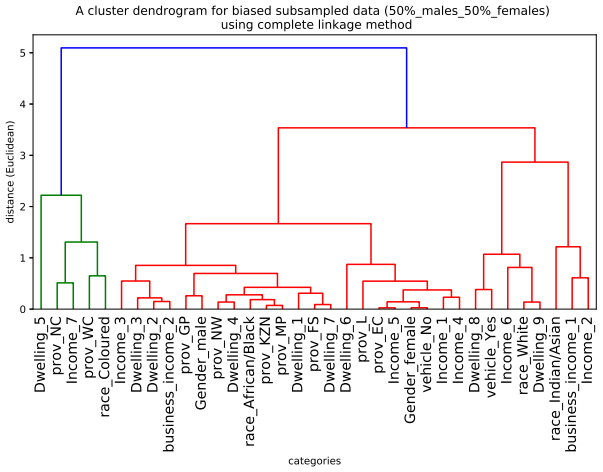


FIGURE B.12: Hierarchical clustering results for 50/50 subsample

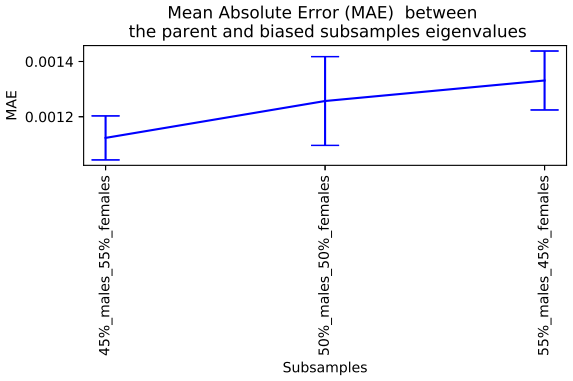


FIGURE B.13: MAE for biased subsampled data

Following that we looked at the MAE in Figure B.13 and the results show that the 50/50 subsample preserves the characteristics of the parent data. The MAE is small for most of our findings. There are some fluctuations in the 2017 GHS.

We also looked at permutation test p-values in Figure B.14 and we notice that the subsample with evenly distributed(50/50) rows yields the high p-values. However the p-values are all close to 1 and that is a good thing, that shows that they are indeed similar to the parent data.

Finally, we looked at the LDA and SVM in the Figures, Figure B.15 and Figure B.16. The highest accuracy attained by both LDA and SVM is associated with 95 % subsample. There is a bit of fluctuation in the accuracy values but the values are all in the range 93.50 % and 93.55 %. This is way better since this is real data.

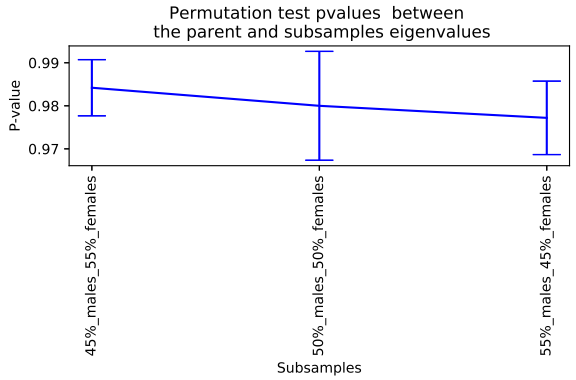


FIGURE B.14: Permutation p-values for biased subsampled data

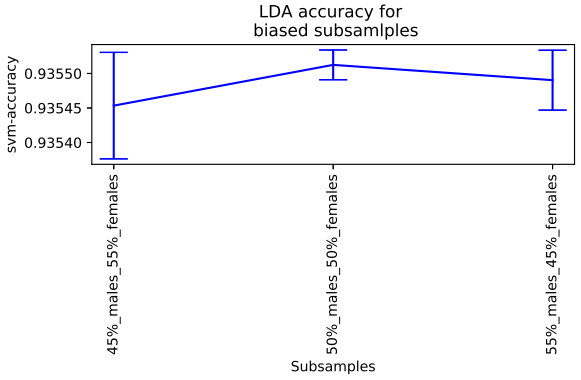


FIGURE B.15: LDA accuracy for biased subsampled data

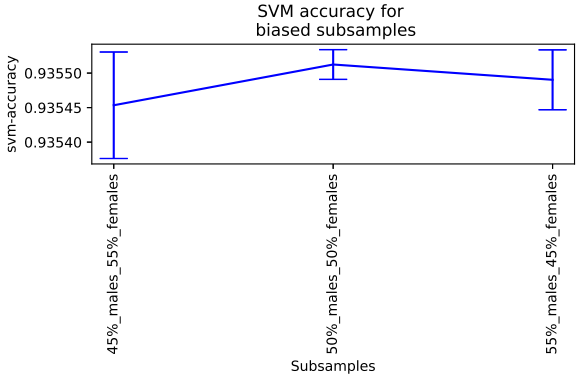
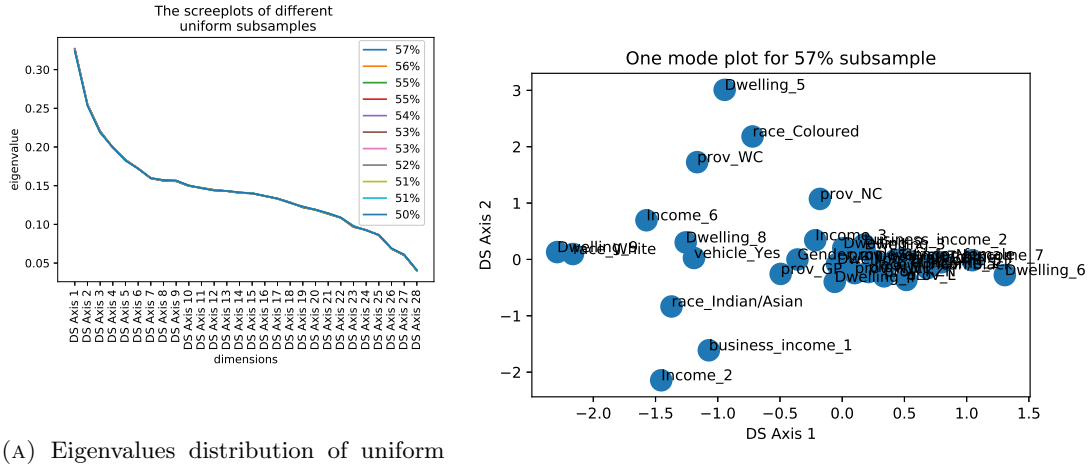


FIGURE B.16: SVM accuracy for biased subsampled data

B.2.4 Experiment 3: Uniform subsampling vs Biased subsampling

In this experiment, we subsampled the GHS parent data using biased subsampling and the subsampled it using the uniform subsampling and the results retained are as follows: Scree plot of eigenvalues, One mode plot, MAE, Permutation tests p-values and classification results (LDA and SVM), respectively.

Figure B.17a report the distribution of the eigenvalues while Figure B.17b shows the one-mode plot for the average column weights. The eigenvalues are similar to those of the parent data and the one-mode plot for 57 % subsample is also similar to what we got in Figure B.1b.



(A) Eigenvalues distribution of uniform subsamples

(B) One-mode plot

FIGURE B.17: A scree plot and one-mode plot for 57 % uniform subsampled data

To strengthen the results Hierarchical clustering was also applied to the subsampled data and results are reported in Figure B.18. The results show that the data has be grouped into two clusters and the first cluster contains 4 categories that are similar to those of the parent data in Figure B.4.

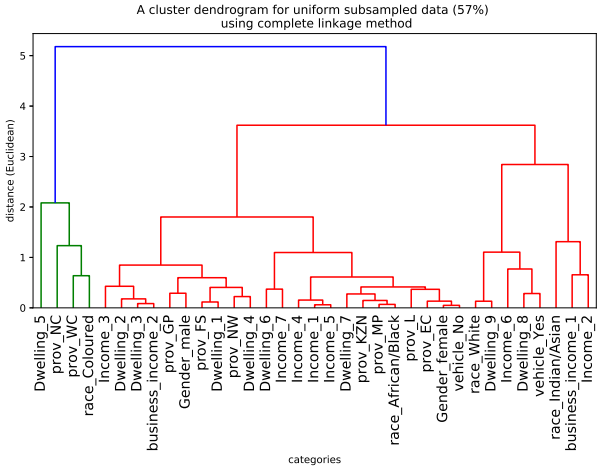


FIGURE B.18: Hierarchical clustering for 57 % uniform subsampled data

In Figure B.19, we considered MAE for the uniform subsampled data and the results show that the MAE decreases as we increase the sample size. On the other hand, Figure B.20 indicates that the p-values increase as we increase the sample size. This is a similar conclusion to what we got in Section B.2.2.

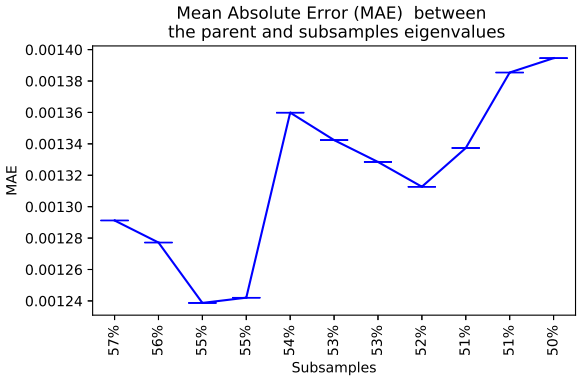


FIGURE B.19: MAE for uniform subsampled data

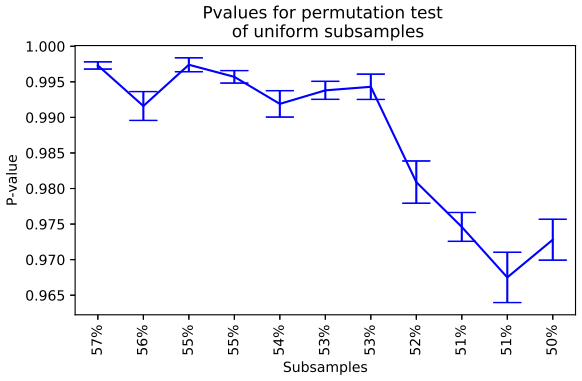


FIGURE B.20: Permutation p-values for uniform subsampled data

Finally, in Figure B.21 and Figure B.22 the classification accuracy results for LDA and SVM, respectively are reported and the results show that the accuracy for both LDA and SVM lies between 93.36 % and 93.52 % at 95% confidence interval.

