CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Cluster sampling in large oral health surveys – issues and implications for design and analysis
Authors	Sheerin, Anthony Joseph
Publication date	2019-10-09
Original Citation	Sheerin, A. J. 2019. Cluster sampling in large oral health surveys – issues and implications for design and analysis. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2019, Anthony Joseph Sheerin. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-16 07:32:53
Item downloaded from	https://hdl.handle.net/10468/10109



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Cluster Sampling in Large Oral Health Surveys – Issues and Implications for Design and Analysis

Anthony Joseph Sheerin
110311727

Thesis submitted for the degree of Doctor of Philosophy



NATIONAL UNIVERSITY OF IRELAND, CORK

Department of Statistics

October 2019

Head of Department: Dr. Michael Cronin

Supervisors: Dr. Michael Cronin
Dr. Mairead Harding

Research supported by Health Research Board (CARG/2012/34)

Contents

List of Figures iv
List of Tables viii
Abbreviations xiii
Abstract xiv
Acknowledgements xv

1 Introduction **1**
1.1 Motivation 1
1.2 Survey Sampling Methods 1
1.3 Standard Error Calculation 4
1.3.1 Taylor Series Linearisation (TSL) 4
1.3.2 Replication Methods 5
1.3.3 Literature Review 7

2 Comparing Variance Estimations in Cluster Samples **10**
2.1 Outline 10
2.2 Motivation 11
2.3 Simulation Data and Procedure 12
2.3.1 Data 12
2.3.2 Implementation 14
2.4 Results 18
2.4.1 Normal Distribution 19
2.4.2 Poisson ($\lambda \sim U(3.5, 6.5)$) Distribution 21
2.4.3 Poisson ($\lambda \sim U(0.5, 1.5)$) Distribution 22
2.4.4 Lognormal Distribution 23
2.5 Concluding Remarks 25

3 Exploratory Analyses of Factors Associated With Design Effects Less Than One **26**
3.1 Outline 26
3.2 Motivation 26
3.3 Design Effect Calculations 27
3.3.1 The Survey Design 27
3.3.2 Aims 28
3.3.3 Notation 29
3.3.4 Methods for Mean and Variance Estimation 30
3.3.5 Some Peculiarities in the Data 32
3.3.6 Design Effect Results 33
3.4 Identifying Factors Associated With Design Effects Less Than One 42
3.4.1 Cluster Analysis 43
3.4.1.1 Types of Measure 44
3.4.1.2 Partitioning Clustering 45
3.4.1.3 Hierarchical Clustering 46
3.4.1.4 Clustering Results 47
3.4.2 Linear Model Analyses 50

3.4.2.1 Linear Regression of dmft DE 50
3.4.2.2 Logistic Regression of dmft DE 52
3.4.2.3 Linear Regression of Caries Free DE 54
3.4.2.4 Logistic Regression of Caries Free DE 55
3.4.3 Concluding Remarks 57

4 Analysing the Effect of Sampling Fewer Clusters on Bias and SE of dmft Estimates 58
4.1 Outline 58
4.2 Motivation 58
4.3 Methods 59
4.4 Results 61
4.4.1 Tables of Percentage Change in SE and Bias by Reduction in Number of Fluoridated Schools Selected 61
4.4.2 Tables of Percentage Change in SE and Bias by Reduction in Number of Non-Fluoridated Schools Selected 75
4.4.3 Summary Figures of Percentage Change in SE and Bias by Reduction in Number of Schools Selected 83
4.5 Concluding Remarks 87

5 Discussion, Future Research and Conclusion 88
5.1 Chapter Two 88
5.2 Chapter Three 92
5.3 Chapter Four 94

A Choosing Number of Replications A1

B SE Estimates for $N = 60, 40$ and 20 B1
B.1 Normal Distribution B1
B.2 Poisson $\lambda \sim U(3.5, 6.5)$ Distribution B4
B.3 Poisson $\lambda \sim U(0.5, 1.5)$ Distribution B7
B.4 Lognormal Distribution B10

C Distributions of Oral Health Variables C1
C.1 Fluoridated dmft Observations C1
C.2 Non-Fluoridated dmft Observations C2
C.3 Fluoridated dmfs Observations C3
C.4 Non-Fluoridated dmfs Observations C4
C.5 Fluoridated BMI Observations C5
C.6 Non-Fluoridated BMI Observations C6

D Caries Free Design Effect Results D1

E Cluster Analysis Results E1
E.1 Fluoridated Results E1
E.1.1 Table Representation of Results E1
E.1.2 Results of k-means (Outliers Removed) E2
E.1.3 Results of Agglomerative Clustering E4

E.2 Non-Fluoridated Results E8
E.2.1 Results of k-means (Raw Data) E8
E.2.2 Results of k-means (Outliers Removed) E10
E.2.3 Results of Agglomerative Clustering E12

F Linear Model Diagnostics **F1**
F.1 Linear Regression Diagnostics (dmft) F1
F.2 Logistic Regression Diagnostics for dmft F6
F.3 Linear Regression Diagnostics (Caries Free) F7
F.4 Logistic Regression Diagnostics for Caries Free F12

G Reducing Number of Clusters Samples (Caries Free) **G1**

List of Figures

1.1 Commonly used Types of Sampling ($n = 8$ observations sampled under each scheme) 2

2.1 Summary of data generation & sampling scheme used 14
2.2 Largest SE for normal distribution across all m_i values for $N = 80, 60, 40$ and 20. 18
2.3 SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes. 19
2.4 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes 21
2.5 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes. 22
2.6 SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes. 23

3.1 Within Sum of Squares against Number of Clusters Used 48
3.2 Cluster Analysis of Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups 49

4.1 Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 5. 83
4.2 Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 8. 84
4.3 Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 12. 85
4.4 Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 15. 85
4.5 Change in Mean SE by Reduction in Number of Clusters Sampled at 1% Bias. 86

A.1 Largest SE for Poisson($\lambda \sim U(3.5, 6.5)$) distribution across all m_i values for $N = 80, 60, 40$ and 20. . . . . A1
A.2 Largest SE for Poisson($\lambda \sim U(0.5, 1.5)$) distribution across all m_i values for $N = 80, 60, 40$ and 20. . . . . A2
A.3 Largest SE for lognormal distribution across all m_i values for $N = 80, 60, 40$ and 20. . . . . A3

B.1 SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes for N=60 clusters. B1

B.2 SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes for N=40 clusters. B2
B.3 SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes for N=20 clusters. B3
B.4 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes for N=60 clusters. B4
B.5 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes for N=40 clusters. B5
B.6 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes for N=20 clusters. B6
B.7 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes for N=60 clusters. B7
B.8 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes for N=40 clusters. B8
B.9 SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes for N=20 clusters. B9
B.10 SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes for N=60 clusters. . . . B10
B.11 SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes for N=40 clusters. . . . B11
B.12 SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes for N=20 clusters. . . . B12

C.1 Histogram of dmft values for fluoridated observations. C1
C.2 Histogram of dmft values for non-fluoridated observations. . . . C2
C.3 Histogram of dmfs values for fluoridated observations. C3
C.4 Histogram of dmfs values for non-fluoridated observations. . . . C4
C.5 Histogram of BMI values for fluoridated observations. C5
C.6 Histogram of BMI values for non-fluoridated observations. . . . C6

E.1 Within Sum of Squares against Number of Clusters Used (Outliers Removed) E2
E.2 Cluster Analysis of Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups (Outliers Removed) E3
E.3 Agglomerative Clustering Dendrogram (Ward Linkage) E4
E.4 Agglomerative Clustering Dendrogram (Average Distance Linkage) E5

E.5 Agglomerative Clustering Dendrogram (Average Distance Linkage) [k=5 Groups] E6
E.6 Plot of Clusters by Principal Components One and Two for the Average Distance Linkage using Agglomerative Clustering . . . E7
E.7 Within Sum of Squares against Number of Clusters Used E8
E.8 Cluster Analysis of Non-Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups E9
E.9 Within Sum of Squares against Number of Clusters Used (Outliers Removed) E10
E.10 Cluster Analysis of Non-Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups (Outliers Removed) E11
E.11 Agglomerative Clustering Dendrogram (Ward Linkage) E12
E.12 Agglomerative Clustering Dendrogram (Average Distance Linkage) E13
E.13 Agglomerative Clustering Dendrogram (Average Distance Linkage) [k=4 Groups] E14
E.14 Plot of Clusters by Principal Components One and Two for the Average Distance Linkage using Agglomerative Clustering . . . E15

F.1 Leverage values from model of $log(DE)$ on n. F1
F.2 Residual values from model of $log(DE)$ on n. F2
F.3 Cook's distance values from model of $log(DE)$ on n. F2
F.4 Scatter plot of $log(DE)$ vs n for dmft. F4
F.5 Plot of residuals vs fitted values for model of $log(DE)$ on n for dmft. F4
F.6 Histogram of residuals for model of $log(DE)$ on n for dmft. . . . F5
F.7 Normal probability plot of residuals for model of $log(DE)$ on n for dmft. F5
F.8 Index plot of Pearson residuals for model $logit(p)$ on $\bar{m}$ for dmft F6
F.9 Plot of Pearson Residuals vs Linear Predictor for model $logit(p)$ on $\bar{m}$ for dmft F6
F.10 Residual values from model of $log(DE)$ on Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{\bar{M}}$ for caries free. F7
F.11 Residual values from model of $log(DE)$ on Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{\bar{M}}$ for caries free. F7
F.12 Residual values from model of $log(DE)$ on Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{\bar{M}}$ for caries free. F8
F.13 Scatter plot of $log(DE)$ vs $FluoridationStatus$ for Caries Free. F9
F.14 Scatter plot of $log(DE)$ vs $\frac{n}{N}$ for Caries Free. F9
F.15 Scatter plot of $log(DE)$ vs $\frac{\bar{m}}{\bar{M}}$ for Caries Free. F10
F.16 Plot of residuals vs fitted values for model of $log(DE)$ on n for Caries Free. F10
F.17 Histogram of residuals for model of $log(DE)$ on n for Caries Free. F11
F.18 Normal probability plot of residuals for model of $log(DE)$ on n for Caries Free. F11
F.19 Index plot of Pearson residuals for model $logit(p)$ on n for Caries Free F12

F.20 Plot of Pearson Residuals vs Linear Predictor for model $logit(p)$ on n for Caries Free F12

List of Tables

3.1 Mean and Variance Estimates of dmft for Fluoridated Observations Age 5 . . . 34
3.2 Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 5 . . . 35
3.3 Mean and Variance Estimates of dmft for Fluoridated Observations Age 8 . . . 36
3.4 Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 8 . . . 37
3.5 Mean and Variance Estimates of dmft for Fluoridated Observations Age 12 . . . 38
3.6 Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 12 . . . 39
3.7 Mean and Variance Estimates of dmft for Fluoridated Observations Age 15 . . . 40
3.8 Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 15 . . . 41
3.9 Initial form of DE Values for Cluster Analysis . . . 47
3.10 Simple Linear Regression Estimates for $log(DE_{(dmft)})$ . . . 52
3.11 Logistic Regression Estimates for $log(DE_{(dmft)})$ . . . 53
3.12 Simple Linear Regression Estimates for $log(DE_{(Caries)})$ . . . 55
3.13 Logistic Regression Estimates for $log(DE_{(Caries)})$ . . . 56
3.14 Final Models and Significant Univariate Parameters for DE's less than one . . . 56

4.1 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5 . . . 62
4.2 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5 . 64
4.3 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8 . . . 66
4.4 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8 . 68
4.5 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12 . . . 69
4.6 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12 71
4.7 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15 . . . 72
4.8 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15 74

4.9 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5 75
4.10 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5 76
4.11 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8 77
4.12 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8 78
4.13 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12 79
4.14 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12 80
4.15 Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15 81
4.16 Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15 82

D.1 Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 5 . . . . . D1
D.2 Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 8 . . . . . D2
D.3 Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 12 . . . . . D3
D.4 Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 15 . . . . . D4
D.5 Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 5 . . . . . D5
D.6 Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 8 . . . . . D6
D.7 Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 12 . . . . . D7
D.8 Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 15 . . . . . D8

E.1 CCA DE values and assigned cluster indicator. E1

G.1 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5 . . . . . G2

G.2 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5 . . . . . G3
G.3 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8 . . . . . G4
G.4 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8 . . . . . G5
G.5 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12 . . . . . G6
G.6 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12 . . . . . G7
G.7 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15 . . . . . G8
G.8 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15 . . . . . G9
G.9 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5 . . . . . G10
G.10 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5 . . . . . G11
G.11 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8 . . . . . G12
G.12 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8 . . . . . G13
G.13 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12 . . . . . G14
G.14 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12 . . . . . G15
G.15 Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15 . . . . . G16
G.16 Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15 . . . . . G17

I, Anthony Joseph Sheerin, certify that this thesis is my own work and has not been submitted for another degree at University College Cork or elsewhere.

Anthony Joseph Sheerin

"You'll be something someday." - Sr. Kathleen Moore

Abbreviations

DE - Design Effect

- DEFF - Design Effect (Ratio of Variances)
- DEFT - Root Design Effect (Ratio of Standard Errors)
- $DEFT = \sqrt{DEFF}$

JK1 - Delete-one Jackknife

PSU - Primary Sampling Unit

SE - Standard Error

SRS - Simple Random Sample

SSU - Secondary Sampling Unit

TSL - Taylor Series Linearisation

Abstract

The use of survey sampling has become routine in almost all aspects of life. If the chosen sample is representative of the population, inferences about the population can be made from the sample. Chapter 1 reviews commonly used survey sampling techniques and discusses the variance estimation methods required for these techniques. Particular attention will be given to cluster sampling, as the data underlying this thesis was gathered in such a manner. The results of the literature review in Chapter 1 identify the direction of analysis for Chapter 2.

Chapter 2 compares the readily available cluster variance estimation methods, namely Taylor Series Linearisation (TSL) and the delete-one jackknife (JK1), on simulated finite populations of data with known characteristics, to ascertain if there are any situations where the estimates differ. Multiple situations are examined systematically, including skewed distributions, large sampling fractions and small sample sizes with these situations occurring both in isolation and simultaneously. These methods provided identical estimates when the sampling fraction was small, and the number of observations, particularly at second-stage, was large. However, when these conditions were violated, diverging estimates occurred. Moreover, design effects less than one are seen in this chapter which is unusual in a cluster sampling setting. Chapter 3 looks at using the above variance estimation techniques on a national oral health dataset.

Chapter 3 analyses a national oral health dataset, in which there are regions with design effects less than one. As the data was collected using cluster sampling, the presence of design effects less than one is extremely unusual and warranted an analysis to try and identify the possible causes. Both cluster analysis and linear model methods are used to identify variables which may indicate the presence of a design effect less than one, with the number of clusters sampled (n) being the most identified variable using different models. Chapter 3 suggests that reducing the number of clusters sampled will reduce the design effect of the data.

Chapter 4 looks at the effect of reducing the number of clusters sampled (n) on the SE and DE estimates of the survey data analysed in Chapter 3. Variance estimates are produced using a jackknife resampling approach, looking at all combinations when $n = 1, 2, \ldots, 10$ clusters are dropped in turn. The results show that a small number of clusters per community care area (CCA) can be dropped, with no impact on the bias of the estimate and a very small increase in the SE of the estimate.

Acknowledgements

Firstly I would like to express my sincere gratitude to my supervisors Dr. Michael Cronin and Dr. Mairead Harding for their help during my thesis. In particular I wish to thank Michael for his patience, advice and encouragement over the last three and a half years for which I am extremely grateful. His guidance and support have been invaluable throughout this thesis.

This project could not have been completed without the help of my family, particularly my parents. I am eternally grateful for their continuous support throughout my research. Their encouragement, optimism and reassurance helped me greatly in bringing this project to completion. Also a special thanks to my girlfriend Sorcha, who has been a pillar of support all the way through this thesis, always having a positive outlook on both the good and bad days. Lets hope the journey wasn't too much like work!

My friends in the lab (2.11) have been a great help, facilitating idea development and discussion, while also providing the occasional much needed distraction from research. I wish to thank you all for making the lab a great environment to work in each day. Also to my friends outside of the lab, a big thanks for your help in showing me there is more than just research to focus on.

I would like to thank all the staff in UCC who have helped me over the past nine years. UCC introduced me to the world of statistics and helped form my passion for the subject from undergraduate level. All the staff across the school have been a great help throughout this journey in both professional and social manners.

Finally, I wish to acknowledge the Health Research Board for their funding throughout this project. It is a privilege to receive a scholarship which allowed me to work full time on this thesis and without the support of the HRB this research would not be possible.

Chapter 1

Introduction

1.1 Motivation

The use of survey sampling has become routine in almost all aspects of life. Survey sampling is a general term for choosing a sample from a population. If the chosen sample is representative of the population, then the attributes of the samples will closely match those of the population, and inferences about the population can be made from the sample. The results of surveys can be used for small scale decisions, to national and international decisions. The topics discussed in this document stemmed from the analysis of a national study of oral health in Ireland, which frequently utilizes surveys and survey sampling methods. The initial chapter of this document aims to review commonly used survey sampling methods and discuss the methods of variance estimation required for these sampling methods. Particular attention will be given to cluster sampling, as the data underlying this document was gathered in such a manner. Finally this chapter compares the available literature on survey variance estimation methods and will outline the direction of analysis for Chapter 2.

1.2 Survey Sampling Methods

Survey sampling (Särndal, Swensson and Wretman, 2003) can be carried out in many different forms, including simple random sampling, stratified sampling, cluster sampling, systematic sampling and multi-stage sampling. During a simple random sample (SRS) each observation of the sample is chosen randomly, with

each observation of the population having the same probability of inclusion in the sample. Stratified sampling involves dividing the population into independent groups (strata) with a sample, usually a SRS, of observations being chosen from each strata. Cluster sampling involves dividing the population into independent groups (clusters), and initially taking a sample of these clusters (usually a SRS). Once a cluster is chosen all of the cluster observations are included in the sample. Systematic sampling involves choosing a starting observation (at random), and then sampling every n^{th} observation after the initial observation. Multi-stage sampling requires more than one stage of sampling, and can be a combination of the above methods or otherwise. Sampling can be carried out with or without replacement of the chosen observations.

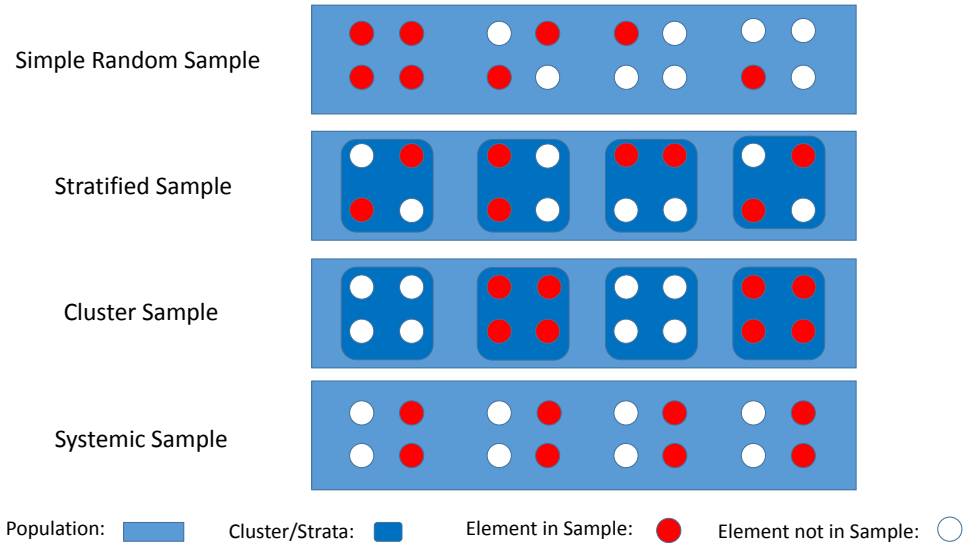


Figure 1.1: Commonly used Types of Sampling ($n = 8$ observations sampled under each scheme)

Survey sampling, and particularly the use of cluster sampling, is frequently used to collect data, as it is generally more cost efficient and it does not require a list of every observation in the population to be available (Särndal, Swensson and Wretman, 2003). In order to conduct cluster sampling, the researcher uses pre-existing distinct groups in the population (clusters) and then selects a sample, generally a simple random sample (SRS), of clusters from the population. In single-stage cluster sampling, once a cluster is chosen, all observations within the cluster are included in the sample. In multi-stage designs, a further stage(s) of sampling is used to select the observations that are included in the sample. The advantages of cluster sampling include time and cost efficiency along with being safer with

regards to data protection (only one venue set up needed), and sometimes it may be the only feasible approach (Eldridge and Kerry, 2012; Lesaffre et al., 2009). The motivation for this paper stemmed from an oral health study (Whelton et al., 2006), in which data is often gathered in clusters due to convenience (i.e. data is gathered from schools which are natural clusters and the individuals within these schools would be the observations of interest). Cluster sampling is used in many study designs outside of oral health, with the use of cluster sampling to survey children being quite common as children are naturally gathered in schools which provide a convenient frame for a sample to be gathered.

Formulating clusters requires the observations to be grouped together based on some pre-defined criteria, and as a result the observations may not be independent (Cornfield, 1978; Killip, Mahfoud and Pearce, 2004). Generally observations within clusters tend to be more alike than observations from different clusters and as a result correlation between the observation responses within a cluster is likely (Eldridge and Kerry, 2012; Wears, 2002). This correlation leads to a decrease in variation amongst the observations in the same cluster (i.e. variance of the within-cluster responses) and this decrease must be accounted for during analysis as it may affect the variance estimates associated with parameter estimates (Carle, 2009). Thus, there are specific variance estimation techniques developed for clustered data which ensure that the variance estimate for the sample is appropriate, and that the resulting confidence intervals and hypothesis tests are valid (Hannigan and Lynch, 2013). The most commonly used survey variance estimation methods are analytical formulae, Taylor Series Linearisation (TSL), Balanced Repeated Replication (BRR) and the Jackknife method.

Using the survey sample variance, calculated using one of the aforementioned methods, design effects can also be calculated. Two related definitions of the design effect are commonly used. The root design effect (DEFT) is the ratio of the survey (design based) standard error calculated compared to the standard error estimate as if the sample was a SRS with replacement. A further definition of design effect, generally referred to as the design effect (DEFF), is also used. This measure represents the ratio of the two variance estimates, the numerator being the survey variance estimate and the denominator representing the hypothetical, with replacement SRS variance estimate. If a study has a design effect less than one, it implies that the survey design is more efficient than the SRS design however, within the reviewed literature using cluster sampling, this result is uncommon. In the case of a clustered sample, a design effect less than one would imply large variation within the clusters (i.e. observations have a large range of

individual values in the clusters), but little variation between clusters (i.e. overall cluster estimates are similar), which is contrary to the belief of clusters having similar characteristics of interest. Design effect estimates are often used in the calculation of sample sizes required for surveys and also can be used as a proxy for measuring the efficiency of a survey sampling method, compared to a SRS.

1.3 Standard Error Calculation

Sampling textbooks (Rao, 2000; Cochran, 2007) present analytical formulae to estimate the standard error (SE) of a statistic for many different survey sample designs, and for simple statistics (e.g. means and totals), these formulae provide the best approach to standard error estimation. However, closed form expressions may not always exist for the SE of an estimator, due to the complexity of the sampling scheme used to gather the observations. The use of advanced weighting procedures and complex sampling schemes can make the variance calculation intractable, even for simple statistics, and thus no exact method need exist. The alternative is to use approaches which approximate the statistic of interest. Approximation of the standard error (SE) for survey sampled data occurs using two main techniques; simplifying assumptions and replication methods.

1.3.1 Taylor Series Linearisation (TSL)

TSL estimates a statistic of interest from a complex survey by approximating this statistic with a linear estimator. Thus, using the first order terms from an appropriate Taylor series expansion, the statistic of interest is approximated by linear forms of the observations (Eurostat, 2002). Higher order terms can be included, but these are usually omitted as a linear approximation is sufficient in most cases, excluding highly skewed populations. The inclusion of higher order terms is not readily available in standard software. TSL provides a method of expressing the statistic of interest in a linear form, and variance estimation techniques are then used to estimate the variance of the linear statistic, which represents the variance of the statistic of interest (which may be linear or non-linear) (Wolter, 2007). Many software packages use TSL as the default method of variance estimation, however, the packages will calculate the variance of a design using analytical formulae when the design is not so complex that a linear approximation is required.

TSL is seen as the gold standard when it comes to complex variance estimation, as it can be applied to general sample designs, and the theory is well established. However, when estimating statistics that are non-smooth (e.g. quantiles), or when the sample size is not large enough, TSL can be biased downwards (Lohr, 2009). The bias is considered negligible for simple statistics regardless of the sample size, and the bias essentially becomes zero when large samples are present (Särndal, Swensson and Wretman, 2003). However, to reduce the bias of complex statistics, large samples are needed. Furthermore, not all statistics can be expressed as smooth functions, and hence do not fit the framework required for estimation. TSL can also be difficult to implement with complex statistics involving weights (Cochran, 2007; Lohr, 2009; Särndal, Swensson and Wretman, 2003; Wolter, 2007).

1.3.2 Replication Methods

Replication methods (Eurostat, 2002) are capable of capturing complex sampling schemes due to their flexibility and robust nature. However, replication methods are generally more computationally intensive than the linearisation method. All replication methods follow a similar process, taking a (usually large) number of sub-samples/replicate samples from the original sample and use estimates of the variability of the statistic of interest from each of the sub-samples to estimate the variability of the full sample. Many replication methods, with many variations on each exist. The most common replication methods in the literature, which are also the default methods implementable in statistical software such as SAS™ V9.4 (SAS Institute, 2016) and R (R Core Team, 2016), are the balanced repeated replication (BRR) method and the jackknife method.

Balanced Repeated Replication (BRR)

The BRR method is applied to cluster designs where each cluster has exactly two final stage units. The BRR method involves selecting balanced half-samples from the original sample of data. Once the half sample is selected, the statistic of interest is calculated for the half-sample. The SE of the full sample is worked out based on the difference between the full-sample and half-sample estimates. The main advantage of the BRR method, compared to the jackknife method, is that it leads to asymptotically valid inferences for both smooth and non-smooth functions, however the design needed for the default BRR is very restrictive (Cochran,

2007; Lohr, 2009; Särndal, Swensson and Wretman, 2003; Wolter, 2007). Extensions of BRR to alternative designs have been developed, however these are not readily available in most software packages.

Jackknife Method

The jackknife method (Eurostat, 2002) can take many different forms, depending on the number of segments dropped at each replicate. This can be beneficial in complex study designs which include strata, clusters and other sampling methods. The delete-one jackknife (JK1) method is the most commonly used variation and is the most appropriate technique for unstratified designs. The JK1 method divides the population into segments, and one segment is dropped from the sample at a time. The statistic of interest is calculated from the incomplete sample, and this dropped segment is reintroduced into the sample. The process continues dropping one segment at a time and recalculating the statistic of interest. In a clustered sample, each cluster needs to be dropped one at a time, as it would be incorrect to drop single observations from within the clusters, as this would divide the cluster unit. The variance of the overall sample is calculated from the difference of the replicates compared to the original sample. The standard jackknife method is more applicable for with-replacement sampling designs. However, when the sampling fraction is small, it will provide a good estimate for without-replacement designs. The JK1 method can be inconsistent for non-smooth functions (e.g. quantiles), but it performs well for smooth functions and is widely applicable to most sampling designs (Cochran, 2007; Lohr, 2009; Särndal, Swensson and Wretman, 2003; Wolter, 2007).

Variance Estimation Methods Choice

Appropriate variance estimation methods must be used when data is gathered using survey sampling methods besides a SRS. The most frequently used survey variance estimation methods, which are also available in most statistical software, are BRR, TSL and JK1, and each offer their own advantages and disadvantages. Some of the requirements and assumptions for the survey methods were mentioned in Section 1.3, and a more technical description of the methods can be found in Section 2.3.2. The next section will present a review of previous work on the assumptions required for each survey variance estimation method, and summarise papers which compared the estimation techniques.

1.3.3 Literature Review

The literature review began from the working title of "quantifying and incorporating the cluster effect in oral health surveys". Initially appropriate methods of quantifying the cluster effect were identified. Oral health papers involving surveys, including surveys on fluoride use, were reviewed, to examine the commonly used methods in clustered oral health surveys. Initially the literature review was conducted in journal databases such as Scopus, Web of Science, JSTOR and Google Scholar, with a further review of the literature was also completed in search engines such as Google and Yahoo to bring up results that may be outside the reach of journal searches. Text books describing the appropriate methods were explored in a search of the library catalogue. Other works, often referenced in text books or journal papers were also examined for relevant content related to the literature search. Once the literature was compiled, a comparison of the three main survey variance estimation methods began, and further information was gathered on the assumptions of each of these methods, along with any papers which directly compared the estimation methods. A summary of the papers comparing survey variance estimation methods is provided below.

Rust (1985) provided a detailed review of the TSL, BRR and JK1 methods, their applicability and draws comparisons between the methods. In almost all of the studies Rust (1985) reviewed, the data was a stratified population with two primary sampling units (PSU's) per stratum. This survey design is a more restrictive design than desired in the context of this paper, however the results still give guidance on the use of survey methods. Lemeshow, Hosmer and Hislop (1980) found that in the presence of non-normality, estimators performed badly when only small sample sizes in three strata were used, but when the strata were increased to 16, or the second-stage sample was large, all three estimators performed comparably and satisfactorily across multiple distributions. In highly skewed populations, first order TSL estimates may be unreliable, though this is rarely an issue if the sample size is large (Wolter, 2007). Koop (1972) showed that to confirm convergence of the TSL method, taking only first order terms, that there should be at least five observations. Kish and Frankel (1970) showed an increase in the number of strata in surveys increased the precision of the standard error (SE) estimates. Kish and Frankel (1974) also showed that for ratio means and regression coefficients TSL consistently gave the smallest mean square error (MSE), but did not always have the smallest bias. Brillinger (1970) showed analytically that the JK1 method can provide a poor estimate for complex statistics

(i.e. median) being biased downwards, and this may extend to all methods which use a linear approximation for a non linear statistic. Ultimately, the review by Rust (1985) concludes that TSL, BRR and JK1 have little difference between them on theoretical grounds, and that practical matters should be considered in the choice of the variance estimation method used.

Kovar, Rao and Wu (1988) discuss the estimation methods with reference to non-linear statistics and medians in their paper, varying the number of strata observed. They suggest that for smooth functions, TSL and JK1 estimates (which are indistinguishable) are more accurate and stable than BRR. For non-smooth functions, the JK1 method can be inconsistent, but the BRR method performs well for variance estimates. Spitznagel (1985) compares the three variance estimation methods for rates of occurrence under extreme conditions. This paper supports the previous conclusion, that the estimates are almost identical under any of the three variance estimation methods. However Spitznagel (1985) also suggests the results may not hold for more complex statistics. It is noteworthy that in this paper the JK1 method performs very well, and this is attributed to the absence of stratification and the presence of large sample sizes. A paper by Paben (1999) further supports the earlier findings, suggesting that for larger samples there is little to no difference between the methods. It also backs up the fact that TSL produces the most stable estimates. Finally, the paper suggests that the JK1 method seemed to be the closest to the true value of the variance for smaller samples. Chowdhury (2013) suggests looking at the different stages of weighting on the estimate of variance (i.e. allowing for post-stratification and non-response weights) and shows that BRR allows these weights to be incorporated, while TSL does not. However, this paper suggests that this causes TSL to over-estimate, which can be desirable, if a source of error is overlooked. The JK1 method does not feature in this paper. Hammer, Shin and Porcellini (2003) compare TSL and JK1 variance estimation techniques using clustered data collected from a large scale survey based on a national probability sample. Their results indicate that under the two variance estimation methods, the difference in estimates was negligible, and had no impact on the results of the statistical tests carried out. However they do note the design of the data is relatively "simple" by the standards of complex sampling designs. Notably, Hammer, Shin and Porcellini (2003) observed a design effect less than one in their paper, which is uncommon under a cluster sampling design. They attribute this results to ”implicit stratification and systematic selection”.

Bruch, Münnich and Zins (2011) gave a detailed review of variance estimation, again presenting findings concurrent with those obtained previously in the literature. Attention is drawn to the BRR method leading to unstable variance estimates when the number of strata is small. Their paper also reiterates that the JK1 method can give inconsistent estimates if the estimator is non-smooth (e.g. quantiles). Furthermore, Bruch, Münnich and Zins (2011) implement a simulation study to see how the JK1 method (along with other variations of the Jackknife, and bootstrap methods) reacted in different simulated circumstances. Their simulation changes the variance in each cluster, the expected values of each cluster and the size of the PSU's in each cluster, with the simulated data being generated from a log-normal distribution. Bruch, Münnich and Zins (2011) also looked at the need to include the variation at the second-stage of sampling. The results show that there is "no cookbook approach" to variance estimation, and that the design of the survey should be considered hand in hand with the analysis before the data is collected.

Throughout the literature review, little information was available on using the survey variance estimation methods in the presence of a large sampling fraction. Furthermore, the definition of "small" or "large" is never estimated under controlled conditions, or set distribution assumptions. The influence of a small number of sample clusters, or a small second stage sample size are discussed, but again not when a large sampling fraction is present. Chapter 2 aims to compare how each of the variance estimation methods are effected by the aforementioned ideas, using a range of sampling fractions in an attempt to see if the definition of a large sampling fraction differs for each method. Furthermore Chapter 2 explores different distributions, number of sampling clusters and number of second stage observations in the presence of different sampling fractions to examine the different estimates produced by the TSL and JK1 methods. Finally Chapter 2 will try and assess if there are any situations where a cluster design can be more efficient than a SRS, and ascertain if this unusual behaviour can occur in dataset simulated to be similar to that of an oral health study.