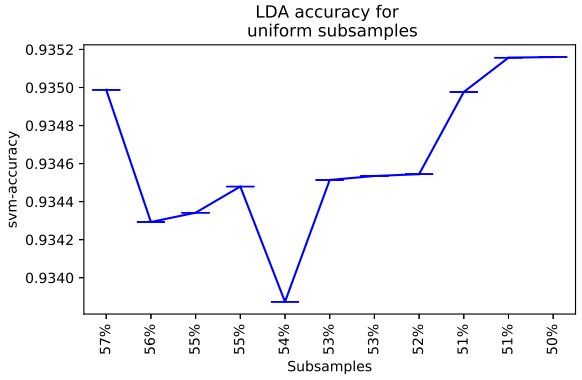


FIGURE B.21: LDA accuracy for uniform subsampled data

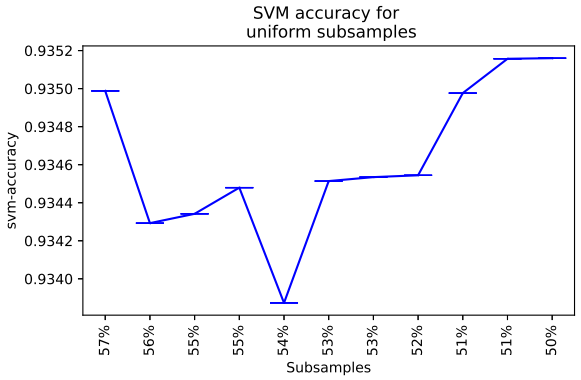


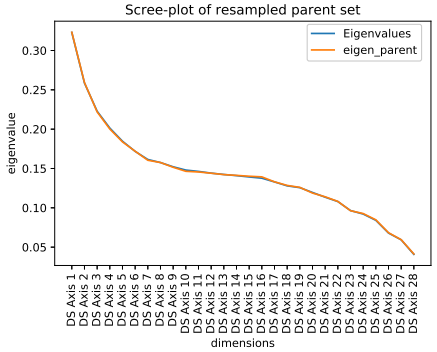
FIGURE B.22: SVM accuracy for uniform subsampled data

B.2.5 Experiment 4: Resampling vs parent data (GHS)

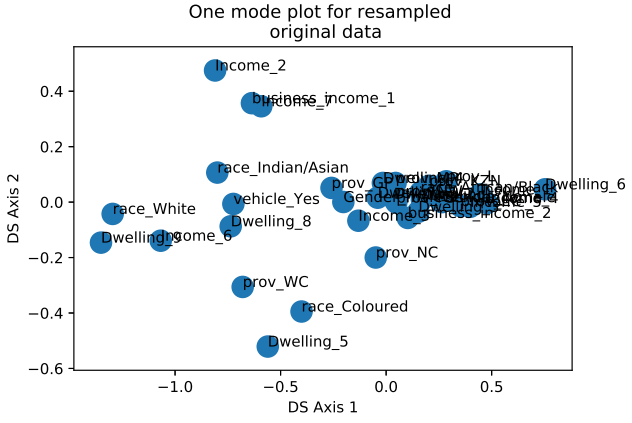
.

In this experiment, we resampled the GHS parent data with replacement and the results retained are as follows: Scree plot of eigenvalues, One mode plot and classification results (LDA and SVM), respectively.

Figure B.23a, shows that the eigenvalues are distributed the same way as the eigenvalues of the parent data. Figure B.23b show similar groupings of categories as those in the parent data (see Figure B.1b).



(A) Eigenvalues distribution of resampled parent data(GHS)



(B) One-mode plot

FIGURE B.23: A scree plot and one-mode plot for resampled parent data (GHS)

Considering Table B.5 we examined classification of this resampled parent data and the results shows similar results to what we got in Section B.2.2 with LDA have higher accuracy and precision values are equal for both LDA and SVM.

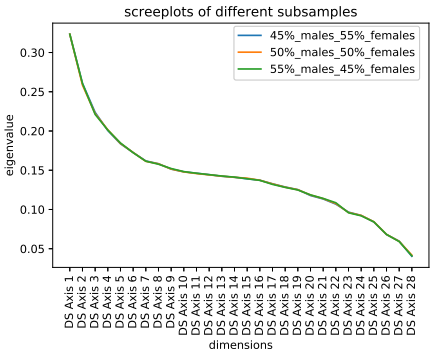
TABLE B.5: Classification results for resampled parent data (GHS)

	svm_accu	svm_preci	svm_reca	lda_accu	lda_preci	lda_reca
Resampled parent data	0.935483	0.970739	0.55431	0.935476	0.970649	0.55431

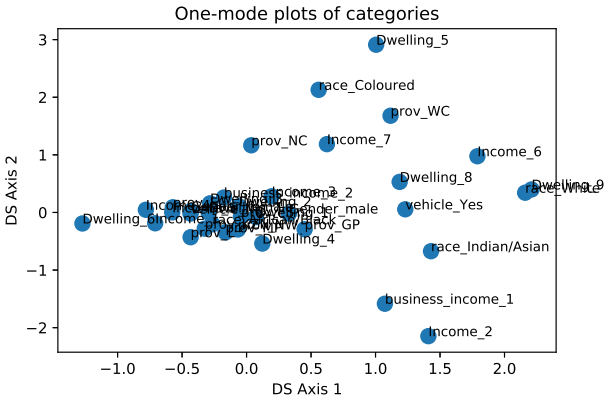
B.2.6 Experiment 5: Resampling vs Biased subsampling

In this experiment, we resampled the biased subsampled datasets of the GHS data and the results retained are as follows: Scree plot of eigenvalues, One mode plot, MAE, Permutation tests p-values and classification results (LDA and SVM), respectively.

Figure B.24a shows the distribution of eigenvalues for resampled biased subsamples. The eigenvalues are very similar throughout all subsamples and do not deviate from the distribution of the parent data eigenvalues. In Figure B.24b we see the one-mode plot of a resampled 50/50 subsample. The groupings in Figure B.24b are similar to those found in Section B.2.3.



(A) Eigenvalues distribution of resampled biased subsamples



(B) One-mode plot

FIGURE B.24: A scree plot and one-mode plot for 50/50 resampled biased subsampled data

The hierarchical clustering was applied to the resampled parent data and the results are reported in Figure B.25. The results show two clusters, one these clusters have about four sub-clusters. The similarities between categories are not altered by the proximity of clusters which means that DS is able to withstand changes in the minor changes in the data.

The results of MAE as reported in Figure B.26 agree with the results obtained in Section B.2.3, although it may seem that there is a decreasing trend in the values of MAE, the 50/50 subsample preserves the properties of a full parent data. We also looked at the

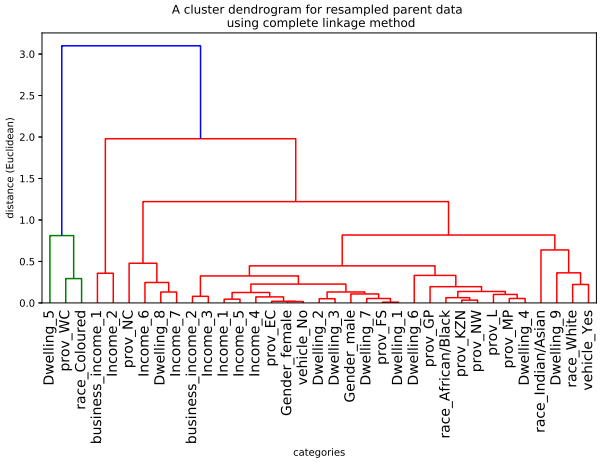


FIGURE B.25: Hierarchical clustering results for resampled parent data

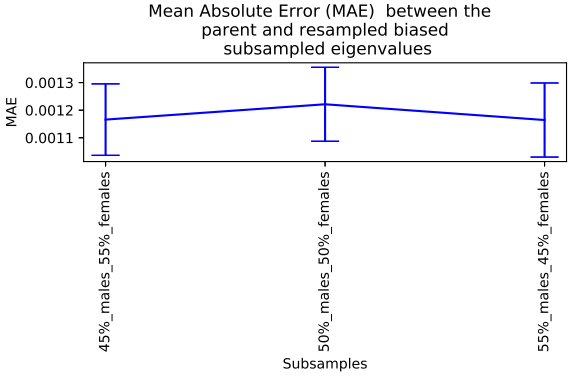


FIGURE B.26: MAE for resampled biased subsampled data

p-values of the permutation test in Figure B.27 and the results show some increasing trend. It is worth noting that these results are exactly what experiment attained.

.