Chapter 2

Comparing Variance Estimations in Cluster Samples

2.1 Outline

This chapter aims to compare the readily available cluster variance estimation methods, namely TSL and JK1, on simulated finite populations of data with known characteristics, to ascertain if there are any situations where the estimates produced differ. Multiple situations are examined systematically, including skewed distributions, large sampling fractions and small sample sizes, and to the best of the knowledge of the authors, no such extensive (simulation) study has been performed before. This chapter explores how these conditions affect estimators when they occur both in isolation and simultaneously in samples, and also looks at the effect of these conditions have on the design effect of a survey. These methods provided identical estimates when the sampling fraction was small, and the number of observations, particularly at second-stage, was large. However when these conditions were violated, the two methods showed diverging estimates. Estimates of when this divergence occurs are given using simulated variables representative of a typical oral health design. Moreover, design effects less than one are seen in this chapter which is unusual in a cluster sampling setting. The cause of these design effects less than one will be examined. Section 2.2 provides results from the existing literature, which were the motivation for this chapter. Section 2.3 provides a description of the data generated and the simulation conducted in this chapter. Section 2.4 and Section 2.5 present the results of the simulation, and discuss some concluding remarks that can be drawn from

this research.

2.2 Motivation

As seen from the literature review there is "no cookbook approach" to variance estimation, and the design of the survey should be considered hand in hand with the analysis before the data is collected. As a result this chapter looks to compare TSL and JK1 methods on a simulated finite population of data with known characteristics, to ascertain if there are any situations where the estimates differ. These situations include varying distributional assumptions, varying sample sizes and varying sampling fractions. This chapter will look at these, sometimes extreme, conditions occurring simultaneously as, while much is known about how these conditions affect variance estimators in isolation, there is little literature on how these conditions affect the estimator when they occur together. The TSL and JK1 methods are chosen as they are the default survey variance estimation procedures in both SAS (SAS Institute, 2016) and R (R Core Team, 2016), and are the most commonly featured methods in the literature. Although balanced repeated replication (BRR) is also a default in the two software packages, and arises frequently in the literature, this method is omitted due to the restrictive design needed. The standard BRR method requires two final stage units per cluster, which was not the form of study this chapter was looking to investigate.

Furthermore, this chapter looks to estimate values for the conditions under which TSL and JK1 estimates differ. The literature suggests that differences may occur when there is a "small" second-stage sample size or a "large" sampling fraction present. The definitions of small and large are not explicitly specified, and this chapter hopes to investigate these definitions. Although Murray (1998) suggests such a definition will never be possible in the general sense, it might be feasible to explain these differences with some restricted assumptions, namely a commonly used sampling plan, such as cluster sampling within schools. The restriction on class sizes and number of clusters may give some informative results on when divergence between the estimates occur.

Finally this chapter examines whether a cluster survey may ever have a design effect less than one, and looks to examine what may cause such a result. Although this would be an unusual result in terms of cluster sampling, the presence of skewed distributions, large sampling fractions and small sample sizes both at first and second stage may influence the design effect estimate.

2.3 Simulation Data and Procedure

2.3.1 Data

To ensure chosen population characteristics could be obtained, code was written to simulate populations in R (R Core Team, 2016) software. Each simulated population was a clustered population with $N = 80, 60, 40$ or 20 clusters. Initially each of the population clusters were simulated from a normal distribution, with a random mean and a fixed standard deviation. The mean of each cluster was generated from a uniform distribution defined on [3.5,6.5] which allowed each of the cluster means to vary from one another (i.e. $\mu_i \sim U(3.5, 6.5)$) with the standard deviation within each cluster (σ) having a fixed value of 1.5. The size of each of the clusters M_i was generated from a uniform distribution defined on [5,30] (i.e. $M_i \sim U(5, 30)$).

Further populations were generated allowing the distribution of the variable of interest to vary. Clusters were generated from two different Poisson distributions, and one lognormal distribution, representing skewed discrete and skewed continuous variables typically observed in a survey. Two Poisson distributions were generated with the Poisson parameters λ_i being obtained from a uniform distribution on the ranges $(3.5, 6.5)$ and $(0.5, 1.5)$ (i.e. $\lambda_i \sim U(3.5, 6.5)$ and $\lambda_i \sim U(0.5, 1.5)$) respectively. In the case of the lognormal distribution, the parameters of the lognormal distribution were chosen such that the expected value of the lognormal distribution $E[X] \sim U(3.5, 6.5)$ and that the standard deviation of the lognormal distribution was 1.5.

Each of the parameters values defined in the simulation were chosen to reflect what an observed oral health dataset may look like, possibly in the context of a school setting. The number of clusters varying from 20 to 80 represented the number of clusters that could be in a particular geographical area. The cluster sizes ranging from 5 to 30 represented the size of typical classrooms, with the large range to capture both urban and rural school classroom sizes. The normal distribution was chosen to represent a continuous symmetric distribution which could represent height or weight. The two versions of the Poisson were chosen to represent discrete, slightly skewed (e.g. decayed missing and filled surfaces (dmfs) in children) and a discrete, heavily skewed (decayed missing and filled teeth (dmft) in children) distributions respectively. The lognormal distribution represented a continuous skewed distribution, which could be used to represent

blood pressure in a survey situation. Thus distribution choices, along with the parameter estimates of the distributions, represented the common types of variables one would gather in a survey. Data gathered from a national oral health survey in Ireland is analysed in Chapter 3 and the distribution of the dmft, dmfs and BMI variables by age group and fluoridation status are presented in Appendix C. The use of multiple distributions also allowed comparison between discrete and continuous distributions with varying levels of skewness. It is noteworthy that in the context of oral health dmft can be calculated in different manners, and both DMFT and dmft refer to different levels of dentition (namely permanent and temporary teeth) associated with age. However for simplicity, dmft will be used for all age levels throughout this thesis, as the numeric values are mostly of interest in this research. If clinical interpretation is given, the appropriate age grouping will be specified.

Once the populations were simulated the next stage was to draw samples from the populations. The chosen sampling procedure was a two-stage sampling scheme involving a random sample of clusters at the first stage (n clusters chosen), followed by a simple random sample of m_i observations from each cluster. For each population, the n clusters sampled were a proportion of the N clusters in the population, with the value of this proportion going from 5% to 95%. While 95% may seem like a large sampling fraction of clusters, it is very possible this may occur in a large scale survey, particularly if one is looking to capture a rare attribute in the population, or the available number of clusters is small. The m_i observations sampled at the second stage took a fixed value for each sample (i.e. $m_i = 5, 10, 15$ or 20). If the desired number of second-stage observations were not in the cluster, then the full cluster was included in the sample (i.e. if $M_i <$ desired number of second-stage observations then $m_i = M_i$).

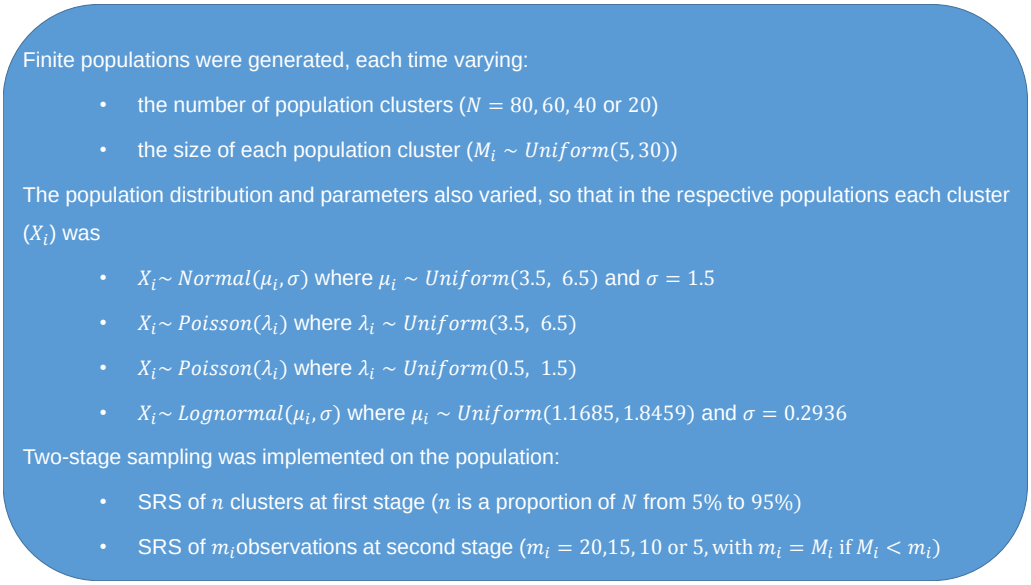


Figure 2.1: Summary of data generation & sampling scheme used

2.3.2 Implementation

Once the populations and samples were generated as described above, the next step was to find the associated simple random sample standard error (SRS SE), Taylor series linearisation standard error (TSL SE) and the delete-one jackknife standard error (JK1 SE) associated with samples from the population. As each of the populations were finite, and due to the sampling proportion being as large as 95%, a finite population correction (FPC) was included in the estimate of the SE. The FPC correction was

$$\sqrt{1-f},$$

with m_0 being the total number of observations in the sample, M_0 being the total number of observations in the population and f being the ratio of these (Lohr, 2009). Thus the formula to calculate the SE was

$$se(y) = \frac{s}{\sqrt{m_0}}\sqrt{1-f},$$

where s was the standard deviation estimate of the sample (Cochran, 2007; Lohr, 2009). It is noteworthy that the SRS SE is only a theoretical estimate and the variance of the survey methods should not be less than this estimate as, during the data generation each of the population clusters were centred around a

differing mean and each cluster had the same standard deviation. Thus, each of the survey variance estimation methods should notice this cluster effect, unless the conditions required for the survey variance method to be implemented are violated.

The estimates of the SE for both the TSL and JK1 methods were also calculated in R, using the "survey" package (Lumley, 2011). The details of these calculations can be found in the Appendix of Lumley (2011) and are sketched in the following two sections. As the design based SE results are returned when using the "survey" package, the respective root design effect (DEFT) denoted TSL DE and JK1 DE were calculated by dividing the SE of the respective survey method by the SE of the SRS method.

Taylor Series Linearization

If samples are gathered using unequal probability sampling, an estimate of the total can be obtained using the Horvitz-Thompson estimator (Lumley, 2011; Lee and Forthofer, 2005). Assuming a without replacement sample of n PSU's, with known inclusion probability $\pi_i = P(\text{unit } i \text{ in sample})$ and joint inclusion probability $\pi_{ij} = P(\text{units } i \text{ and } j \text{ are both in the sample})$ the Horvitz-Thompson (HT) estimator of the population total is

$$\hat{T}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i},$$

where $\hat{t}_i$ is an unbiased estimator of the PSU total. Given a function of one or more totals that is a continuously differentiable function (Lumley, 2011), the asymptotic variance of the statistic under the delta method is

$$\hat{var}[\hat{f}(\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_p)] \approx \sum_{i,j=1}^{p} \frac{\partial f}{\partial T_i} \hat{cov}[\hat{T}_i, \hat{T}_j] \frac{\partial f}{\partial T_i}.$$

Systems of equations that are estimated population totals can be used to estimate statistics that are not explicit functions of totals. Let $U_i(\theta)$ be a function of the data for observation i and a parameter θ. Given$\sum U_i()$ for the true population statistic over the whole population (N) is approximately zero, then $\sum U_i()$ for the complex sample statistic ($\hat{\theta}$) over the sample should also be approximately zero. Thus, $\hat{\theta}$ is a function of $\sum_i U_i(\theta)$ if $\sum_i U_i(\theta)$ is sufficiently smooth, and implying

the delta method one obtains:

$$\hat{var}[\hat{\theta}] \approx (\sum_{i=1}^{n} \frac{\partial U_i(\hat{\theta})}{\partial \theta})^{-1} \hat{cov}[\sum_{i=1}^{n} U_i(\hat{\theta})](\sum_{i=1}^{n} \frac{\partial U_i(\hat{\theta})}{\partial \theta})^{-1}.$$

Binder (1983) and Wolter (2007) describes this method in detail, which is called linearization (or TSL). Lohr (2009) details formulae to estimate totals, means, and variances using the Horvitz-Thompson estimator in one and two stage unequal probability sampling schemes, under both with and without replacement designs. When looking at multi-stage sampling (Lohr, 2009; Lumley, 2011), sometimes it is mathematically (and computationally) easier to consider each sampled cluster as a population for further sampling. The sum of the variance estimates from each stage of sampling can be used as the overall variance estimate. Treating multi-stage sampling as sequential is computationally simpler than using the direct Horvitz-Thompson formula, and allows any number of stages of sampling. Moreover, in simple designs, with equal probability sampling at each stage, formulae may be used directly to estimate the variance associated with statistics such as means and totals, and linearization is only required when the design becomes very complex or the statistic to be estimated is complex, such as a median or ratio. Thus, the variance of a mean estimate for a two-stage cluster sample with SRS at each level can be calculated directly using an analytical formula (Section 3.3.4) and linearisation is not be required. Statistical software will usually implement the required method directly without user specification.

Delete-One Jackknife

Suppose that (Lohr, 2009) we have k clusters. The jackknife procedure for clustered samples varies very little from the default procedure, with the only difference being the obligation to work with full clusters, rather than individual units. Thus each cluster (PSU) must be dropped one at a time, and division of the cluster unit is not allowed. Suppose $\hat{\theta}$ is an estimater of an arbitrary parameter θ. Using the JK1 methodology one of the k clusters (cluster α) will be dropped each time and the statistic

$$\hat{\theta}_{(\alpha)} = \frac{\sum_{i \neq \alpha} \theta_i}{k-1},$$

which is the estimate of θ following the same functional form as $\hat{\theta}$ without the

cluster α, will be calculated for each value of α ($\alpha = 1, 2, \ldots, k$).

The jackknife estimator is the mean of the $\hat{\theta}_{(\alpha)}$ given by:

$$\hat{\bar{\theta}} = \frac{\sum_{\alpha=1}^{k} \hat{\theta}_{(\alpha)}}{k}.$$

And, the variance of the jacnknife estimator is given by

$$\hat{var}[\hat{\bar{\theta}}] = \frac{k-1}{k} \sum_{\alpha=1}^{k} (\hat{\theta}_{(\alpha)} - \hat{\bar{\theta}})^2.$$

The JK1 method is most suitable for unstratified designs and further detail of the jackknife estimation technique can be found in Lohr (2009) and Wolter (2007).

Number of Simulations

The simulation and variance estimation procedure described above was completed for each combination of $N = 80, 60, 40$ and 20 clusters with $m_i = 20, 15, 10$ and 5, thus obtaining 16 different combinations of N and m_i. Once the choice of N and m_i were specified, the above estimation of SE was completed for $50,000$ independent simulations with the same characteristics of N and m_i. The mean of the $50,000$ simulations was chosen as the SE estimate for that combination of N and m_i to ensure a robust SE estimate. The $50,000$ replications avoided an extreme population influencing the SE estimates and the number of replications was chosen to ensure that the difference in SE was consistently less than 1%, when adding more replications (in blocks of 000's) for each of the normal, Poisson and lognormal distributions. Figure 2.2 shows the difference in SE caused by adding an extra 1000 observations compared to the original SE estimate for the sample without the extra 1000 observations (as a percentage). Thus Figure 2.2 shows the largest difference in SE at each proportion (i.e. 5% to 95%) for each m_i under four different choices of N for the normal distribution. The figures for the two versions of the Poisson distribution, along with the lognormal distribution are shown in Appendix A.