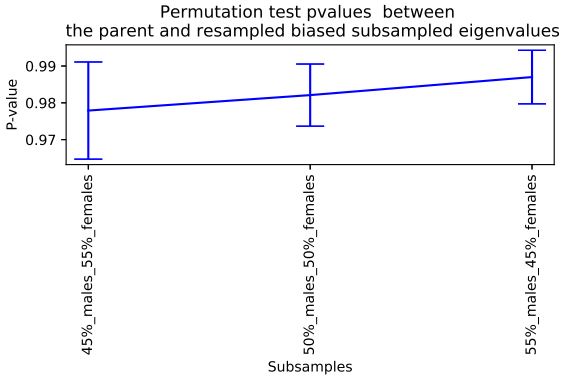


FIGURE B.27: Permutation p-values for resampled biased subsampled data

Finally we looked at the LDA and SVM in Figure B.28 and Figure B.29, respectively. The results of LDA in Figure B.28 together with accuracy results of SVM in Figure B.29 show and increase in the p-values.

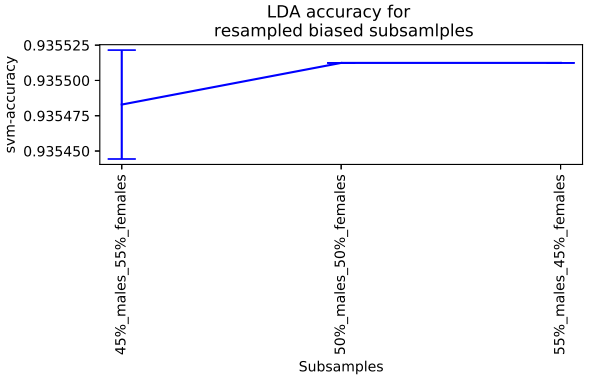


FIGURE B.28: LDA accuracy for resampled biased subsampled data

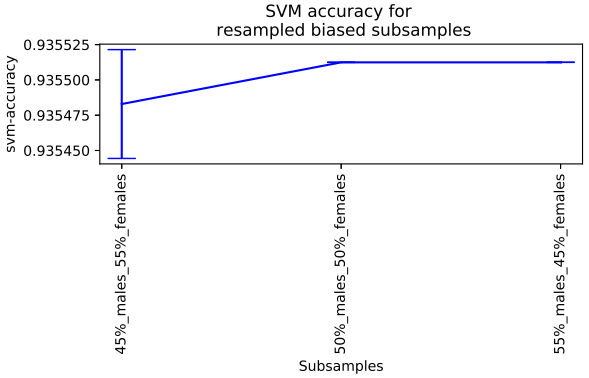
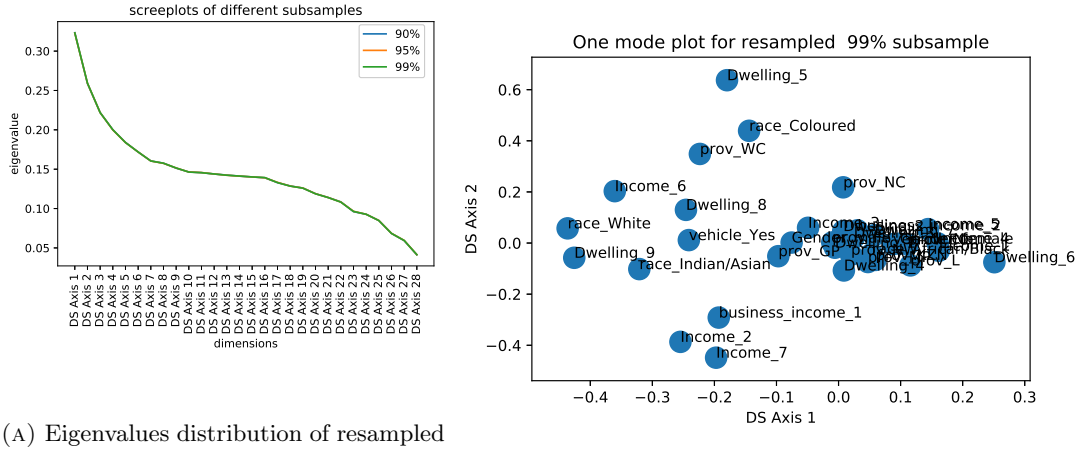


FIGURE B.29: SVM accuracy for resampled biased subsampled data

B.2.7 Experiment 6: Resampling vs Uniform subsampling

In this experiment, we resampled the subsampled datasets of the GHS t data and the results retained are as follows: Scree plot of eigenvalues, One mode plot, MAE, Permutation tests p-values and classification results (LDA and SVM), respectively.

Figure B.30a Shows the distribution of eigenvalues for three subsamples that were resampled, the distribution resembles that of the parent data in Figure B.1b. Figure B.30b shows one-mode plot of the resampled 99 % data. The groupings of Figure A.7 are much similar to the grouping in Section B.2.2.



(A) Eigenvalues distribution of resampled uniform subsamples

(B) One-mode plot resampled 99 % subsample

FIGURE B.30: A scree plot and one-mode plot for 99 % resampled uniform subsampled data

The resampled uniform subsampled data was also subjected to the hierarchical clustering and the results are reported in Figure B.31. The hierarchical clustering resulted in 2 clusters and these results are similar to the results obtained for normal (without resampling) uniform subsampling in Figure B.6. Thus, DS proves stability under uniform resampling.

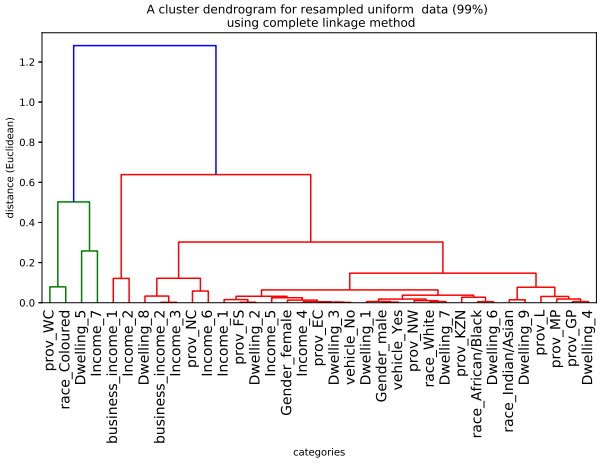


FIGURE B.31: Hierarchical clustering of resampled uniform subsampled GHS data

Considering the MAEs in Figure B.32, we notice that the 95 % subsampled data is more similar to the parent data than all other subsamples since it has the lowest MAE. Secondly, the permutation test p-values in Figure B.33 an increasing trend, the p-value of 97.25% is associated with the 95 % subsample while the highest p-value is associated with the subsample of 90 %. It is worthwhile to note that at 95 % the p-values seems to converge.

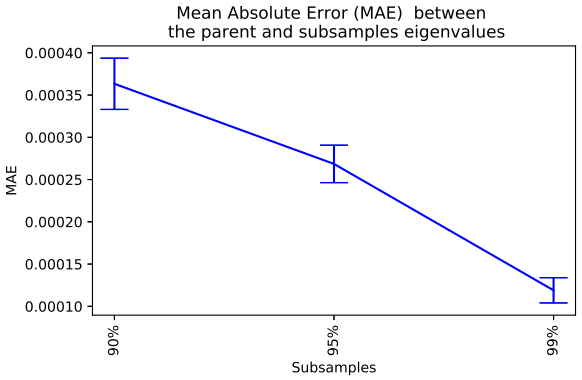


FIGURE B.32: MAE for resampled uniform subsampled data

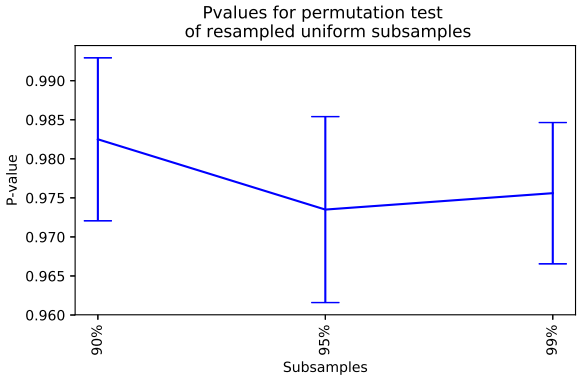


FIGURE B.33: Permutation p-values for resampled uniform subsampled data

Finally, we looked at the accuracy of LDA and SVM and the results are reported in Figure B.34 and Figure B.35, respectively. The accuracy decreases with an increase in the sample size. Although, this trend may be evident, the values of accuracy for LDA are still within the range reported in Section B.2.2, that is, between 93.25 % and 93.55 %.

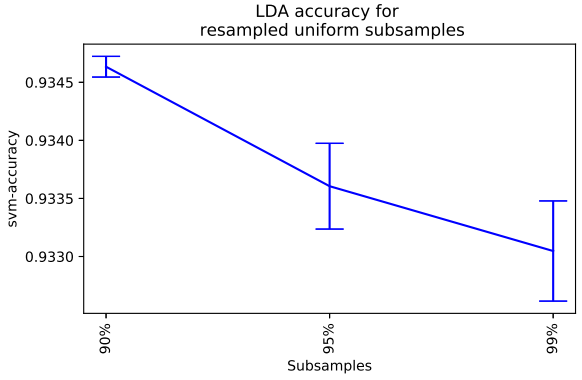


FIGURE B.34: LDA accuracy for resampled uniform subsampled data

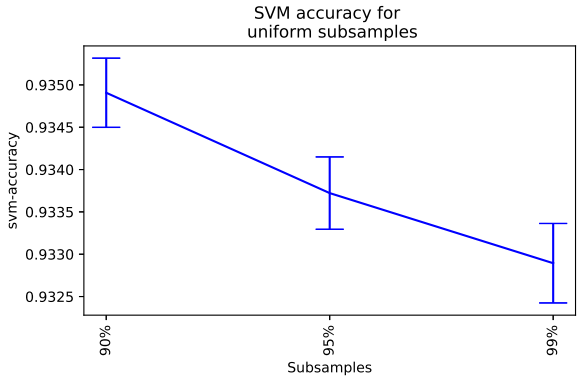


FIGURE B.35: SVM accuracy for resampled uniform subsampled data

B.2.8 Summary of results

We have applied the experiments of our study to real data that was obtained from STATS SA. The Dual scaling results are consistent throughout, from experiment 1 to experiment 6. This is some indication that dual scaling is indeed stable, the groupings remain the same on all of the undertaken experiments. The results of classification, in particular, accuracy was between 93 % and 97 % in this data, it remained between 93 % and 94 % on the subsamples. This is appealing, it means LDA and SVM can effectively classify the real multiple-choice datasets.

Appendix C

Similarity and Distance Measures

C.1 Cosine Similarity of the original parent data

In this section, the cosine similarity measure was further extended for the remaining categories and the results are reported in Figure C.1. These similarities are grouped by the initial groups obtained in the one-mode plot reported in Table 4.2.

C.2 Euclidean distance of the parent data

In this section, we present the remaining Euclidean distances of column weights for the parent data. These distances are grouped by the initial groups obtained in the one-mode plot reported in Table 4.2.

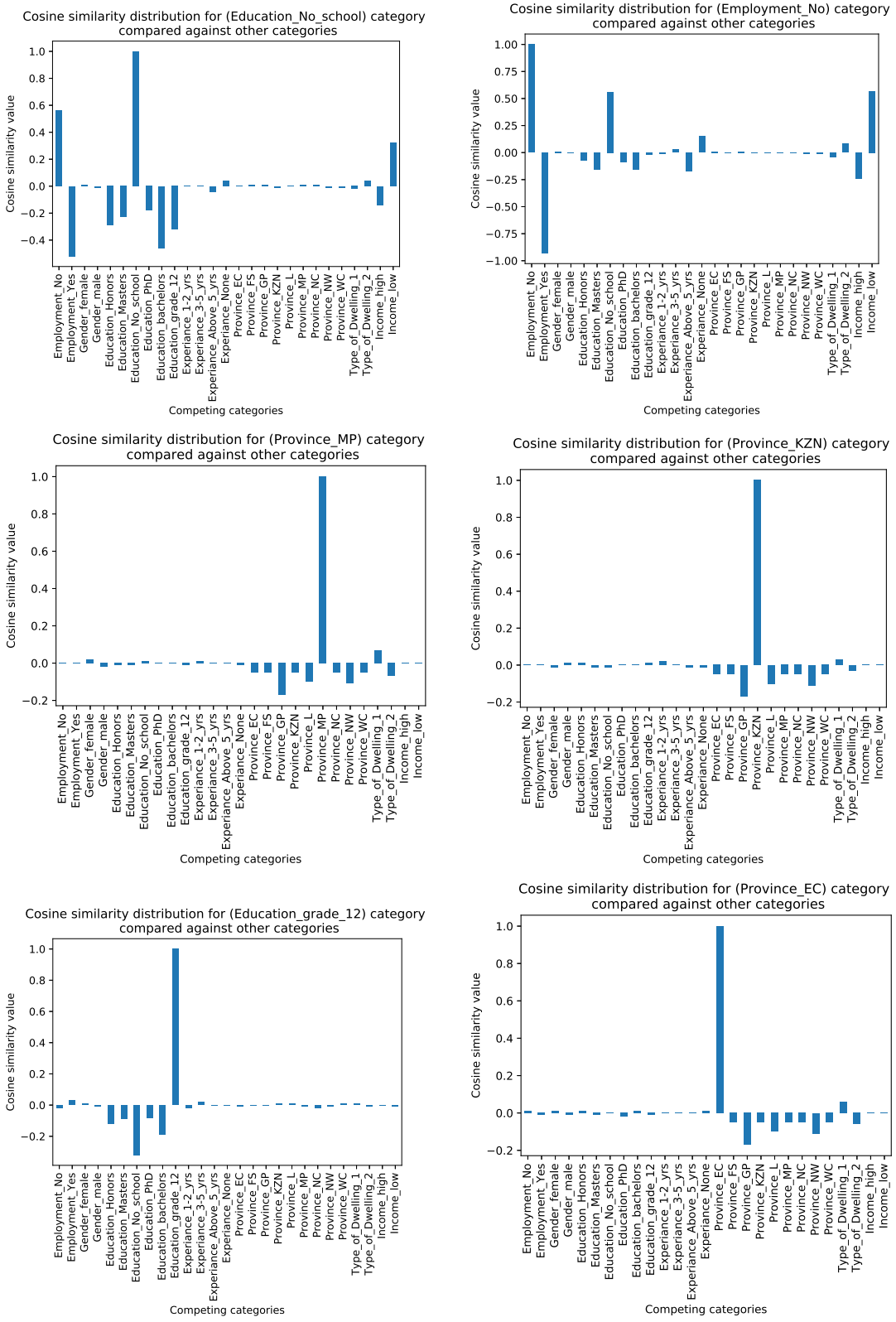


FIGURE C.1: Detailed cosine similarity results for the original parent data

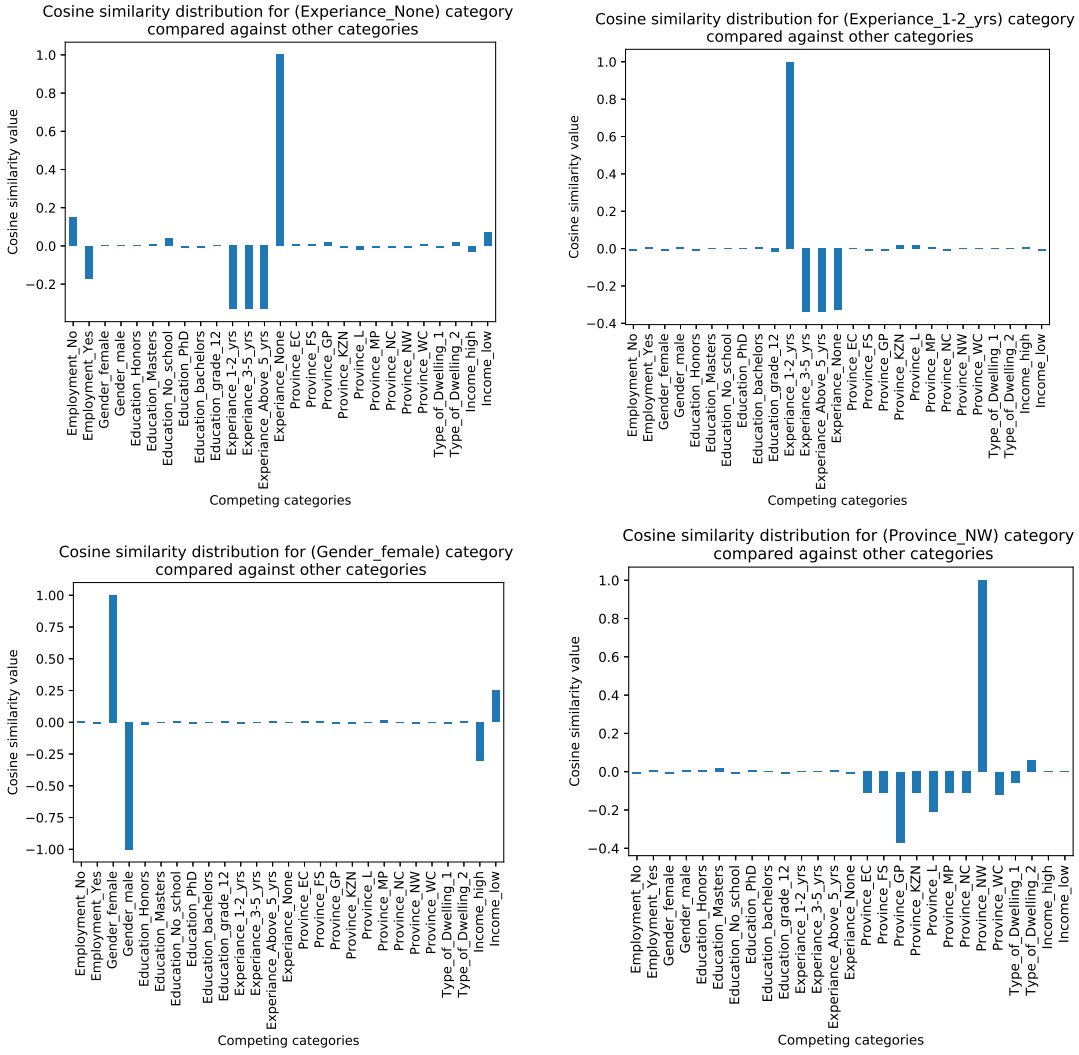


FIGURE C.2: Detailed cosine similarity results for Group 1 categories in Table 4.2 the parent data (continuation...)