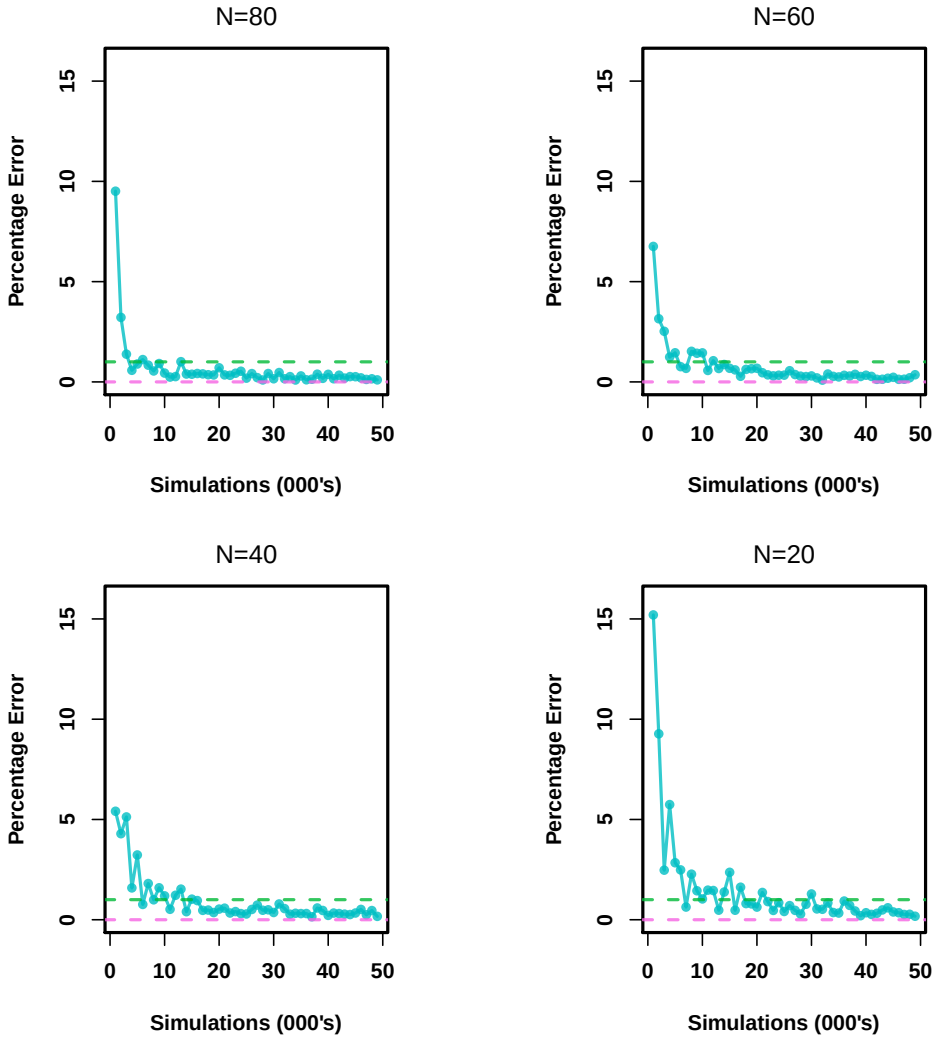


Figure 2.2: Largest SE for normal distribution across all m_i values for $N = 80, 60, 40$ and 20.

2.4 Results

Although the simulation was carried out for $N = 80, 60, 40$ and 20 population clusters the results are presented only for $N = 80$ population clusters scenario. $N = 80$ was chosen as, in a geographical divide of primary schools, it is representative of a typical number of primary schools per geographical area in Ireland. The mean number of secondary schools would, in general, be smaller as the schools are larger. The analyses/figures for the $N = 60, 40$ and 20 estimates showed almost identical patterns as the $N = 80$ cases and are presented in Appendix B. The only difference was that the SE estimates increased as the number of population clusters decreased, which one would expect.

For all of the following results the SRS SE is only a theoretical estimate of SE for reference. As the population was a clustered population, the SE estimates under the cluster methodologies would be expected to always be higher than this estimate (i.e. as clusters were each generated around a particular mean value and thus the values within each cluster would be expected to be more alike than values from different clusters). The TSL SE estimates would be calculated using a formula in this setting as the design of the survey was relatively simple, and the mean is also a simple statistic.

2.4.1 Normal Distribution

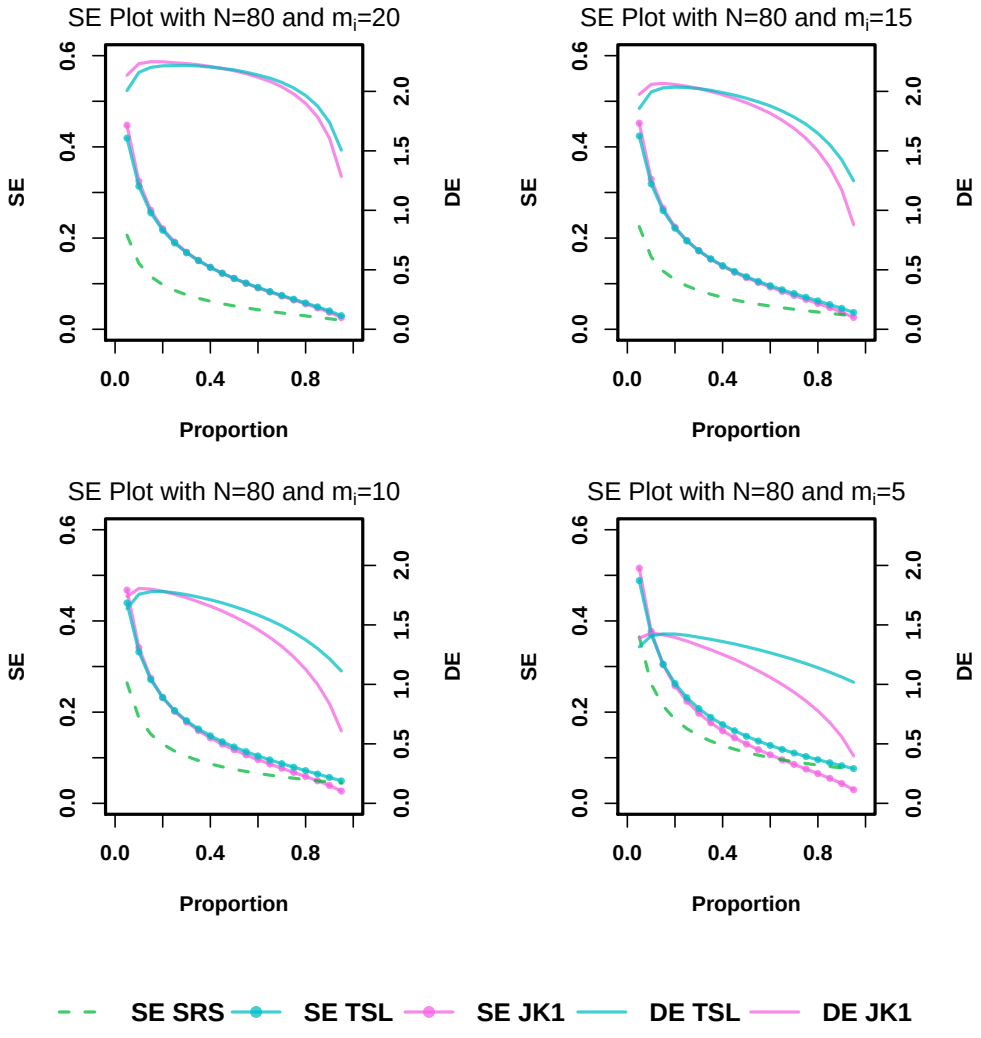


Figure 2.3: SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes.

As expected, the SE estimates decrease as the proportion of clusters sampled increase as more of the population is being sampled. This holds for each of the variance estimation techniques under the normal distribution (Figure 2.3). Observing Figure 2.3 there is essentially no difference in the SE estimates under both TSL and JK1, when the number of elements at second-stage sampling (m_i) is large or the sampling fraction (proportion) is low. Divergence in estimates between the two methods occurs when the sampling fraction is large or the number of elements at second-stage sampling is low. This result is not surprising as neither method is recommended for large sampling fractions (Wolter, 2007). It is noteworthy that the TSL DE never becomes less than one even when the sampling fraction is large, even though the use of the TSL method is not advised. Increasing the sampling fraction and decreasing the number of second-stage elements does cause the JK1 DE to become less than one. This may be attributed to the JK1 methods development for a with-replacement design, and thus when the sampling fraction is large this method becomes inappropriate to use (Wolter, 2007). One possible explanation for this result is the small number of second-stage observations sampled lead to large variation within each cluster (as only a few elements), but the large sampling fraction will increase the number of cluster mean estimates, thus reducing the variation between cluster mean estimates. This results in a DE less than one. Thus, as the TSL method estimates highlight the cluster effect in all cases for the normal distribution it can be seen as the more prudent estimator of the two. However, when the sampling fraction is small and the number of elements sampled at second-stage is large both methods provide almost identical estimates.

2.4.2 Poisson ($\lambda \sim U(3.5, 6.5)$) Distribution

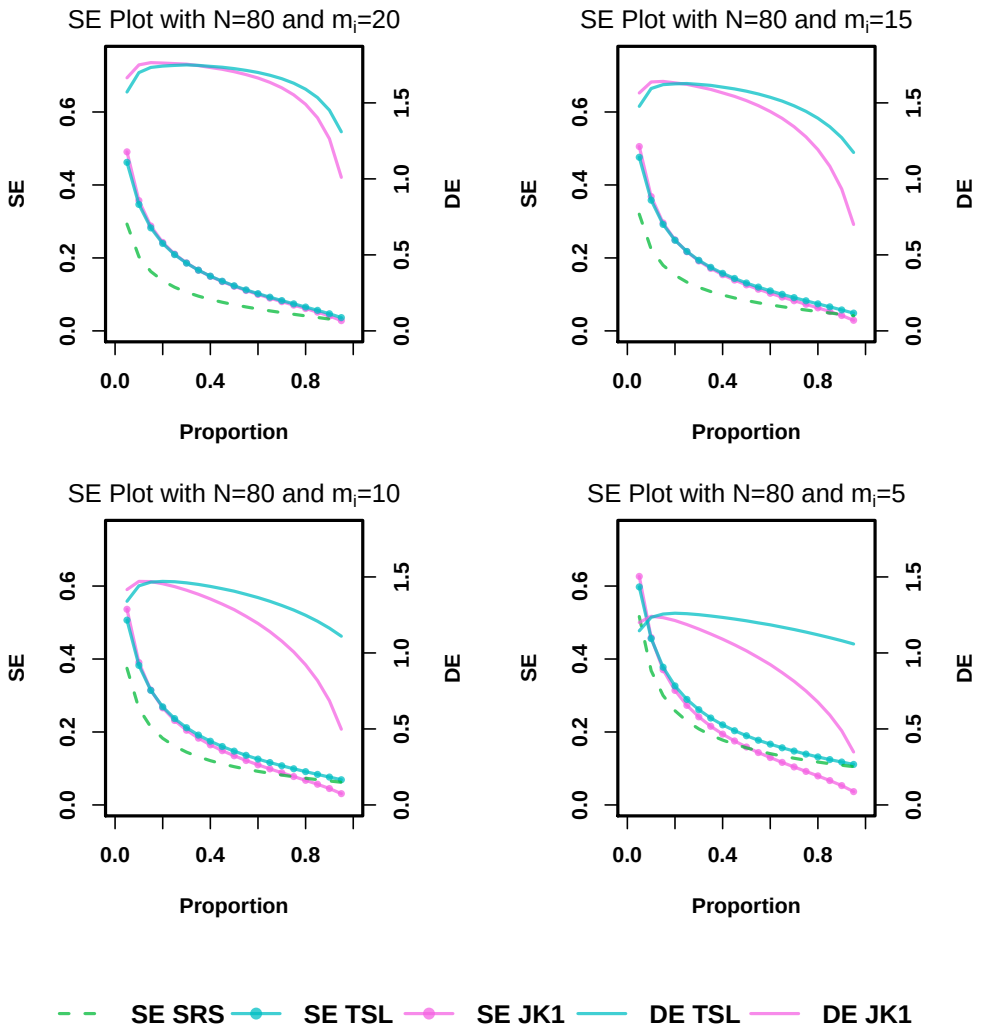


Figure 2.4: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes

The Poisson($\lambda_i \sim U(3.5, 6.5)$) distribution SE estimates (Figure 2.4) follow a similar pattern to the normal distribution, with the SE estimates decreasing as the proportion of clusters increased. However, the estimates of SE for the survey methods are closer to the SRS estimates than the normal distribution. The JK1 estimates again diverge from the TSL estimates as the sampling fraction becomes large or the number of second-stage elements becomes small. Despite the TSL estimates being closer to the SRS estimates, the TSL DE again does not drop below one, and thus the cluster effect is highlighted at each sampling fraction for each number of second-stage elements. However, the JK1 method shows an

increased number of estimates with DE less than one for each second-stage m_i value, compared to the normal distribution. One possible explanation for this is the increase in skewness of the Poisson distribution compared to the normal distribution. The JK1 was known to underestimate SE in the presence of skewed distributions. Again however, where the sampling fraction is low, and the number of second-stage sampling elements is large, both the TSL and JK1 produce almost identical estimates.

2.4.3 Poisson ($\lambda \sim U(0.5, 1.5)$) Distribution

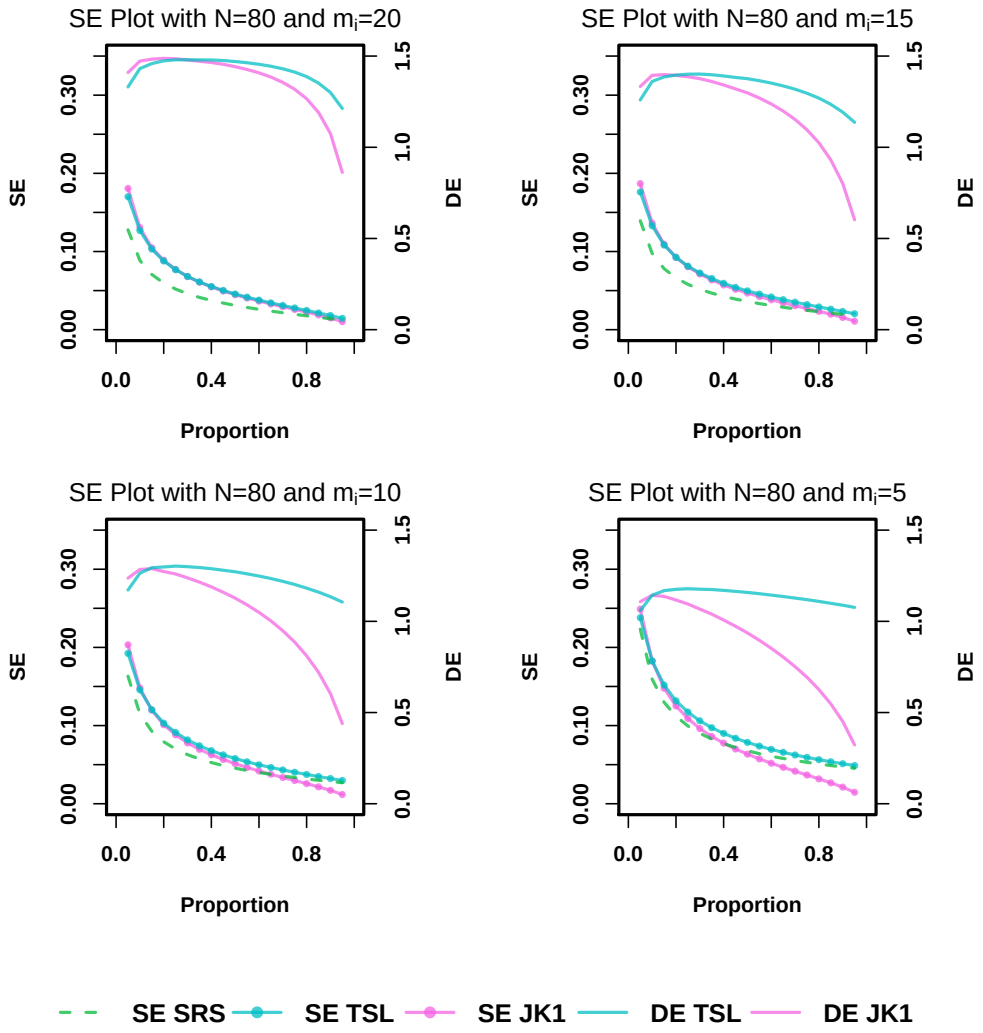


Figure 2.5: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes.

The results of the Poisson($\lambda_i \sim U(0.5, 1.5)$) distribution are shown in Figure 2.5. These results are almost identical to those of the Poisson($\lambda_i \sim U(3.5, 6.5)$) shown in Figure 2.4. However, due to the increase in skewness, the JK1 method underestimates SE even more severely, thus leaving a larger difference between the curves. Thus, there is an increased number of DE's less than one for each m_i value compared to Figure 2.4. Again, the TSL method notices the cluster effect at all sampling fractions for each second-stage sample size.