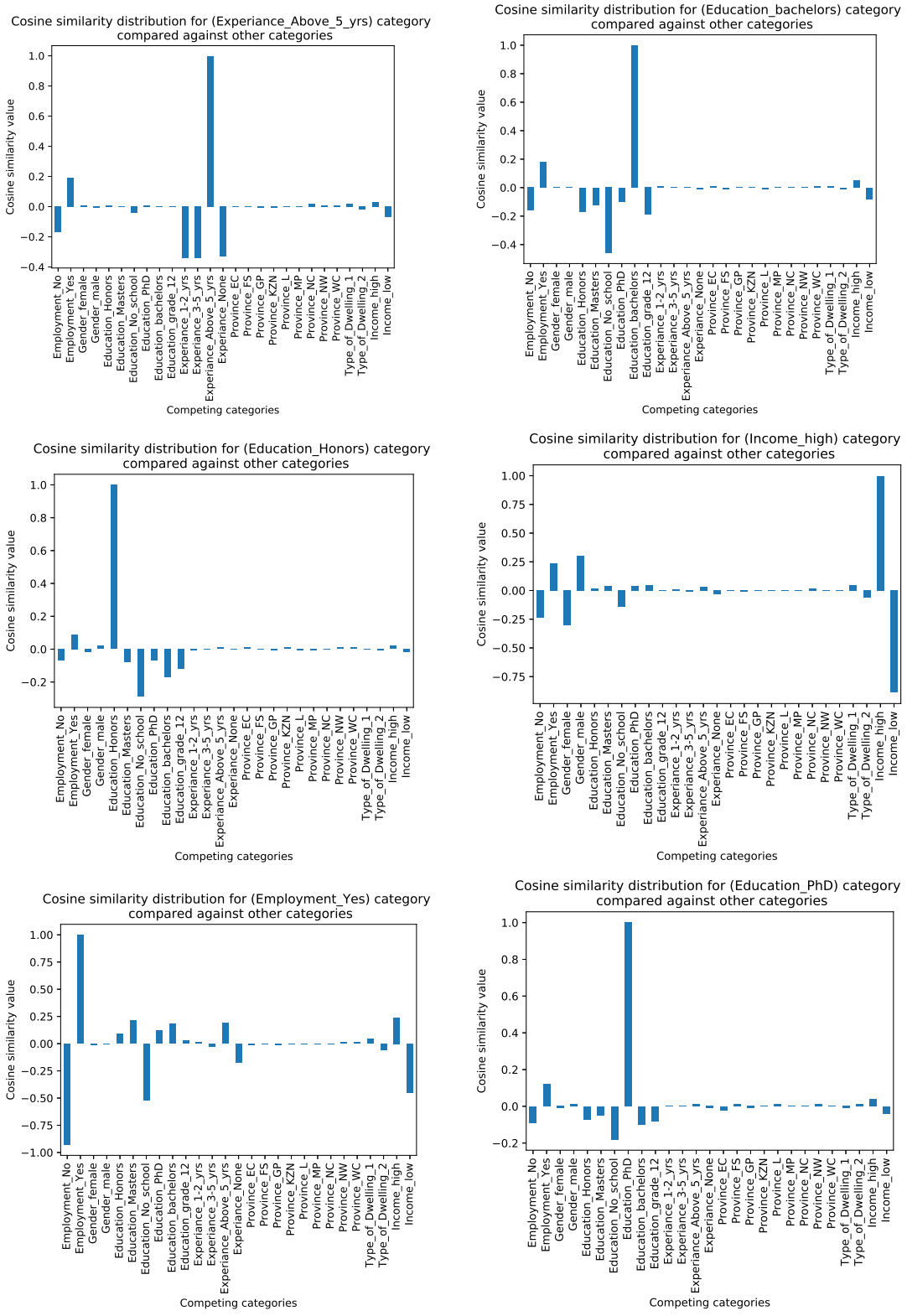
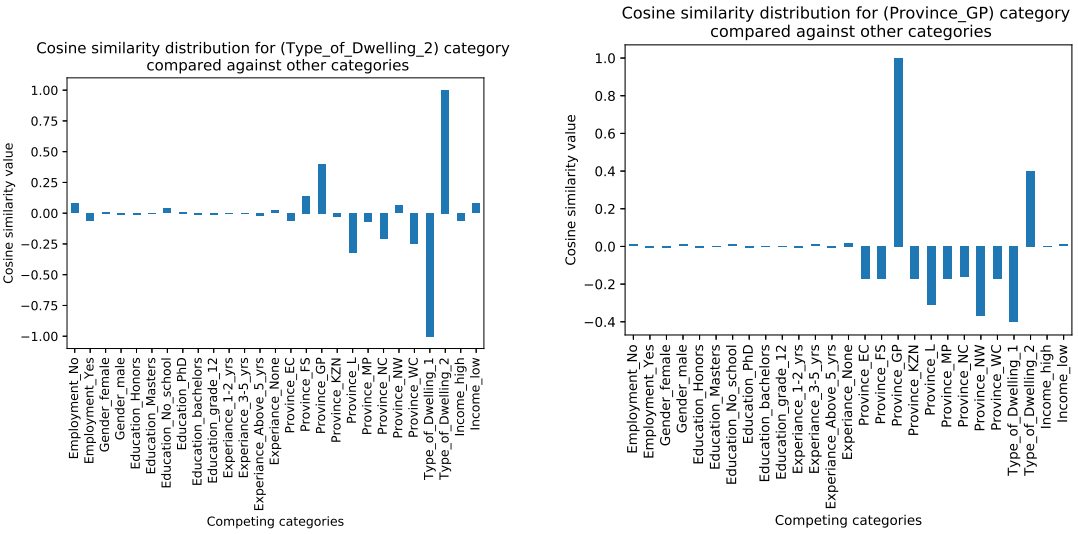


FIGURE C.3: Detailed cosine similarity results for the categories in Group 4 in Table 4.2 based on the original parent data