2.4.4 Lognormal Distribution

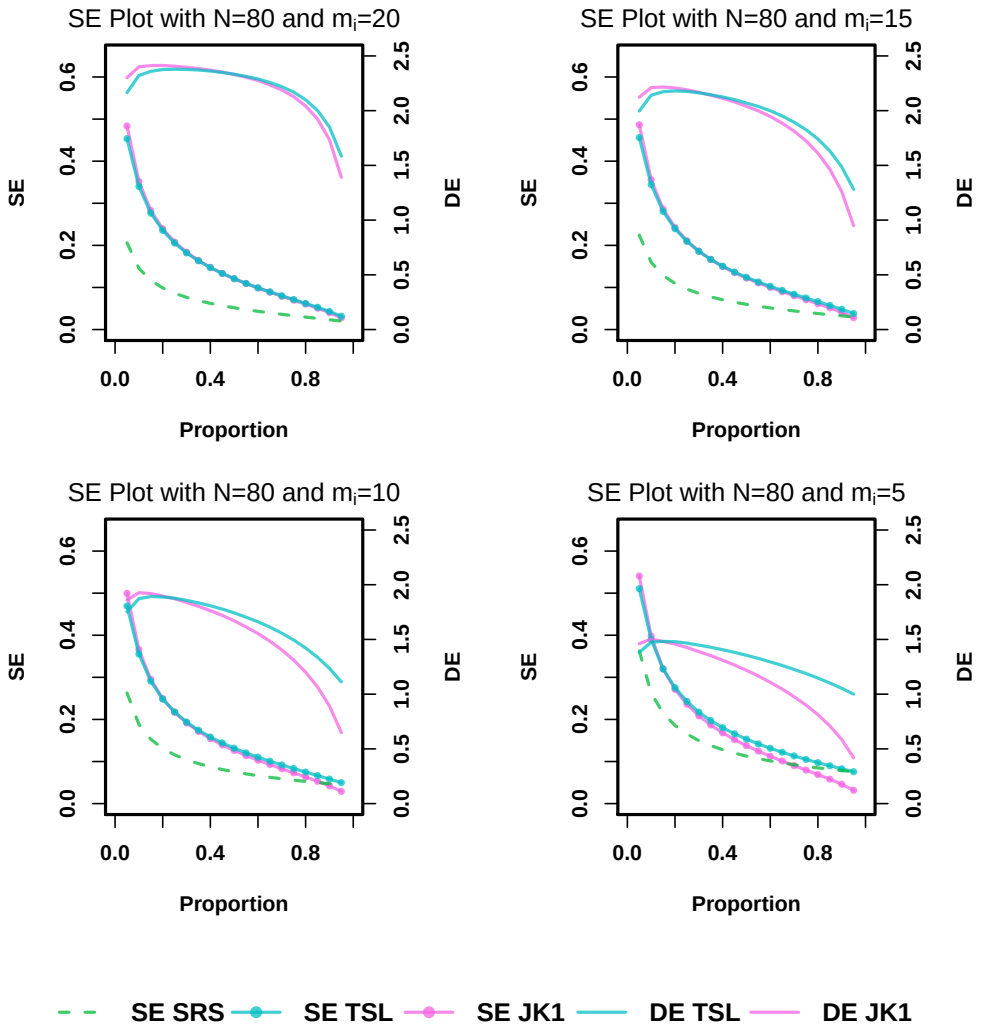


Figure 2.6: SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes.

The results of the lognormal distribution are presented in Figure 2.6, and are comparable to the results of the normal distribution (Figure 2.3). The SE estimates reduce as the sampling proportion at first stage increases. The DE estimates only become less than one for the JK1 estimation method, when the number of second-stage units become small, and the sampling fraction is large. This is similar to the Normal distribution. Even though the lognormal distribution is skewed, the JK1 method does not suffer the same large number of DE's less than one, as in the case of the Poisson. Similarly to the normal distribution there is little difference in the estimates between TSL and JK1 until the number of second-stage observations becomes small.

It is noteworthy that for both Poisson distributions, with $N = 40$ and 20 clusters and with small second-stage sample sizes (Figures B.5, B.6, B.8 & B.9 respectively, shown in Appendix B), the TSL method showed DE's less than one for small sampling fractions (5% and 10%). This could be attributed to the small number of clusters, causing unreliable estimates of the SE. As TSL is not recommended for less than five clusters, this may explain this occurrence. This underestimation may also be due to the heavy skewed nature of the population for which TSL may underestimate the SE. Thus, under heavily skewed populations, neither TSL nor JK1 may be appropriate to obtain SE estimates, as both give DE estimates less than one for a clustered population. Other than this notable point, the results were similar for $N = 60, 40$ and 20 population clusters.

2.5 Concluding Remarks

The simulations in Chapter 2 have examined the effect of an increasing sampling fraction on multiple distributions, numbers of clusters and numbers of second stage sampling elements. The simulations show that at large sampling fractions the estimates under the two methods diverge, and the definition of this 'large' sampling fraction varies based on the choice of the underlying distribution. Estimates of the sampling fraction where divergence occurs can be obtained by looking at the results section, and may be used as a guide for surveys with similar characteristics. Furthermore this chapter shows that small numbers of clusters and small second stage sample sizes affect the estimates under both methods, as was described in the literature review, even in the presence of increasing sampling fractions. It is also noteworthy that design effects less than one have appeared. However, for the TSL method, these occur only when $n < 5$ and there is doubt whether the variance estimation method is accurate. The design effects less than one appear more frequently for the JK1 method, suggesting it becomes inappropriate when the sampling fraction is large, and again this estimate of large can be obtained from the results of the simulation, in the context of an oral health design.

Chapter 3

Exploratory Analyses of Factors Associated With Design Effects Less Than One

3.1 Outline

This chapter analyses a national oral health dataset, in which there are regions with design effects less than one. As the data was collected using cluster sampling, the presence of design effects less than one is extremely unusual and warranted an exploratory analysis to try and identify the possible causes. Section 3.3 describes the design of the survey (Section 3.3.1), outlines the methods used to calculate the standard errors (Section 3.3.4), and presents the standard error results per CCA (Section 3.3.6). Section 3.4 presents exploratory analyses which aimed to determine the factors associated with the design effects less than one, including cluster analysis (Section 3.4.1) and linear models (Section 3.4.2).

3.2 Motivation

After reviewing and comparing the properties and requirements of the most routinely implementable survey variance calculations in the previous chapter, the next step was to calculate variance estimates for a national survey conducted in Ireland, with variables of a similar form to Chapter 2. The aim of the survey was to assess the effect of water fluoridation on children's oral health. The

multi-stage, seemingly simple, clustered sampling design of this survey suggested using a closed form estimators for both the mean and variance of the survey data, with these estimators being identical to those termed TSL estimators in Chapter 2. Estimates were produced by manually scribing code, and also produced using the inbuilt packages R and SAS (where applicable). Calculation of design effects (DE's) were then carried out to assess the effect of the cluster sampling design on the variance estimates. The DE estimates obtained on the real data set were substantially different to Chapter 2, with many CCA's having design effects less than one. A large number of design effects less than one, under a cluster sampling design, was never observed in the literature reviewed and warranted an exploration of the possible causes.

3.3 Design Effect Calculations

3.3.1 The Survey Design

The survey followed a multi-stage sampling procedure involving stratified and cluster sampling. Children were selected randomly on the basis of age, gender, geographical location of the school attended, and whether they attended a school with a fluoridated or non-fluoridated water supply. Firstly the population was stratified by age group, which following the World Health Organization (WHO) standards, consisted of four classes (Age 5, Age 8, Age 12 and Age 15). Age stratification was implemented by choosing the class of the child (Junior Infants, 2nd Class, 6th Class and Junior Certificate), which represented the aforementioned age groups. This method of using classes to divide children into appropriate age groups is common practice in survey sampling. Stratification based on age is important in oral health as dental caries is a cumulative condition, and thus levels increase with age. Furthermore the number of teeth and type of teeth (deciduous or permanent) change with age.

After stratification by age a further level of stratification was implemented, based on a geographical divide termed Community Care Areas (CCAs). Within each CCA, a cluster sampling technique was used with schools as the primary sampling unit (PSU). Schools were stratified according to size (to ensure representation of schools of various sizes) along with whether the school was located in a fluoridated or non-fluoridated area. Once the PSU's were chosen, a list of children in each year was obtained from the selected schools.

The required number of children was selected randomly from each year and consent forms were issued only to those children. In cases where there was a number of different classes within one year, a class was randomly selected and the children were randomly selected within this class. If insufficient numbers of children were presented in the first class selected, another class was randomly selected until the required number of children to issue consent forms to was obtained. Histograms of the dmft, dmfs and BMI variables for the children sampled by age group and fluoridation status are presented in Appendix C.

3.3.2 Aims

The initial aim of this analysis was to obtain mean and variance estimates of the number of decayed, missing and filled teeth (dmft), stratified by age group, CCA and fluoridation status of the child. To simplify the variance estimation calculation, the stratification of school sizes within CCA's was omitted from the variance calculations. This omission should lend to larger variance estimates as stratification generally decreases the variability in a survey design due to a more representative sample of the population (Särndal, Swensson and Wretman, 2003; Wolter, 2007). Thus 240 cluster level estimates of mean number of dmft per child, along with the associated variability were produced (4 Age Groups $\times$ 30 CCA's $\times$ 2 Fluoridation Status'). The variance of each of these 240 groups was also calculated treating each grouping as a simple random sample (SRS). Comparison of the two variance estimates demonstrated the change in variability introduced by the multi-stage sampling design, and also allowed calculation of the design effect (DE) of each grouping. The design effect for a given variable is the ratio of the variance calculated using survey variance methods relative to the variance of the design calculated as if it was a SRS. The DE definition varies slightly from that of Chapter 2, which uses the root design effect (ratio of SE's). The ratio of variances is used here as it is more commonly encountered within survey designs, however both methods are comparable by taking the respective square or square root. As mentioned previously, cluster samples are, in general, less efficient (i.e. have more variability per unit) than simple random samples, and thus have design effect estimates greater than one.

After age, fluoridation and CCA stratification the sampling design consisted of a SRS of clusters followed by a SRS of observations at second stage, and thus closed form estimators were used to calculate mean and variance estimates. Initially two estimators were explored, namely the unbiased estimator and the ratio estimator.

Due to peculiarities within the data collected during the survey, code needed to be written in R to allow calculation of the variance estimates, and the details of these calculations, along with the details of the peculiarities will be described in the following sections.

3.3.3 Notation

N = number of primary sampling units (PSUs) in the population (e.g. available schools in the population).

n = the number of PSUs selected in the sample (e.g. sampled schools).

M_i = the number of secondary sampling units (SSU's) in PSU i (e.g. the children eligible for sampling in the chosen school).

m_i = the number of SSU's selected in the sample in PSU i (e.g. the children sampled in the chosen school).

$M = \sum_{i=1}^{N} M_i$ = the total number of SSUs in the population (e.g. total number of children in the population).

$\bar{M} = \frac{M}{N}$ = the average cluster size for the population (e.g. average number of children per school in the population).

y_{ij} = the jth obseravtion in the sample from the ith cluster (e.g. the dmft value of the child).

$\bar{y_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ = the sample mean for the ith cluster (e.g. the mean dmft value for the chosen school).

Before presenting the estimator and variance formulae it is noteworthy that when the sampling fraction at the first stage of sampling is small (i.e. $\frac{n}{N}$ is small), the variance contribution of the second stage of sampling is often ignored as its contribution to the overall variance is negligible relative to the first stage variability (Lohr, 2009). This is the default many software packages. The definition of a small sampling fraction is not well defined in the literature. However, as some of our sampling fractions exceed 50% it was decided to include the contribution of second stage variability. Furthermore, the survey data gathered may have some CCA's with a relatively small number of population clusters (N) at first stage, leading to high sampling fractions, and including the second stage variability may

give a more accurate representation of the variability within the CCA. Thus all the formulae presented in the next section are in terms of a two-stage sampling procedure, calculating variance with contributions from both the first and second stage (i.e. between cluster variability and within cluster variability), and this calculation is identical to that labelled TSL in Chapter 2. The following sections will briefly introduce the notation for cluster sampling, and describe both estimators.

3.3.4 Methods for Mean and Variance Estimation

Unbiased Estimator

In two stage cluster sampling, assuming that simple random sampling is used at both stages to choose PSUs and SSUs, the unbiased estimator of the mean μ is given by

$$\bar{y}_{unb} = \left(\frac{N}{M}\right)\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{n} = \frac{1}{\bar{M}}\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{n}$$

with the variance of $\hat{\mu}$ given by

$$\hat{V}(\bar{y}_{unb}) = \left(1-\frac{n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)s_b^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$$

where

$$s_b^2 = \frac{\sum_{i=1}^{n}(M_i\bar{y}_i - \bar{M}\bar{y}_{unb})^2}{n-1}$$

and

$$s_i^2 = \frac{\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2}{m_i-1} \text{ for } i = 1,2,\ldots,n$$

Notice that in the above, s_b^2 is simply the sample variance among the terms $M_i\bar{y}_i$ and s_i^2 is the sample variance for the sample selected from cluster i. The unbiased estimator of the mean can be inefficient when the values of M_i are unequal (Lohr, 2009). Furthermore, the above estimator of $\hat{\mu}$ depends on M, the number of

elements in the population. A method of estimating μ when M is unknown is given by the ratio estimator in the next section.

Ratio Estimator

When M, the total number of elements in the population is unknown, it must be estimated from the sample data. Multiplying the average cluster size ($\sum_{i=1}^{n} \frac{M_i}{n}$) by the number of clusters (N) in the population, an estimate of M can be obtained. Then, using the estimator of M a ratio estimator of the mean $\hat{\mu_r}$ is given by

$$\bar{y}_r = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

This estimator is a ratio estimator as both the numerator and denominator are random variables, and the estimate $\hat{\mu_r}$ has and estimated variance given by

$$\hat{V}(\bar{y}_r) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)s_r^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2\left(1 - \frac{m_i}{Mi}\right)\left(\frac{s_i^2}{m_i}\right)$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n} M_i^2(\bar{y}_i - \bar{y}_r)^2}{n-1} = \frac{\sum_{i=1}^{n}(M_i\bar{y}_i - M_i\bar{y}_r)^2}{n-1}$$

and

$$s_i^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2}{m_i - 1} \text{ for } i = 1, 2, \ldots, n$$

The variance of the ratio estimator depends on the variability of the means per element in the clusters, and can be much smaller than the unbiased estimator. Although the estimator $\hat{\mu_r}$ is biased, the bias is negligible when n is large (Lohr, 2009). When there is a correlation between cluster sizes and the variable of interest, or when cluster sizes vary wildly the ratio estimator may provide more accurate estimates than the unbiased estimator (Lohr, 2009).

3.3.5 Some Peculiarities in the Data

One obstacle when analyzing the data collected is that the fluoridation status of the child is only known after sampling, and this causes issues with the above calculations. The number of clusters that contain fluoridated children is not known in advance, and similarly for the number of non-fluoridated clusters. While the number of fluoridated and non-fluoridated schools are known, this does not guarantee the children's fluoridation status will match. This can be attributed to rural children travelling to urban schools or vice versa. Furthermore, these proportions are not known in advance of sampling, which would allow a poststratification adjustment to be implemented. The results presented in the next section are calculated taking N to be the total number of clusters in the CCA. Some of these clusters will contain no fluoridated observations, and some will contain no non-fluoridated observations and thus the number of population clusters is overestimated leading to larger variance estimates as the FPC correction will not be as influential. Furthermore the finite population correction at second stage is not known in advance. Once a school is selected, the divide of fluoridated and non-fluoridated children is not known for the population, and is only known for the sample once the sampling process is complete. The second stage correction is less of a worry as its contribution to the variance is often negligible compared to the first.

Using the above adjustment for the number of population clusters, estimates of the dmft per CCA and their associated variability were also calculated in R using the "survey" package. These calculations were also attempted in SAS (using PROC SURVEYMEANS). However, at the time of analysis SAS was unable to produce variance estimates including variability at second stage. As mentioned previously, the second stage variance contribution is omitted in some software as it is in general small relative to the first stage variation. The "survey" package has the capabilities to produce two-stage variance estimates. However, due to the time involved in planning and implementing the survey data collection, the list of population clusters and the respective class sizes at population level were outdated by one year. Thus in some scenarios population class size estimates (M_i) from the previous year were less than the number sampled in the survey (m_i). This was adjusted in the hand written code, allowing $m_i = M_i$ if the data showed that $m_i > M_i$. However, based on the format required for the statistical software this adjustment required a lot of data modification, and the decision was made to write the code in R. The previous year class size estimates were used

as, in general, the class sizes would have been similar, and failing to use these would have caused the FPC at second stage to be omitted. The handwritten code in R provides identical estimates to the "survey" package in R, allowing for the FPC adjustments in some troublesome cases. These cases included CCA's with population sizes smaller than sample sizes (explained previous) and second stage sample consisting of just one observation within a class (i.e. $m_i = 1$). This could happen in a rural school or a student commuting and attending a school with a different fluoridation status than their home status. The adjustment was to assign a value of zero to their variance at second stage.

It is also worth noting that by default, the ratio estimator is used in both statistical software rather than the unbiased estimator. This is not always initially obvious to the user of the statistical package. While in most cases the ratio estimator is more desirable the user should be aware the estimator is biased, and ensure they wish to use the ratio estimator. Otherwise the reader may have to look for alternative statistical software commands or to scribe their own code to obtain the unbiased estimator. Furthermore the statistical software can have some issues with calculation, such as when one observation appears at second stage, and some corrections need to be implemented to address this. While it is uncommon the second stage sample size will be one ($m_i = 1$), this can be a real difficulty in the data presented in this chapter due to some children commuting to schools from a non-fluoridated area, or small classes in rural schools.