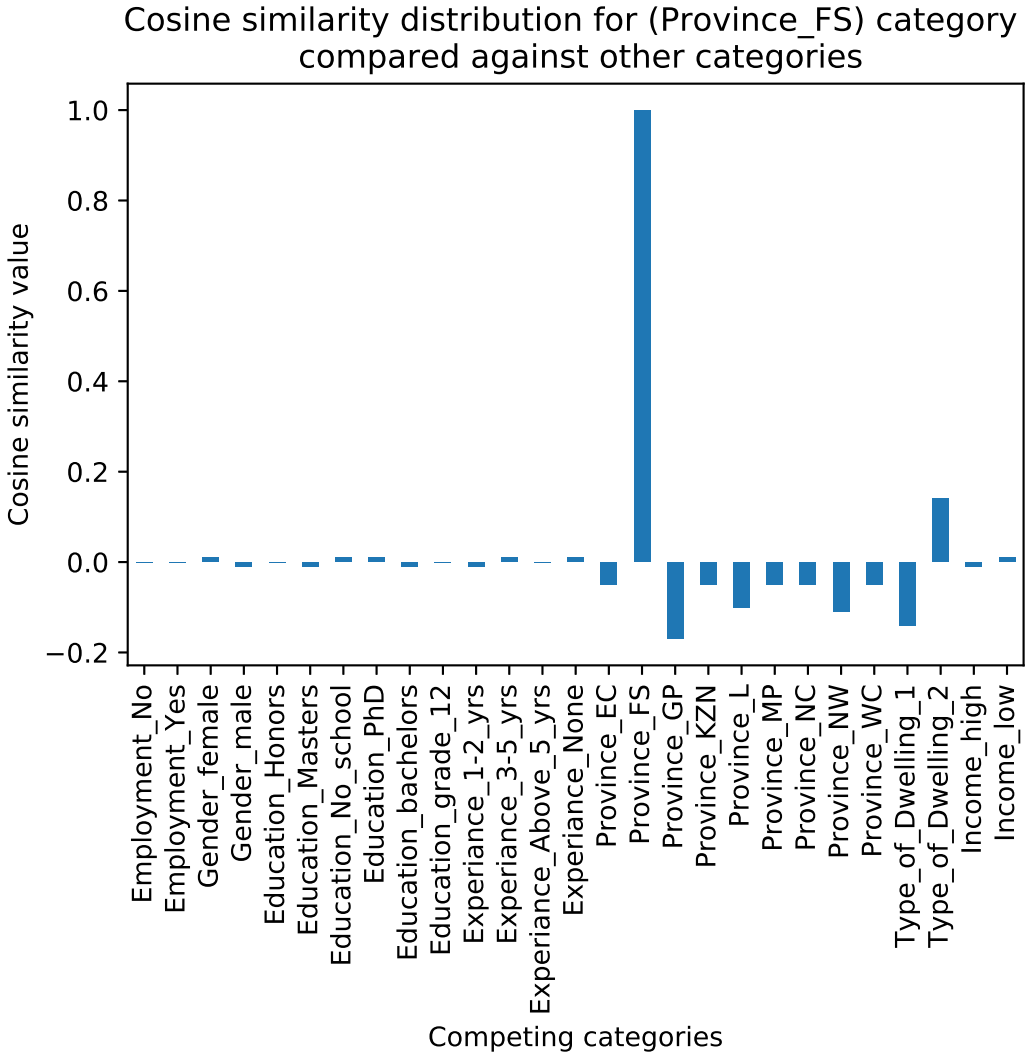


FIGURE C.4: Cosine similarity results for the categories in Group 3 in Table 4.2

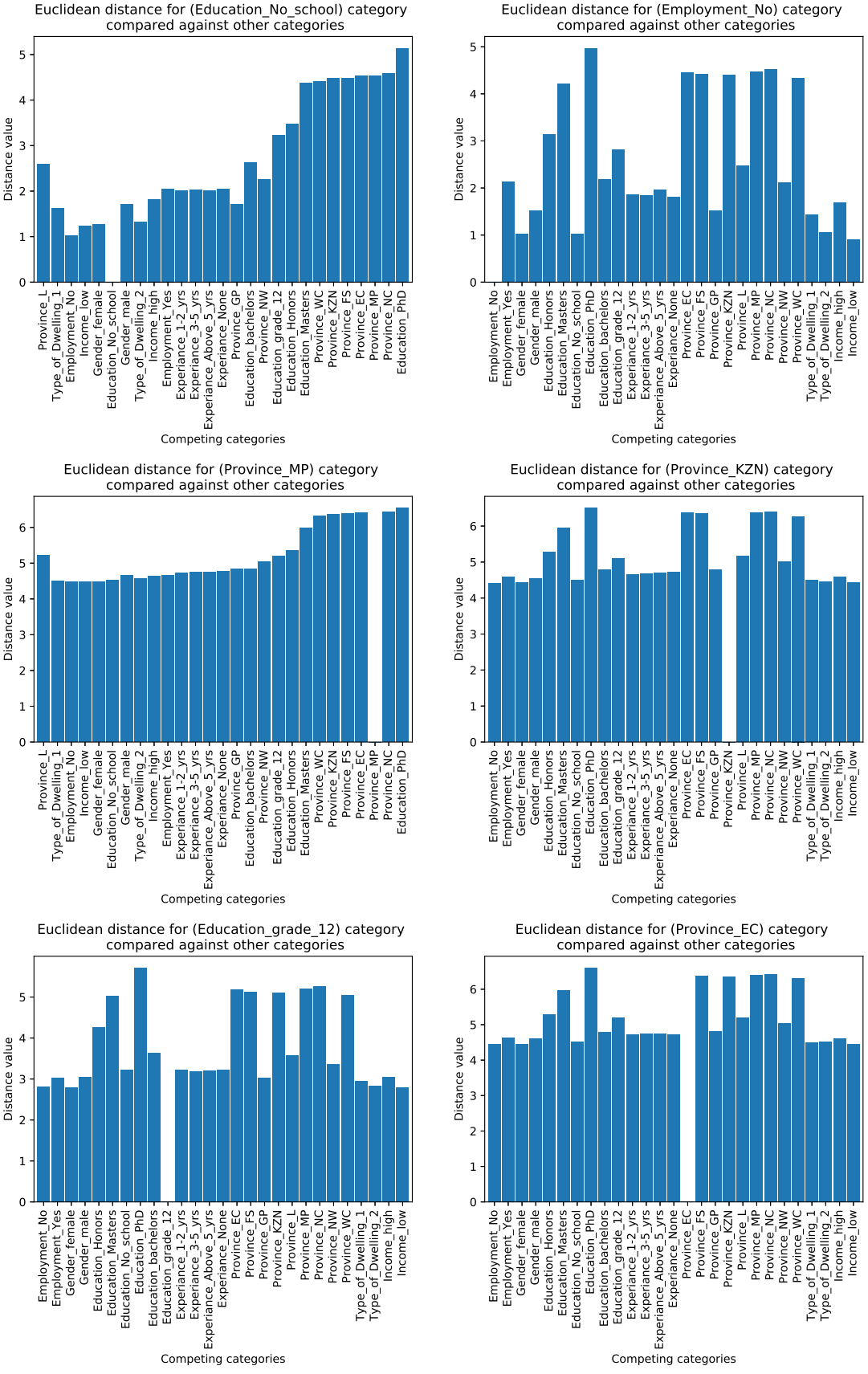


FIGURE C.5: Detailed Euclidean distance results for the original parent data

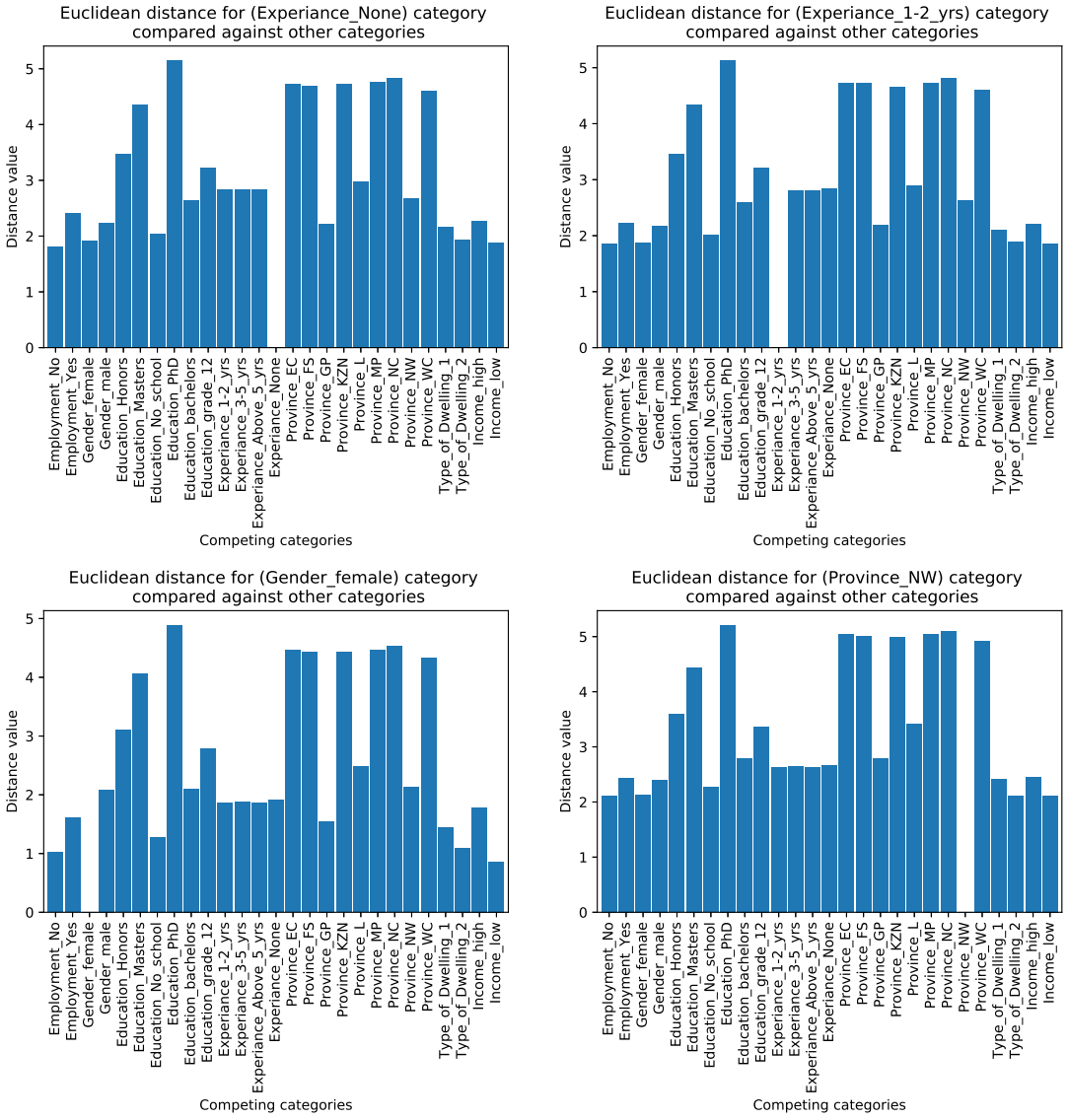


FIGURE C.6: Detailed Euclidean distance results for the parent data (continuation...)

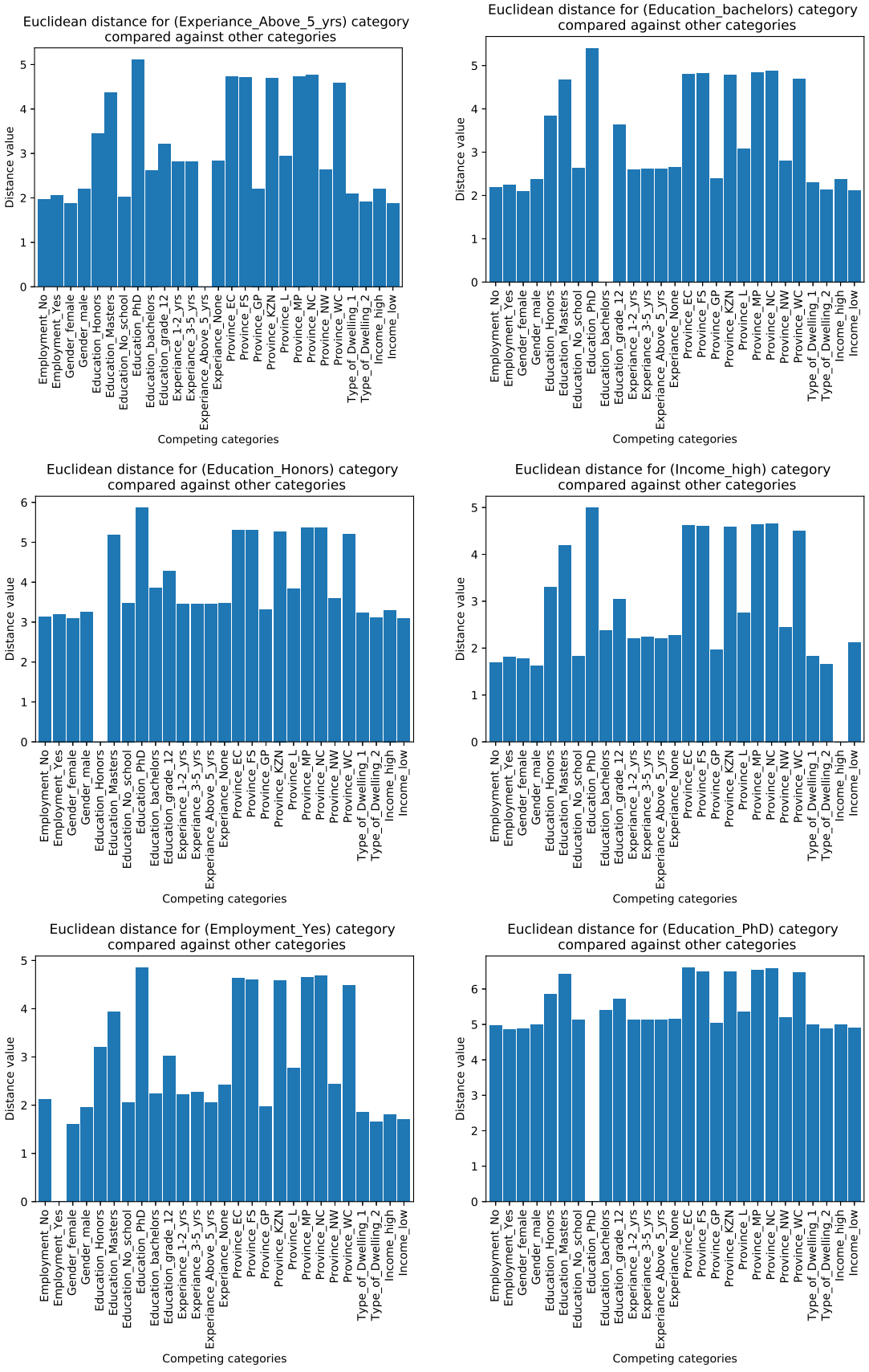


FIGURE C.7: Detailed Euclidean distance results associated with Group 4 categories for the original parent data (see Table 4.2)