3.3.6 Design Effect Results

Using the above formulae, and also obtaining estimates from R and SAS software where possible, estimates of the mean number of dmft stratified by fluoridation status, age group and CCA were obtained. The results for the 5-year old age group are shown in Table 3.1 and Table 3.2. Note, CCA one to eight do not appear in the NF region as no non-fluoridated observations were present in the sample for these CCA's as they were contained in Dublin city where the public water supply is exclusively fluoridated. Furthermore, CCA 4 for the fluoridated, aged 5 grouping has no estimates as there were no observations in this particular grouping, and this will be represented using N/A in the table of results. The total number of clusters at both sample (n) and population (N) level (estimated) are also given along with the total number of observations sampled (m_i) and the total number of observations available (M_i) in the CCA for a particular age group. The main focus of the results will be the ratio estimator, as in our data we have

some large variation between the cluster sizes within some CCA's. The unbiased estimator results will also be presented to see if the difference in estimates is large, and also to observe the DE values using the (in general) larger variance estimate.

Table 3.1: Mean and Variance Estimates of dmft for Fluoridated Observations Age 5

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	12	64	107	493	0.623	0.623	0.136	0.135	0.118	1.334	1.306
2	11	59	75	449	0.374	0.374	0.126	0.160	0.178	0.502	0.807
3	12	46	130	670	1.382	1.382	0.446	0.235	0.191	5.458	1.510
4	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
5	13	67	140	624	0.567	0.567	0.118	0.118	0.150	0.612	0.618
6	11	54	101	485	1.213	1.213	0.227	0.246	0.223	1.040	1.221
7	14	56	83	869	0.289	0.289	0.124	0.116	0.097	1.646	1.435
8	12	102	87	524	1.114	1.114	0.495	0.532	0.199	6.214	7.166
9	16	123	92	459	0.918	0.918	0.281	0.143	0.156	3.261	0.841
10	16	102	82	401	1.286	1.286	0.351	0.235	0.290	1.465	0.658
11	10	112	49	263	0.841	0.841	0.239	0.182	0.292	0.672	0.391
12	10	108	53	241	1.664	1.664	0.609	0.454	0.335	3.303	1.834
13	15	94	77	355	1.239	1.239	0.280	0.274	0.301	0.866	0.832
14	47	122	182	732	1.024	1.024	0.192	0.170	0.137	1.980	1.557
15	45	75	626	1,435	1.388	1.388	0.136	0.096	0.083	2.685	1.340
16	70	95	744	1,618	1.011	1.011	0.103	0.101	0.061	2.793	2.683
17	14	168	57	290	1.929	1.929	0.654	0.321	0.372	3.087	0.746
18	9	98	29	167	0.772	0.772	0.270	0.224	0.386	0.492	0.338
19	16	99	80	324	1.183	1.183	0.369	0.404	0.324	1.303	1.557
20	12	83	103	252	1.338	1.338	0.350	0.264	0.225	2.423	1.383
21	15	70	66	370	0.977	0.977	0.276	0.200	0.251	1.214	0.636
22	11	97	39	309	1.786	1.786	0.584	0.381	0.660	0.784	0.333
23	17	133	86	315	1.039	1.039	0.286	0.215	0.224	1.625	0.922
24	15	74	54	266	1.655	1.655	0.501	0.421	0.336	2.220	1.574
25	14	100	68	403	0.686	0.686	0.260	0.270	0.242	1.155	1.248
26	14	79	89	530	0.484	0.484	0.133	0.145	0.145	0.846	1.003
27	12	71	56	236	1.218	1.218	0.247	0.201	0.227	1.180	0.787
28	11	213	58	201	0.337	0.337	0.173	0.184	0.191	0.820	0.923
29	24	173	110	307	1.618	1.618	0.351	0.300	0.219	2.563	1.869
30	14	90	71	187	0.654	0.654	0.241	0.169	0.161	2.242	1.094

Interpreting the DE for the ratio estimator for CCA 1 ($DE(\bar{y}_r) = 1.306$), the cluster sample design inflates the standard error by 1.306 time that if the data was gathered using a simple random sample. The results show eight DE's less than one for the unbiased estimator and thirteen DE's less than one for the ratio estimator. Seven of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator. Given that in general, cluster sampling is less efficient than SRS (as clusters are usually homogeneous) this

seems a very large number of DE's less than one from 29 CCA's. The results of the non-fluoridated observations are shown in Table 3.2.

Table 3.2: Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 5

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	19	123	63	460	2.959	2.959	1.376	1.038	0.359	14.706	8.370
10	14	102	40	238	1.244	1.244	0.487	0.531	0.405	1.446	1.716
11	13	112	51	199	0.905	0.905	0.275	0.356	0.459	0.359	0.602
12	19	108	68	272	1.355	1.355	0.332	0.315	0.293	1.286	1.157
13	15	94	34	266	1.104	1.104	0.587	0.616	0.327	3.231	3.554
14	88	122	534	1,166	1.502	1.502	0.184	0.176	0.091	4.093	3.745
15	55	75	284	1,167	1.617	1.617	0.276	0.258	0.134	4.253	3.699
16	77	95	337	1,413	1.453	1.453	0.175	0.171	0.131	1.790	1.711
17	14	168	39	233	1.139	1.139	0.339	0.478	0.546	0.386	0.768
18	24	98	100	303	1.345	1.345	0.216	0.228	0.198	1.186	1.333
19	15	99	32	275	1.393	1.393	0.421	0.267	0.417	1.015	0.410
20	13	83	67	174	1.935	1.935	0.710	0.328	0.293	5.894	1.260
21	17	70	54	288	1.986	1.986	0.968	0.608	0.393	6.072	2.393
22	19	97	57	422	1.922	1.922	0.712	0.776	0.514	1.920	2.279
23	14	133	53	201	1.750	1.750	0.548	0.472	0.296	3.425	2.539
24	20	74	48	273	2.513	2.513	0.570	0.365	0.316	3.259	1.334
25	13	100	44	209	1.383	1.383	0.382	0.308	0.355	1.161	0.753
26	6	79	24	145	1.038	1.038	0.503	0.444	0.450	1.249	0.972
27	24	71	78	340	1.904	1.904	0.431	0.364	0.328	1.724	1.231
28	15	213	56	208	2.554	2.554	0.524	0.491	0.436	1.443	1.266
29	11	173	15	140	2.493	2.493	0.651	0.584	0.590	1.219	0.979
30	13	90	36	144	1.085	1.085	0.681	0.584	0.209	10.636	7.836

Out of the 22 CCA's observed for non-fluoridated observations aged 5, there are two DE's less than one for the unbiased estimator and six DE's less than one for the ratio estimator. Two of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator. Thus even with the unbiased estimator, the presence of some DE's less than one remains. Throughout the literature review only one paper (Hammer, Shin and Porcellini, 2003) reported a DE less than one for cluster sampling, and this was only one case in a survey with many design effects greater than one. Also, most of the papers reviewed used the ratio estimator, and in the above results there are six DE's less than one out of the 22 cases. These results are very unusual and, if the remaining age groups show similar DE estimates, it would be useful to attribute the cause of this result to some reason. The results of the fluoridated observations aged 8 are shown in Table 3.3.

Table 3.3: Mean and Variance Estimates of dmft for Fluoridated Observations Age 8

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	13	67	114	515	0.219	0.219	0.061	0.058	0.060	1.025	0.918
2	9	60	76	343	0.170	0.170	0.074	0.061	0.084	0.776	0.521
3	11	47	123	560	0.371	0.371	0.090	0.098	0.078	1.347	1.599
4	15	38	90	488	0.151	0.151	0.058	0.061	0.072	0.657	0.725
5	12	72	130	540	0.215	0.215	0.069	0.060	0.051	1.835	1.405
6	14	61	130	526	0.335	0.335	0.090	0.075	0.074	1.470	1.043
7	14	60	87	725	0.070	0.070	0.032	0.033	0.033	0.906	1.003
8	11	107	81	480	0.408	0.408	0.118	0.114	0.087	1.826	1.728
9	15	129	81	326	0.273	0.273	0.064	0.062	0.082	0.609	0.565
10	12	107	62	309	0.281	0.281	0.143	0.129	0.113	1.601	1.302
11	10	115	65	254	0.169	0.169	0.068	0.068	0.076	0.795	0.811
12	9	112	66	227	0.640	0.640	0.182	0.157	0.135	1.807	1.343
13	16	101	79	295	0.103	0.103	0.035	0.041	0.048	0.544	0.744
14	13	126	39	221	0.468	0.468	0.224	0.227	0.218	1.054	1.084
15	14	77	63	323	0.150	0.150	0.073	0.071	0.083	0.785	0.737
16	11	99	34	320	0.461	0.461	0.210	0.152	0.209	1.008	0.533
17	11	168	58	257	0.219	0.219	0.083	0.080	0.069	1.444	1.340
18	4	104	19	54	0.167	0.167	0.141	0.120	0.085	2.733	1.980
19	17	106	86	352	0.473	0.473	0.199	0.144	0.081	6.018	3.139
20	12	88	77	235	0.403	0.403	0.141	0.145	0.104	1.828	1.926
21	12	74	60	292	0.336	0.336	0.148	0.091	0.091	2.624	0.995
22	9	98	47	261	0.750	0.750	0.387	0.236	0.170	5.205	1.932
23	17	141	61	295	0.257	0.257	0.123	0.128	0.108	1.288	1.409
24	12	81	53	147	0.961	0.961	0.233	0.186	0.172	1.838	1.168
25	16	104	82	380	0.084	0.084	0.055	0.056	0.062	0.795	0.816
26	14	79	88	526	0.132	0.132	0.038	0.043	0.046	0.690	0.878
27	12	75	47	175	0.467	0.467	0.173	0.174	0.135	1.651	1.665
28	11	224	68	240	0.175	0.175	0.075	0.058	0.069	1.200	0.704
29	22	181	77	312	0.420	0.420	0.149	0.149	0.085	3.075	3.060
30	11	93	50	149	0.148	0.148	0.072	0.058	0.083	0.761	0.488

The fluoridated age 8 results are presented for the 30 CCA's as CCA 4 now has observations sampled. There are 10 CCA's with DE less than one using the unbiased estimator, and 13 CCA's with DE less than one using the ratio estimator. Nine of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator. The unbiased estimator has more DE's less than one for 8 year olds vs 5 year olds. The ratio estimator has the same amount of DE's less than one compared with 5 year olds, however it has an extra CCA result (30 clusters not 29). The CCA's with DE's less than one are not all the same CCA's that were identified in the case of 5 year olds.

Table 3.4: Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 8

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	17	129	54	306	0.233	0.233	0.106	0.107	0.094	1.272	1.295
10	11	107	21	138	0.157	0.157	0.082	0.089	0.135	0.371	0.431
11	8	115	11	123	0.089	0.089	0.065	0.074	0.194	0.114	0.143
12	14	112	71	213	0.629	0.629	0.249	0.229	0.107	5.391	4.554
13	9	101	23	168	0.500	0.500	0.240	0.224	0.209	1.312	1.142
14	17	126	74	247	0.261	0.261	0.117	0.121	0.089	1.703	1.823
15	12	77	54	189	0.251	0.251	0.098	0.086	0.100	0.961	0.742
16	12	99	28	234	0.253	0.253	0.150	0.145	0.177	0.717	0.670
17	14	168	47	255	0.306	0.306	0.164	0.179	0.090	3.329	3.967
18	20	104	59	247	0.165	0.165	0.058	0.051	0.065	0.794	0.622
19	15	106	30	225	0.896	0.896	0.712	0.545	0.127	31.404	18.416
20	9	88	20	121	0.130	0.130	0.086	0.088	0.081	1.110	1.186
21	8	74	14	117	0.150	0.150	0.102	0.110	0.096	1.110	1.291
22	12	98	19	304	0.128	0.128	0.100	0.103	0.128	0.606	0.645
23	16	141	46	219	0.184	0.184	0.113	0.114	0.105	1.165	1.173
24	13	81	41	160	0.397	0.397	0.119	0.119	0.100	1.416	1.398
25	13	104	49	194	0.301	0.301	0.151	0.122	0.145	1.084	0.705
26	8	79	20	278	0.138	0.138	0.070	0.086	0.281	0.063	0.093
27	22	75	76	308	0.444	0.444	0.126	0.129	0.142	0.788	0.831
28	6	224	20	98	0.125	0.125	0.099	0.112	0.109	0.817	1.052
29	5	181	7	90	0.400	0.400	0.394	0.403	0.285	1.910	1.995
30	6	93	9	89	0.322	0.322	0.236	0.200	0.147	2.599	1.863

The 8 year old non-fluoridated results show the unbiased estimator has nine DE's less than one, while the ratio estimator also has nine DE's less than one. Eight of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator. The number of DE's less than one under both unbiased and ratio estimation show a substantial increase compared with the number of DE's less than one for the 5 year old case.

Table 3.5: Mean and Variance Estimates of dmft for Fluoridated Observations Age 12

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	12	64	92	472	0.790	0.790	0.228	0.202	0.135	2.857	2.243
2	11	57	71	479	0.809	0.809	0.280	0.201	0.155	3.253	1.675
3	11	47	131	562	1.175	1.175	0.194	0.161	0.131	2.197	1.517
4	13	38	70	422	0.524	0.524	0.155	0.154	0.132	1.375	1.349
5	12	67	115	602	1.108	1.108	0.221	0.123	0.119	3.446	1.068
6	14	62	98	535	0.767	0.767	0.137	0.161	0.124	1.212	1.683
7	13	57	93	812	0.649	0.649	0.143	0.118	0.129	1.223	0.830
8	12	103	74	472	1.126	1.126	0.365	0.230	0.175	4.354	1.733
9	18	131	92	346	1.463	1.463	0.405	0.277	0.142	8.184	3.812
10	9	104	48	220	0.804	0.804	0.214	0.253	0.193	1.237	1.719
11	13	116	89	298	1.041	1.041	0.219	0.258	0.158	1.924	2.667
12	10	108	53	231	1.280	1.280	0.310	0.237	0.249	1.544	0.905
13	12	100	74	235	0.922	0.922	0.289	0.245	0.163	3.136	2.255
14	14	127	40	279	0.610	0.610	0.206	0.165	0.184	1.258	0.809
15	12	77	68	381	1.169	1.169	0.233	0.156	0.179	1.706	0.762
16	11	97	34	236	1.596	1.596	0.526	0.581	0.314	2.813	3.435
17	12	170	51	309	1.030	1.030	0.292	0.174	0.184	2.512	0.897
18	5	100	15	45	1.539	1.539	0.436	0.226	0.355	1.510	0.404
19	18	107	71	375	0.982	0.982	0.190	0.248	0.151	1.581	2.708
20	15	85	95	330	1.427	1.427	0.481	0.183	0.161	8.946	1.294
21	12	74	58	251	0.979	0.979	0.247	0.198	0.155	2.563	1.646
22	9	98	44	273	1.070	1.070	0.326	0.125	0.215	2.291	0.337
23	16	141	58	265	1.058	1.058	0.202	0.288	0.274	0.542	1.108
24	16	84	61	224	1.286	1.286	0.269	0.175	0.185	2.100	0.895
25	17	102	75	375	0.780	0.780	0.310	0.225	0.143	4.717	2.483
26	11	80	80	398	0.736	0.736	0.155	0.149	0.161	0.923	0.856
27	10	75	47	152	1.706	1.706	0.385	0.280	0.276	1.947	1.033
28	10	227	50	209	1.001	1.001	0.341	0.343	0.180	3.567	3.627
29	16	181	78	262	1.583	1.583	0.287	0.178	0.216	1.776	0.677
30	10	95	51	140	0.790	0.790	0.158	0.058	0.146	1.179	0.160

There are two CCA's with DE less than one using the unbiased estimator, and eleven CCA's with DE less than one using the ratio estimator. Just one of the CCA's presenting with DE's less than one is common to both the unbiased and the ratio estimator for the fluoridated age 12 results. This is the first fluoridated age group where there is a large difference in the number of DE's less than one when comparing the unbiased and ratio estimator. This may have occurred as a DE of one was taken as a hard cut off. Observing the individual DE estimates, many are close to one, and thus the biased smaller variance estimate given by the ratio estimator may just bring these borderline DE estimates below one. Therefore in a real world setting the design effect values close to one may be almost negligible

in a real world setting.