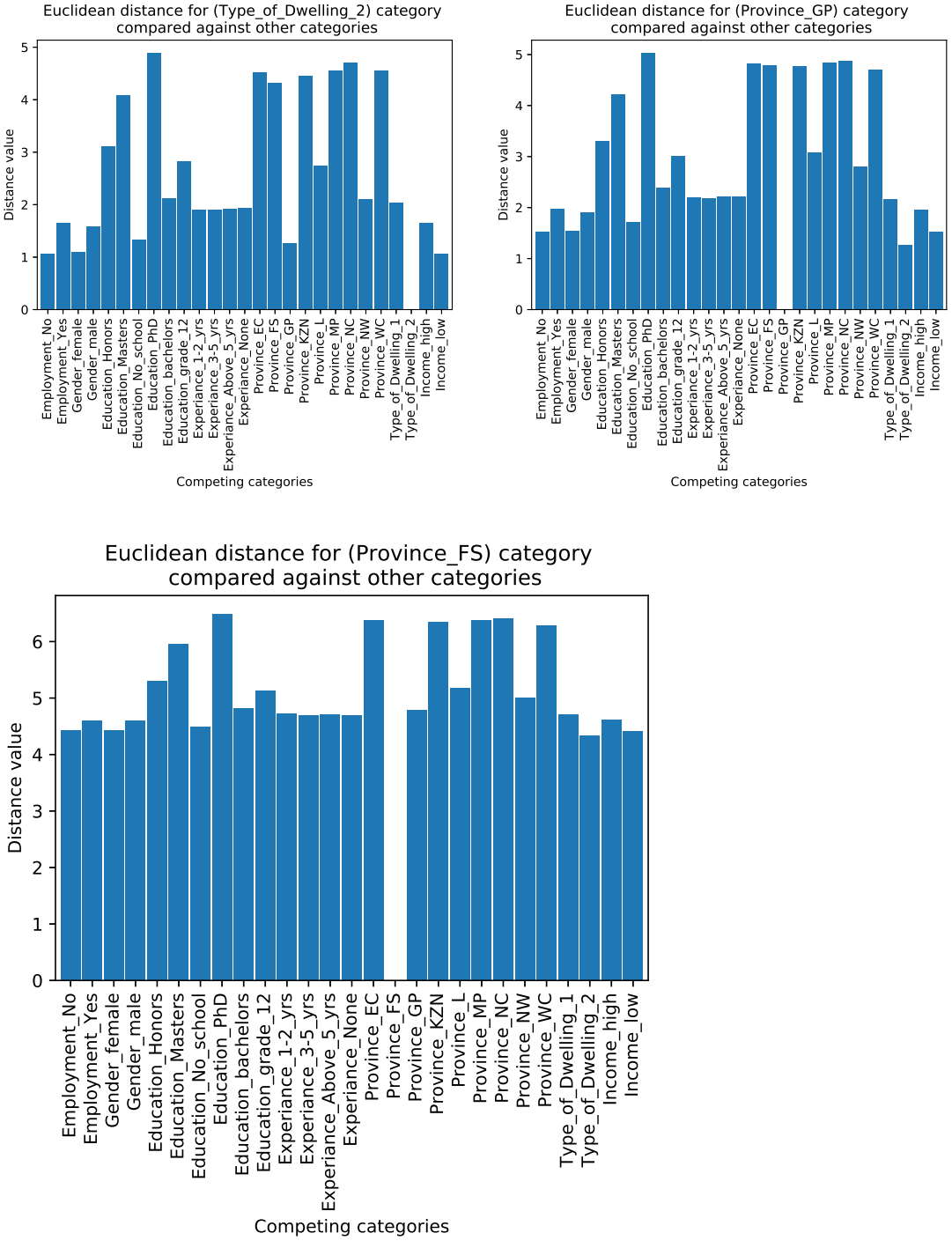


FIGURE C.8: Euclidean distance results associated with Group 3 categories in Table 4.2

Bibliography

Abdi, H. and Béra, M. (2017), *Correspondence Analysis*, Springer New York.

Aggarwal, C. C. and Sathe, S. (2017), *Outlier Ensembles: An Introduction*, Springer.

Anandan, V. and Uthra, G. (2017), 'Extension of topsis using l1 family of distance measures', *Advances in Fuzzy Mathematics* **12**(4), 897–908.

Ayele, D., Zewotir, T. and Mwambi, H. (2014), 'Multiple correspondence analysis as a tool for analysis of large health surveys in african settings', *African health sciences* **14**(4), 1036–1045.

Bakk, Z. and le Roux, N. J. (2017), 'Visualizing latent class models with analysis-of-distance biplots', *Sociological Methodology* **47**(1), 345–378.

Balbi, S. (1992), 'On stability in non-symmetrical correspondence analysis using bootstrap', *Statistica Applicata* **4**(4), 543–552.

Blasius, J. and Greenacre, M. (2006), 'Correspondence analysis and related methods in practice', *Multiple correspondence analysis and related methods* pp. 3–40.

Boriah, S., Chandola, V. and Kumar, V. (2008), Similarity measures for categorical data: A comparative evaluation, *in* 'Proceedings of the 2008 SIAM International Conference on Data Mining', SIAM, pp. 243–254.

Cha, S.-H. (2007), 'Comprehensive survey on distance/similarity measures between probability density functions', *City* **1**(2), 1.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., Initiative, A. D. N. et al. (2012), 'Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images', *Neuroimage* **60**(1), 59–70.

Clavel, J. G. and Nishisato, S. (2012), Reduced versus complete space configurations in total information analysis, *in* 'Challenges at the Interface of Data Analysis, Computer Science, and Optimization', Springer, pp. 91–99.

Connelly, L. M. (2016), 'Fisher's exact test', *MedSurg Nursing* **25**(1), 58–60.

Costantini, P., Linting, M. and Porzio, G. C. (2010), 'Mining performance data through nonlinear pca with optimal scaling', *Applied Stochastic Models in Business and Industry* **26**(1), 85–101.

De Koker, N., Viljoen, C. and Day, P. (2020), 'Risk-optimal sampling for reliability-based design', *Structural Safety* **83**, 101896.

Di Franco, G. (2016), 'Multiple correspondence analysis: one only or several techniques?', *Quality & Quantity* **50**(3), 1299–1315.

Efron, B. (1992), *Bootstrap methods: another look at the jackknife*, Springer.

Engel, D., Hüttenberger, L. and Hamann, B. (2012), A survey of dimension reduction methods for high-dimensional data analysis and visualization, *in* 'OASIcs-OpenAccess Series in Informatics', Vol. 27, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Fan, X. and Wang, L. (1996), 'Comparability of jackknife and bootstrap results: An investigation for a case of canonical correlation analysis', *The Journal of experimental education* **64**(2), 173–189.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S. and Ngo, L. H. (2012), 'Predicting sample size required for classification performance', *BMC medical informatics and decision making* **12**(1), 8.

Fisher, N. I. (1995), *Statistical analysis of circular data*, Cambridge University Press.

Fisher, R. A. (1940), 'The precision of discriminant functions', *Annals of Eugenics* **10**(1), 422–429.

Gaul, W. A. and Pfeifer, D. (2013), *From data to knowledge: Theoretical and practical aspects of classification, data analysis, and knowledge organization*, Springer Science & Business Media.

Gifi, A. (1990), 'Nonlinear multivariate analysis', *John Wiley and Sons. Chichester, England* .

Govender, P. and Sivakumar, V. (2020), 'Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)', *Atmospheric Pollution Research* **11**(1), 40–56.

Gower, J. C., Lubbe, S. G. and Le Roux, N. J. (2011), *Understanding biplots*, John Wiley & Sons.

Greenacre, M. (2006), Tying up the loose ends in simple, multiple, joint correspondence analysis, *in* 'Compstat 2006-Proceedings in Computational Statistics', Springer, pp. 163–185.

Greenacre, M. (2007), *Correspondence analysis in practice*, CRC press.

Greenacre, M. (2017), *Correspondence analysis in practice*, CRC press.

Greenacre, M. and Blasius, J. (2006), *Multiple correspondence analysis and related methods*, CRC Press.

Greenacre, M. and Hastie, T. (1987), 'The geometric interpretation of correspondence analysis', *Journal of the American statistical association* **82**(398), 437–447.

Greenacre, M. J. (1984), *Theory and applications of correspondence analysis*, Academic Press.

Guttman, L. (1941), 'The quantification of a class of attributes: A theory and method of scale construction', *The prediction of personal adjustment* pp. 319–348.

Haukoos, J. S. and Lewis, R. J. (2005), 'Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions', *Academic emergency medicine* **12**(4), 360–365.

Hayashi, C. (1950), 'On the quantification of qualitative data from the mathematico-statistical point of view', *Annals of the Institute of Statistical Mathematics* **2**(1), 35–47.

Hayashi, C. (1951), 'On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view', *Annals of the Institute of Statistical Mathematics* **3**(1), 69–98.