Table 3.6: Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 12

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	12	131	34	177	0.842	0.842	0.287	0.267	0.211	1.856	1.603
10	10	104	14	221	1.088	1.088	0.518	0.545	0.416	1.552	1.717
11	13	116	27	223	0.891	0.891	0.400	0.390	0.241	2.762	2.632
12	16	108	60	247	1.425	1.425	0.248	0.263	0.231	1.151	1.296
13	6	100	13	85	1.042	1.042	0.397	0.180	0.328	1.469	0.302
14	19	127	54	282	0.920	0.920	0.293	0.258	0.205	2.051	1.580
15	14	77	50	275	1.116	1.116	0.174	0.147	0.225	0.598	0.427
16	12	97	35	212	1.759	1.759	0.624	0.468	0.331	3.562	2.003
17	20	170	43	351	0.757	0.757	0.235	0.135	0.180	1.691	0.561
18	18	100	72	287	1.038	1.038	0.220	0.273	0.201	1.201	1.855
19	11	107	25	191	2.334	2.334	0.775	0.482	0.344	5.079	1.962
20	8	85	16	156	1.929	1.929	0.730	0.607	0.433	2.842	1.964
21	7	74	9	112	2.737	2.737	1.164	0.777	0.757	2.367	1.053
22	7	98	9	162	0.864	0.864	0.417	0.361	0.351	1.415	1.061
23	14	141	43	189	1.425	1.425	0.398	0.312	0.280	2.026	1.244
24	17	84	31	240	1.462	1.462	0.340	0.328	0.401	0.719	0.668
25	16	102	53	248	0.970	0.970	0.206	0.208	0.233	0.781	0.800
26	6	80	25	144	1.011	1.011	0.393	0.363	0.278	2.000	1.700
27	23	75	93	301	2.311	2.311	0.432	0.347	0.231	3.492	2.244
28	4	227	6	69	1.370	1.370	0.789	0.794	1.024	0.593	0.602
29	3	181	6	91	3.205	3.205	1.438	1.146	0.846	2.888	1.834
30	6	95	16	69	1.176	1.176	0.528	0.381	0.404	1.707	0.889

The 12 year old non-fluoridated results show the unbiased estimator has four DE's less than one, while the ratio estimator has seven DE's less than one. Both of these are comparable to the age 5 results for the non-flouridated results, but substantially less than the DE's less than one for the aged 8 age group. Four of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator.

Table 3.7: Mean and Variance Estimates of dmft for Fluoridated Observations Age 15

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	12	45	97	982	1.991	1.991	0.375	0.335	0.209	3.212	2.555
2	10	17	74	981	1.738	1.738	0.292	0.230	0.249	1.378	0.852
3	10	19	147	1,102	2.098	2.098	0.270	0.266	0.201	1.796	1.739
4	6	18	54	554	1.444	1.444	0.334	0.254	0.353	0.900	0.521
5	11	32	126	1,299	1.972	1.972	0.340	0.199	0.191	3.182	1.089
6	11	27	122	918	1.696	1.696	0.304	0.278	0.194	2.448	2.044
7	10	32	78	1,157	2.329	2.329	0.503	0.308	0.227	4.915	1.840
8	13	30	103	1,083	2.556	2.556	0.432	0.341	0.262	2.721	1.695
9	9	26	79	696	2.238	2.238	0.618	0.444	0.262	5.563	2.872
10	8	25	55	718	1.813	1.813	0.402	0.457	0.325	1.536	1.985
11	9	18	60	778	2.170	2.170	0.467	0.371	0.287	2.658	1.676
12	7	30	43	744	1.895	1.895	0.449	0.394	0.324	1.921	1.484
13	7	19	42	637	1.074	1.074	0.285	0.279	0.270	1.119	1.073
14	8	21	26	767	2.473	2.473	0.549	0.447	0.477	1.322	0.875
15	9	17	65	1,090	2.666	2.666	0.462	0.371	0.326	2.007	1.295
16	12	20	35	967	2.542	2.542	0.525	0.399	0.379	1.917	1.109
17	10	24	40	1,143	1.604	1.604	0.381	0.327	0.335	1.289	0.949
18	7	23	40	500	1.759	1.759	0.689	0.654	0.352	3.830	3.451
19	8	26	55	615	2.661	2.661	0.693	0.389	0.323	4.601	1.450
20	10	16	137	926	2.270	2.270	0.244	0.209	0.191	1.634	1.198
21	8	20	77	813	2.576	2.576	0.490	0.235	0.257	3.648	0.835
22	7	20	32	674	2.099	2.099	0.636	0.541	0.303	4.391	3.177
23	6	29	30	435	4.208	4.208	1.896	1.416	0.455	17.369	9.683
24	10	17	55	796	2.622	2.622	0.358	0.359	0.364	0.968	0.972
25	7	34	70	544	2.417	2.417	0.395	0.206	0.295	1.798	0.488
26	8	25	79	913	2.298	2.298	0.590	0.462	0.237	6.184	3.787
27	6	12	50	380	3.129	3.129	0.560	0.569	0.461	1.475	1.527
28	3	48	15	349	1.657	1.657	0.528	0.261	0.430	1.507	0.368
29	9	28	73	730	2.620	2.620	0.542	0.380	0.267	4.122	2.028
30	6	10	94	502	2.098	2.098	0.377	0.220	0.217	3.017	1.031

There are two CCA's with DE less than one using the unbiased estimator, and eight CCA's with DE less than one using the ratio estimator for the fluoridated aged 15 results. Two of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator. This fluoridated age group also shows a large difference in the number of DE's less than one when comparing the unbiased and ratio estimator. Once again observing the individual DE estimates shows many are close to one, and thus the generally smaller ratio estimator may just bring these borderline DE estimates below one.

Table 3.8: Mean and Variance Estimates of dmft for Non-Fluoridated Observations Age 15

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	5	26	16	476	1.773	1.773	0.510	0.662	0.561	0.827	1.392
10	6	25	17	572	3.110	3.110	1.177	0.699	0.604	3.802	1.339
11	7	18	11	607	5.811	5.811	1.537	1.324	1.174	1.713	1.272
12	8	30	43	756	2.944	2.944	0.586	0.543	0.520	1.270	1.093
13	8	19	16	697	3.555	3.555	0.700	0.545	0.649	1.164	0.707
14	8	21	82	767	2.765	2.765	0.604	0.496	0.331	3.336	2.247
15	8	17	15	994	2.581	2.581	0.775	0.684	0.772	1.009	0.786
16	10	20	42	830	2.124	2.124	0.549	0.470	0.343	2.562	1.878
17	9	24	39	994	2.735	2.735	1.013	0.682	0.583	3.021	1.370
18	10	23	79	631	2.811	2.811	0.778	0.787	0.349	4.960	5.066
19	6	26	14	416	2.804	2.804	0.782	0.473	1.017	0.592	0.216
20	5	16	6	384	1.904	1.904	0.674	0.715	0.631	1.140	1.286
21	2	20	4	251	3.289	3.289	2.737	2.382	1.778	2.369	1.794
22	6	20	24	492	2.365	2.365	0.530	0.585	0.520	1.038	1.264
23	8	29	44	640	1.887	1.887	0.770	0.676	0.522	2.180	1.679
24	11	17	33	869	4.510	4.510	0.730	0.764	0.792	0.851	0.932
25	4	34	25	237	1.432	1.432	0.673	0.733	0.363	3.440	4.082
26	6	25	28	705	3.596	3.596	1.772	1.002	0.487	13.245	4.237
27	6	12	60	380	4.934	4.934	1.240	1.063	0.530	5.468	4.012
28	3	48	3	340	6.656	6.656	1.683	1.322	1.202	1.961	1.211
29	3	28	3	276	3.377	3.377	2.130	2.558	3.381	0.397	0.572
30	3	10	8	210	1.931	1.931	0.608	0.700	0.587	1.075	1.423

The 15 year old non-fluoridated results show the unbiased estimator has four DE's less than one, while the ratio estimator has five DE's less than one. Three of the CCA's presenting with DE's less than one are common to both the unbiased and the ratio estimator. Throughout the results above the ratio estimator has at least as many DE's less than one as the unbiased estimator and in most cases more. This is due to the ratio estimator's smaller variance estimate as it uses auxiliary information in its estimation. Thus, the large variability within some of the class sizes may be the underlying cause of the difference between the unbiased and ratio estimate. This variability suggests the ratio is the better estimator to use, however the unbiased results are presented to show that there are many DE's less than one using unbiased estimation also, and many of them overlap with the ratio estimates also less than one.

Tables of results, following the same format as above, were also calculated for the caries free variable. Caries free is a binary representation of the dmft variable and it was of interest if the DE's less than one would also appear for a proportion,

rather than a mean. Estimates of the proportion of caries free within each CCA across the differing age groupings along with SE and DE estimates were obtained and are presented in Appendix D. The results are very similar to the results for the dmft estimates.

Regardless of the variance estimator used (unbiased or ratio), or the representation of dmft (count or binary), the numerous CCA's with DE's less than one are very interesting results and have not been encountered by the author elsewhere in the literature. In general, when cluster sampling is implemented, there are usually no cluster groupings with DE's less than one (Särndal, Swensson and Wretman, 2003; Wolter, 2007). These results are very unusual in terms of a cluster sampling scheme and worthy of investigation to ascertain if some cause can be attributed. If the cause of these results could be found in general, or if the DE's were always less than one in general for a particular CCA, this would have some big implications for future survey planning. Given the cluster sample from a particular CCA always yielded a DE less than one, this would allow the planner to sample less observations in that cluster than needed for a SRS and have the same reliability of estimates in the results. This reduction in children sampled would create savings in the survey, as estimated number of children would not need to be inflated for the cluster effect, and as mentioned above could be reduced relative to a SRS.

3.4 Identifying Factors Associated With Design Effects Less Than One

Attempting to find the cause of these DE less than one will be the primary aim of the remained of this chapter. The initial idea was to look at comparing the characteristics of the data above to the results of the simulation in the previous chapter (Chapter 2). Recalling the results presented earlier, the DE became less than one only for a very small number of clusters under the Taylor Series Linearization (TSL) method. The results presented above have no CCA with less than four clusters sampled, and thus the TSL DE being less than one is not comparable to the simulation in Chapter Two and must have some other underlying cause. The DE became less than one in the presence of large sampling fractions and small second stage sample sizes for the delete-one Jackknife (JK1) method. Although we do not calculate the SE estimates using the JK1 method,

the scenarios in which it produces a DE less than one were also be investigated. Once again the DE's less than one above do not have the same properties as the simulation as the sampling fraction is in general quite small, however the real data above does have some small second stage sample sizes. Having implemented a cross-tabulation of second stage sample size against DE the CCA's with small second stage samples do not correlate with the CCA's with DE's less than one and thus the exploration to identify an underlying cause for DE's less than one will involve more analysis.

As we have many different age groupings, fluoridation status' and CCA's, looking at the results individually may not be easy to see some underlying cause of DE's less than one. Thus, this chapter will implement a cluster analysis to assess which regions are similar, and this may give an indication to the cause of the DE's less than one. It may be the case regions with a mix of fluoridated and non fluoridated observations that commute may be more likely to have DE's less than one compared to regions composed of mainly one type of observation. Furthermore, linear model methods will be implemented to assess if any measurable attribute of a CCA can be identified as significant in obtaining a DE less than one.

3.4.1 Cluster Analysis

Cluster analysis (Kassambara, 2017) aims to identify patterns or form groups of similar objects within a dataset. There are many different approaches to cluster analysis, with the two most common types being partitioning clustering and hierarchical clustering. Each of these broad headings contains many different methods using different algorithms to achieve groupings within the data.

It should be noted that before cluster analysis is implemented (EasyGuides, 2018), a preliminary analysis of the data should be conducted to assess if cluster analysis is appropriate. The two more commonly used methods include computing the Hopkins statistic, or doing a visual assessment on the dataset based on the dissimilarity matrix of the data. The Hopkins statistic (H) measures the probability that a given dataset is generated by a uniform data distribution, and thus can be used to assess clustering tendency. Let D be the real dataset and let D_{rand} be a simulated dataset from a random uniform distribution with the same variation as D. Taking a sample of n points $(p_1, \ldots, p_n)$ from D and another sample of n points $(q_1, \ldots, q_n)$ from D_{rand}, one finds each points nearest neighbour and computes the distance between them for each of the two samples. Thus,

$x_i = dist(p_i, p_j)$ is the distance between each point and its nearest neighbour for the real dataset (D), and $y_i = dist(q_i, q_j)$ is the distance between each point and its nearest neighbour the random (simulated) dataset (D_{rand}). The Hopkins statistic is

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}.$$

A value of H close to 0.5 indicates there are no meaningful clusters in the data. The *hopkins*() function in R can be used to perform this calculation, and returns the value $(1-H)$. Ideally the value of $(1-H)$ will be as close to 0 as possible. It is important to implement these preliminary checks, as cluster analysis methods will return a solution, even on data with no clustering effect. Once these conditions have been checked cluster analysis is appropriate, and usually implemented using one of the aforementioned methods, however many alternative clustering methods may also be used.

All methods of cluster analysis require some method of computing the measure between each pair of observations to form groups. This measure gives the similarity or dissimilarity between groupings. The choice of measure is critical when implementing cluster analysis as each of the groupings formed will be based on this measure.

3.4.1.1 Types of Measure

The classical definition of distance is the Euclidean distance which is expressed as

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where x and y are two vectors of length n. Other measures include the Manhattan distance or correlation based distances such as the Pearson correlation distance, the Eisen cosine correlation distance and the Spearman correlation distance. Correlation based measures consider two objects to be similar if their features are highly correlated, regardless of how large the Euclidean distance between them. This is useful for identifying observations with similar profiles regardless of their magnitude. However, in most cases the classical Euclidean distance measure is

used, and is the default in most software for implementing cluster analysis, particularly when the magnitude of the observations is important.

Once the choice of measure is selected, a dissimilarity matrix is calculated to help form the clusters. To ensure the equality of distance measures across different variables, a preliminary step before cluster analysis involves standardizing each of the variables, especially when the variables are measured on different scales. If this step is overlooked, the dissimilarity matrix will be severely affected. When scaling variables the following formula is commonly used

$$\frac{x_i - centre(x)}{scale(x)}$$

where $centre(x)$ can be the mean or median, and $scale(x)$ can be the standard deviation or the interquartile range. A plot of the dissimilarity matrix can often help in assessing the relationship between variables.

3.4.1.2 Partitioning Clustering

Partitioning clustering classifies the dataset into a user specified number of groups (clusters). Common methods for partitioning clustering include K-means clustering, K-medoids or PAM (Partitioning Around Medoids) clustering and CLARA (Clustering Large Applications) algorithm clustering. K-means clustering is the most commonly used method in which the mean of the data points belonging to each cluster represents each cluster centre (value). PAM allows one object within each cluster to represent the cluster, and is thus less sensitive to outliers compared to k-means while CLARA is an extension of PAM to larger datasets.

K-means (MacQueen, 1967) clustering is implemented by defining groups which minimize the intra-cluster (within-cluster) variation. K-means is the most commonly used unsupervised machine learning algorithm for partitioning a dataset into k groups and has many different algorithms for implementation. The standard algorithm uses the sum of the squared Euclidean distances between observations and the corresponding cluster centroid to measure the within-cluster variation (Hartigan and Wong, 1979). Thus,

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

where x_i design a data point belonging to cluster C_k and μ_k is the mean value of the points assigned to cluster C_k.

The compactness of the cluster analysis is assessed using the total within cluster sum of squares. The total within-cluster sum of squares is measured using

$$SSW = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

and ideally this will be as small as possible. The smaller the value of the SSW the better the fit of the cluster analysis.

Steps of K-means Algorithm: A brief description of the steps in the k-means algorithm is as follows. Once the number of clusters (k) is chosen (user input), the algorithm randomly assigns k cluster means (centroids) using random values in the dataset. Next, using the Euclidean distances, each of the remaining observations are assigned to its closest centroid. Once all observations have been assigned and clusters formed, the algorithm computes a new centroid value. Using this recalculated centroid, all observations are then reassigned (if necessary) to the closest cluster centroid. This recalculation of centroids is repeated until the cluster assignments converge to a solution (or the maximum number of iterations is achieved).

3.4.1.3 Hierarchical Clustering

Hierarchical clustering is an alternative approach to partitioning clustering, usually subdivided into two types namely agglomerative clustering and divisive clustering. Agglomerative clustering begins with each cluster on its own, and merges the two most similar clusters at each step until all clusters are merged into one overall cluster. Divisive clustering begins with one overall cluster and successively divides the most heterogeneous clusters until all observations are in their own cluster. Hierarchical methods offer the advantage that the number of clusters need not be specified, and groupings can be implemented after the cluster method is completed. Hierarchical cluster methods are usually presented in dendrograms, which is a tree based representation of the clusters merged/split at each step. Once again, the choice of measure is an important feature that will influence the results, and in general Euclidean distance is the default method used.

Steps of Agglomerative Clustering Algorithm: Agglomerative clustering begins by computing the dissimilarity information between each pair of observations in the data. Using this information the clusters that are closest in proximity to each other are merged at each step. Once the dendrogram is formed, the clusters can be formed by cutting the tree at the user decided level.

3.4.1.4 Clustering Results

Initially the cluster analysis was completed on the dmft DE estimates at CCA level, for the fluoridated observations across all age groups. Fluoridated and non-fluoridated regions were treated separately as CCAs 1 to 8 had no non-fluoridated values present. Note also that the fluoridated aged 5 DE value for CCA4 (Kilbarrack) was imputed as the mean of the three other DE values, due to missing data and this appears in bold in Table 3.9. Thus the input data for the first six CCA's looked like:

Table 3.9: Initial form of DE Values for Cluster Analysis

CCA	DE_age5	DE_age8	DE_age12	DE_age15
Dun Laoghaire	1.306	0.918	2.243	2.555
Wicklow	0.807	0.521	1.675	0.852
Coolock	1.510	1.599	1.517	1.739
Kilbarrack	**0.865**	0.725	1.349	0.521
Roselawn	0.618	1.405	1.068	1.089
Cornmarket	1.221	1.043	1.683	2.044

The input data was not scaled, as each of the variables were measured on the same scale. Euclidean distances were calculated to form the dissimilarity matrix. The Hopkins test statistic was used to assess if cluster analysis would be useful on this dataset, and to assess if meaningful clusters could be formed. The resulting value was $(1 - H) = 0.27$ indicating some potential for the formation of meaningful clusters. Next some preliminary calculations were completed to assess the number of clusters to use. Choosing the number of clusters (k) involved computing the k-means clustering for different values of k, and calculating the within sum of squares (wss) according to the number of clusters. The location of the bend in the plot generally is considered as the appropriate number of clusters.