Hayashi, C., Yajima, K., Bock, H. H., Ohsumi, N., Tanaka, Y. and Baba, Y. (2013), *Data Science, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996*, Springer Science & Business Media.

Hazra, A. (2017), 'Using the confidence interval confidently', *Journal of thoracic disease* **9**(10), 4125.

Hesterberg, T. C. (2015), 'What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum', *The American Statistician* **69**(4), 371–386.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A. and Epstein, R. (2005), 'Bootstrap methods and permutation tests', *Introduction to the Practice of Statistics* **5**, 1–70.

Hua, J., Xiong, Z., Lowey, J., Suh, E. and Dougherty, E. R. (2004), 'Optimal number of features as a function of sample size for various classification rules', *Bioinformatics* **21**(8), 1509–1515.

Johansson Fernstad, S. (2011), 'Algorithmically guided information visualization: Explorative approaches for high dimensional, mixed and categorical data'.

Jolliffe, I. (2002), 'Outlier detection, influential observations, stability, sensitivity, and robust estimation of principal components', *Principal Component Analysis* pp. 232–268.

Jose, J. and Scholar, P. G. (2014), 'A Survey on Feature Selection Techniques', **2**(5), 44–46.

Kalhori, H., Alamdari, M. M. and Ye, L. (2018), 'Automated algorithm for impact force identification using cosine similarity searching', *Measurement* **122**, 648–657.

Kaplan, D. (2004), *The Sage handbook of quantitative methodology for the social sciences*, Sage Publications.

Knox, R. G. and Peet, R. K. (1989), 'Bootstrapped ordination: a method for estimating sampling effects in indirect gradient analysis', *Plant Ecology* **80**(2), 153–165.

Koller, D. and Friedman, N. (2009), 'Probabilistic graphical models: Principles and techniques-adaptive computation and machine learning'.

Lebart, L. (2006), 'Validation techniques in multiple correspondence analysis', *Multiple correspondence analysis and related methods* pp. 179–195.

Leeuw, J. d. (1973), 'Canonical analysis of categorical data', *Unpublished Ph. D. Thesis, University of Leiden, the Netherlands* .

Linting, M., Meulman, J. J., Groenen, P. J. and van der Kooij, A. J. (2007), 'Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap.', *Psychological methods* **12**(3), 359.

Linting, M. and van der Kooij, A. (2012), 'Nonlinear principal components analysis with CATPCA: a tutorial', *Journal of personality assessment* **94**(1), 12–25.

Lockyear, K. (2013), 'Applying bootstrapped correspondence analysis to archaeological data', *Journal of Archaeological Science* **40**(12), 4744–4753.

Lombardo, R., Ringrose, T. and Beh, E. J. (2012), 'Bootstrap confidence regions in classical and ordered multiple correspondence analysis', *Analysis and Modeling of Complex Data in Behavioural and Social Sciences* p. 53.

Maraun, M. D., Slaney, K. and Jalava, J. (2005), 'Dual scaling for the analysis of categorical data', *Journal of personality assessment* **85**(2), 209–217.

Markos, A., Menexes, G. and Papadimitriou, T. (2009), 'Multiple correspondence analysis for" tall" data sets', *Intelligent Data Analysis* **13**(6), 873–885.

Markus, M. T. (1994), *Bootstrap confidence regions in nonlinear multivariate analysis*, Vol. 28, DSWO Press, Leiden University.

McHugh, M. L. (2013), 'The chi-square test of independence', *Biochemia medica* **23**(2), 143–149.

Michailidis, G. and de Leeuw, J. (1998), 'The gifi system of descriptive multivariate analysis', *Statistical Science* pp. 307–336.

Mondal, A. and Mandal, A. (2020), 'Stratified random sampling for dependent inputs in monte carlo simulations from computer experiments', *Journal of Statistical Planning and Inference* **205**, 269–282.

Mooney, C. Z. (1996), 'Bootstrap statistical inference: Examples and evaluations for political science', *American Journal of Political Science* pp. 570–602.

Murtagh, F. and Contreras, P. (2017), 'Algorithms for hierarchical clustering: an overview, ii', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **7**(6), e1219.

Nenadic, O. and Greenacre, M. (2007), 'Correspondence analysis in r, with two-and three-dimensional graphics: The ca package', *Journal of Statistical Software* **20**(3).

Nguyen, H., Moayedi, H., Jusoh, W. A. W. and Sharifi, A. (2019), 'Proposing a novel predictive technique using m5rules-pso model estimating cooling load in energy-efficient building system', *Engineering with Computers* pp. 1–10.

Nishisato, S. (1978), 'Optimal scaling of paired comparison and rank order data: An alternative to guttman's formulation', *Psychometrika* **43**(2), 263–271.

Nishisato, S. (1980), *Analysis of Categoncal Data: Dual Scaling and Its Applications*, JSTOR.

Nishisato, S. (1993), 'On quantifying different types of categorical data', *Psychometrika* **58**(4), 617–629.

Nishisato, S. (1995), 'Graphical representation of quantified categorical data: its inherent problems', *Journal of statistical planning and inference* **43**(1), 121–132.

Nishisato, S. (1996), 'Gleaning in the field of dual scaling', *Psychometrika* **61**(4), 559–599.

Nishisato, S. (1998), Exploring multidimensional quantification space, *in* 'Data Science, Classification, and Related Methods', Springer, pp. 441–451.

Nishisato, S. (2006), *Multidimensional nonlinear descriptive analysis*, CRC Press.

Nishisato, S. (2014), *Elements of dual scaling: An introduction to practical data analysis*, Psychology Press.

Nishisato, S. and Clavel, J. G. (2009), 'Total information analysis: Comprehensive dual scaling', *Behaviormetrika* **37**(1), 15–32.

Nishisato, S. and Hillsdale, N. (1994), 'Elements of dual scaling: an introduction to practical data analysis', *Applied Psychological Measurement* pp. 379–382.

Nishisato, S. and Sheu, W.-J. (1980), 'Piecewise method of reciprocal averages for dual scaling of multiple-choice data', *Psychometrika* **45**(4), 467–478.

Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H. and Ralescu, A. (2017), 'Confusion-matrix-based kernel logistic regression for imbalanced data classification', *IEEE Transactions on Knowledge and Data Engineering* **29**(9), 1806–1819.

Patil, V. H., Singh, S. N., Mishra, S. and Donavan, D. T. (2008), 'Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one'criterion', *Journal of Business Research* **61**(2), 162–170.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M. and Stamatakis, A. (2009), How many bootstrap replicates are necessary?, *in* 'Annual International Conference on Research in Computational Molecular Biology', Springer, pp. 184–200.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M. and Stamatakis, A. (2010), 'How many bootstrap replicates are necessary?', *Journal of computational biology* **17**(3), 337–354.

Popovici, V., Chen, W., Gallas, B. D., Hatzis, C., Shi, W., Samuelson, F. W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T. et al. (2010), 'Effect of training-sample size and classification difficulty on the accuracy of genomic predictors', *Breast Cancer Research* **12**(1), R5.

Rebentrost, P., Mohseni, M. and Lloyd, S. (2014), 'Quantum support vector machine for big data classification', *Physical review letters* **113**(13), 130503.

Saporta, G. and Keita, N. N. (2006), 'Correspondence analysis and classification', *Multiple correspondence analysis and related methods* pp. 371–392.

Scutari, M., Vitolo, C. and Tucker, A. (2018), 'Learning bayesian networks from big data with greedy search: Computational complexity and efficient implementation', *arXiv preprint arXiv:1804.08137* .

Singh, P. (2017), 'Distance and similarity measures for multiple-attribute decision making with dual hesitant fuzzy sets', *Computational and Applied Mathematics* **36**(1), 111–126.

Sordo, M. and Zeng, Q. (2005), 'On sample size and classification accuracy: a performance comparison', *Biological and medical data analysis* pp. 193–201.

StatsSA (2017), 'General household survey 2017', *Statistics South Africa, Pretoria, South Africa* .

Takane, Y. (1994), 'Book Review: Elements of Dual Scaling: An Introduction to Practical Data Analysis by Shizuhiko Nishisato Hillsdale NJ: Eribaum, 1994, 381 pp., ', *Applied Psychological Measurement* **18**(4), 379–382.

Takane, Y. and Hwang, H. (2006), 'Regularized multiple correspondence analysis', *Multiple correspondence analysis and related methods* pp. 259–279.

Torres, A. and Greenacre, M. (2002), 'Dual scaling and correspondence analysis of preferences, paired comparisons and ratings', *International Journal of Research in Marketing* **19**(4), 401–405.

van de Velden, M., D'Enza, A. I. and Palumbo, F. (2017), 'Cluster correspondence analysis', *Psychometrika* **82**(1), 158–185.

van de Velden, M. and Neudecker, H. (2000), 'On an eigenvalue property relevant in correspondence analysis and related methods', *Linear Algebra and its Applications* **321**(1), 347–364.

Venugopal, D. and Gogate, V. (2012), On lifting the gibbs sampling algorithm, *in* 'Advances in Neural Information Processing Systems', pp. 1655–1663.

Wang, B., Liu, X., Yu, B., Jia, R. and Gan, X. (2019), 'An improved wifi positioning method based on fingerprint clustering and signal weighted euclidean distance', *Sensors* **19**(10), 2300.

Xu, L. (1999), Bootstrap for dual scaling of rankings, PhD thesis, National Library of Canada.

Yu, C. H. (2003), 'Resampling methods: concepts, applications, and justification', *Practical Assessment, Research & Evaluation* **8**(19), 1–23.

Zhu, S., Wu, J., Xiong, H. and Xia, G. (2011), 'Scaling up top-k cosine similarity search', *Data & Knowledge Engineering* **70**(1), 60–83.