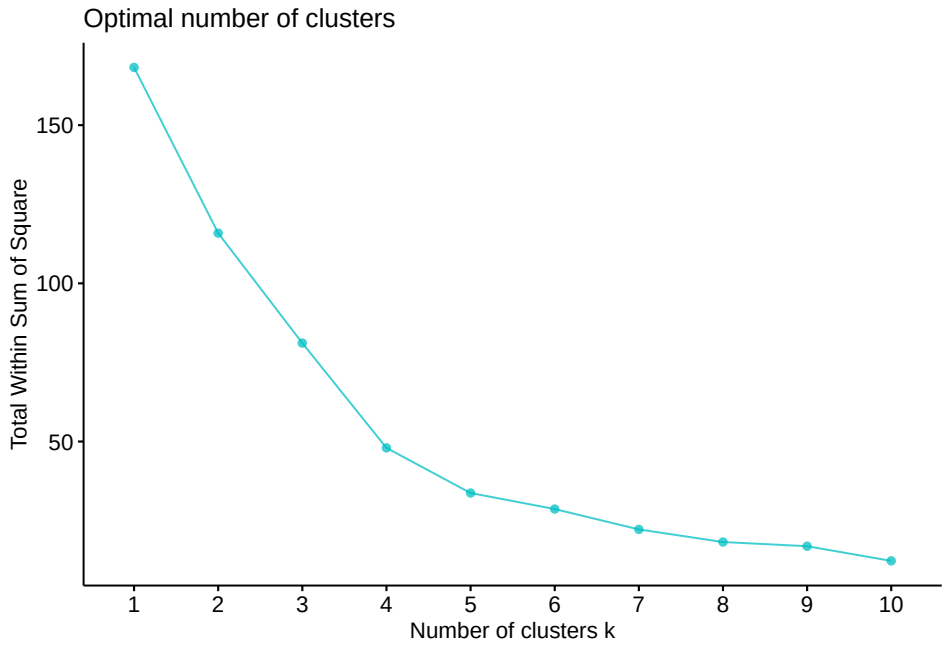


Figure 3.1: Within Sum of Squares against Number of Clusters Used

Based on the plot above, 5 clusters was chosen as the knee of the plot. The cluster analysis was implemented with k=5 clusters and 100 different starting points. As the choice of starting clusters is important it is useful to try many random start points. Figure 3.2 is a plot of the $k = 5$ cluster groupings computed using the k-means algorithm. As the input data contained 4 variables, Figure 3.2 shows the two principal components for each cluster. For reference, the raw input data along with the assigned cluster from the k-means analysis are presented in Table E.1 in Appendix E.

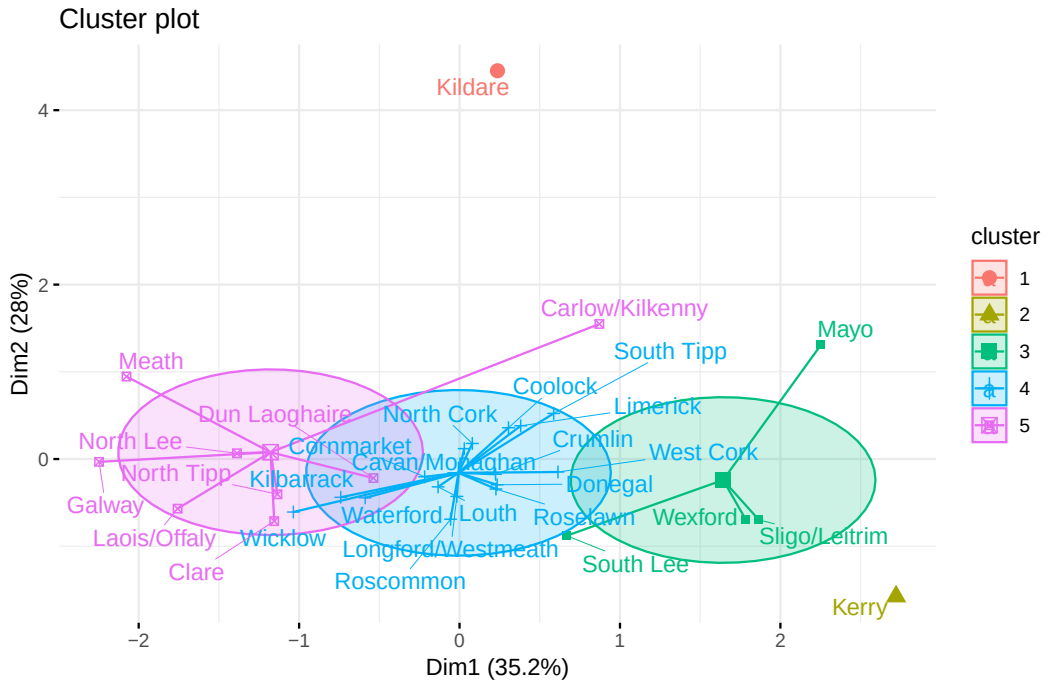


Figure 3.2: Cluster Analysis of Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups

Observing Figure 3.2 and Table E.1, two clusters (Kerry and Kildare) are formed with just one CCA in each. Looking at the raw input data, this is due to large DE values for both these CCA's under one age group. As k-means is quite sensitive to outliers, and as this analysis is exploratory looking for some meaningful cluster groupings which might help indicate the presence of a DE less than one, these large values were replaced with the mean of the DE estimates for the other three age groupings and the cluster analysis was re-run. The results are presented in Appendix E (Section E.1.2).

Further variables frequently used in oral health studies, including BMI and decayed missing and filled surfaces (dmfs), were included into the cluster analysis, however, their addition did not help the clustering process form defined, interpretable clusters. From a practical standpoint, this exploratory analysis did not show any geographical groupings of interest, or any distinct cluster groupings that could help explain the DE's less than one.

As there are many clustering techniques and algorithms, a more conventional approach to cluster analysis, hierarchical clustering, was also implemented, namely agglomerative clustering. This algorithm starts with all clusters being individual and joins the closest two based on the chosen measure and continues this until all observations form one cluster. As the results from this method of cluster

analysis also showed no meaningful groupings, the results are relegated to Appendix E (Section E.1.3), where one can find the description of implementation and the associated results. Moreover, an identical analysis was conducted for the 22 non-fluoridated CCAs, however again no useful groupings were formed using both partitioning and hierarchical clustering methods. The results of the non-fluoridated cluster analysis are presented in Appendix E (Section E.2). As no clear groupings which may account for the presence of DE's less than one were apparent from the cluster analysis the next step was to use linear model methods and assess if any coefficient was statistically significant in identifying when the design effects were less than one.

3.4.2 Linear Model Analyses

3.4.2.1 Linear Regression of dmft DE

The initial method of exploration looked at fitting numerous simple linear regression models to the raw data. These models took the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

where

$$\begin{aligned} Y &= \text{Dependent Variable} \\ X_1 &= \text{Independent Variable} \\ \varepsilon_i &\sim NID(0, \sigma^2) \end{aligned}$$

The purpose of this exploratory analysis was to assess which of the independent variables, if any, would be of use to indicate a design effect value less than one. Many different explanatory variables were used including:

- the number of clusters sampled in the CCA (n)
- the number of population clusters in the CCA (N)
- the average number of observations sampled per cluster in the CCA ($\bar{m}$)
- the average number of population observations per cluster in the CCA ($\bar{M}$)
- Fluoridation Status (binary)
- Age Group (categorical) {Age5, Age8, Age12, Age15}

- the sampling fraction at stage one of sampling in the CCA ($\frac{n}{N}$)
- the average sampling fraction at stage two of sampling in the CCA ($\frac{\bar{m}}{M}$)

Some of the independent variables are combinations of the others, however they are included as their interpretation can be useful within a clustered sample setting. The regression was carried out in R software using dummy variables to represent fluoridation status, medical card status, and age group.

Results

The initial level of significance in these simple linear regression models was set at 20%. Once the initial predictors were chosen from the simple regression models, multiple regression models were fitted (Hair, 2006; Weisberg, 2005). The level of significance required was then lowered to 5% and manual backwards elimination was used to select the final model. The backwards selection technique involved fitting a multiple regression model and then removing the independent variable with the largest p-value and refitting multiple regression model again, until only significant variables were left in the model.

The dependent variable ($Y = DE$) was transformed using a log transformation, as the DE values had a skewed distribution, due to some large DE values present in the data. The selection procedure described above was implemented on the transformed dependent variable $Y = log(DE)$, using each of the response variables in simple linear regression models. Each of the simple linear models were fitted initially using a 20% significance level. The transformed dependent variable showed three of the predictors to be statistically significant at the 20% level. Fitting the multiple regression model with predictors n, $\frac{n}{N}$ and $\bar{m}$, and using backwards selection, the final model reduced to a simple linear regression model with n, the number of clusters sampled per CCA, as the chosen predictor. The results are shown in Table 3.10. The R-squared value for the model was 0.03. Checks for collinearity in the multiple regression model were carried out as including both n and $\frac{n}{N}$ would suggest some degree of collinearity. However, the correlation values were weak (<0.45) and the variance inflation factors were 1.27, 1.55 and 1.26 respectively suggesting collinearity was not an issue.

Table 3.10: Simple Linear Regression Estimates for $log(DE_{(dmft)})$

Simple Models	β	SE	p-value
n	0.012	0.005	0.012
$\frac{n}{N}$	0.819	0.319	0.011
$\bar{m}$	0.035	0.018	0.054
Chosen Model	β	SE	p-value
n	0.012	0.005	0.012

It is noteworthy that the simple linear regression model using sampling fraction at stage one of a CCA ($\frac{n}{N}$) as the independent variable was also significant at the 5% level suggesting some degree of interchangeability amongst the predictors, and thus suggesting some degree of collinearity between the predictors. The residual checks of the chosen simple linear regression model ($log(DE) \sim n$) are presented in Appendix F. The residuals look to satisfy the residuals assumptions, and thus this model is chosen as the most useful of the regression models.

Finding a statistically significant parameter which could help identify whether the DE is less than one would be a very a useful result. The estimate from the above regression model indicates that one may need to sample less clusters than originally thought to reduce the design effect. However one must bear in mind that the R-squared value for the model is quite low, and of course some restrictions will need to be implemented to optimize n, as it cannot be reduced to zero.

3.4.2.2 Logistic Regression of dmft DE

Logistic regression was also implemented on the DE value, assessing if the magnitude of the DE had any affect when looking to identify a DE less than one. The response variable took the form

$$y_i = \begin{cases} 1 \text{ if } DE < 1. \\ 0 \text{ Otherwise.} \end{cases}$$

The same predictors as used in the linear regression model $\{n, N, \bar{m}, \bar{M}, \frac{n}{N}, \frac{\bar{m}}{\bar{M}}$, fluoridation status (binary) and age group (categorical)$\}$ will be considered again.

Similar to the linear regression, the logistic regression will again conduct a preliminary analysis at 20% significance. The logistic model took the form

$$\text{Random Component: } y_i|x_i \sim B(n_i, p_i)$$
$$\text{Systematic Component } g(\mu_i) = \eta_i = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x$$

for the continuous predictors, with additional dummy variables being used to represent the categorical predictors.

The simple logistic regression models showed six of the predictors to be statistically significant at the 20% level, and these are shown in Table 3.11. Fitting the multiple logistic regression model with predictors Age Group (categorical), $\frac{n}{N}$, n, N, $\bar{m}$ and $\bar{M}$ and using backwards selection, the final model reduced to a simple logistic regression model with $\bar{m}$, the average number of observations at second stage sampled per CCA, as the chosen predictor. It is noteworthy that the sampling fraction at stage one of a CCA ($\frac{n}{N}$) was also significant at the 5% level suggesting some degree of interchangeability amongst the predictors. The residual deviance values of both the logistic regression models was again quite large. The residual diagnostic plots for the model $logit(p) \sim \bar{m}$ are presented in Appendix F.

Table 3.11: Logistic Regression Estimates for $log(DE_{(dmft)})$

Simple Models	β	SE	p-value
Age 8 (Factor)	0.211	0.403	0.601
Age 12 (Factor)	−0.115	0.411	0.780
Age 15 (Factor)	−0.577	0.432	0.181
$\frac{n}{N}$	−2.481	1.117	0.026
n	−0.031	0.022	0.146
N	0.005	0.003	0.098
$\bar{m}$	−0.134	0.059	0.025
$\bar{M}$	−0.007	0.005	0.155
Chosen Model	β	SE	p-value
$\bar{m}$	−0.134	0.059	0.025

As both regression analyses highlighted some potential predictors that might be useful indicating the presence of a DE less than one, these regression analyses were

also carried out on the variable caries free to assess if the same characteristics would be useful indicating the presence of DE's less than one. As mentioned earlier, caries free is a binary representation of an observations dmft values with:

$$Caries = \begin{cases} 1 \text{ if dmft } = 0. \\ 0 \text{ Otherwise.} \end{cases}$$

The estimates of the proportion of caries free observations, along with SE and DE estimates were calculated for each CCA and presented in Appendix D. The regression analyses will be carried out on the caries free DE estimates again using the same independent variables used in the dmft regression analyses.

3.4.2.3 Linear Regression of Caries Free DE

Again, the simple linear regression models for the caries free data were fitted on the raw DE estimates $Y = log(DE)$, due to the skewed nature of the DE estimates. The results of the transformed dependent variable showed six predictors to be statistically significant at the 20% level. Fitting the multiple logistic regression model with predictors Fluoridation Status (categorical), Age Group (categorical), stage-one sampling fraction ($\frac{n}{N}$), stage-two sampling fraction ($\frac{\bar{m}}{\bar{M}}$), number of clusters sampled (n) and average number of population SSU's ($\bar{M}$), then using backwards elimination, the final model was a multiple regression model containing three variables, Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{\bar{M}}$. The parameter estimates, standard errors and p-values are given in Table 3.12.

Table 3.12: Simple Linear Regression Estimates for $log(DE_{(Caries)})$

Simple Models	β	SE	p-value
Fluoride = Yes (Factor)	-0.204	0.083	0.016
Age 8 (Factor)	-0.225	0.118	0.058
Age 12 (Factor)	-0.030	0.118	0.801
Age 15 (Factor)	-0.143	0.119	0.231
$\frac{n}{N}$	0.463	0.271	0.089
$\frac{\bar{m}}{\bar{M}}$	1.093	0.442	0.014
n	0.012	0.004	0.003
$\bar{M}$	-0.002	0.001	0.193
Chosen Model	β	SE	p-value
Fluoride = Yes (Factor)	-0.284	0.083	0.001
$\frac{n}{N}$	0.678	0.264	0.011
$\frac{\bar{m}}{\bar{M}}$	1.591	0.444	< 0.001

3.4.2.4 Logistic Regression of Caries Free DE

Looking at the binary representation of the caries free DE values, the simple logistic regression models at the 20% significance level showed four predictors to be statistically significant. Fitting the multiple logistic regression model with predictors stage-one sampling fraction ($\frac{n}{N}$), stage-two sampling fraction ($\frac{\bar{m}}{\bar{M}}$), number of clusters sampled (n) and average number of population SSU's ($\bar{M}$), then using backwards elimination, the final model reduced to a simple logistic regression model with predictor n. Again it is noteworthy that $\frac{\bar{m}}{\bar{M}}$ is also significant at the 5% significance level possibly suggesting some interchangeability. The parameter estimates, standard errors and p-values of the logistic regression are given in Table 3.13, with the residual checks in Apprendix F.

Table 3.13: Logistic Regression Estimates for $log(DE_{(Caries)})$

Simple Models	β	SE	p-value
$\frac{n}{N}$	-1.729	1.294	0.181
$\frac{\bar{m}}{\bar{M}}$	-5.105	2.032	0.012
n	-0.174	0.048	< 0.001
$\bar{m}$	-0.112	0.071	0.113
Chosen Model	β	SE	p-value
n	-0.174	0.048	< 0.001

Once again a one predictor model was chosen as the best fit for the DE data. The residual deviance value of the logistic regression model was again quite large. The residual plots of the logistic regression model for caries free are presented in Appendix F. Table 3.14 gives a summary of the final regression model chosen under both linear and logistic regression for both the dmft and caries free variable.

Table 3.14: Final Models and Significant Univariate Parameters for DE's less than one

Variable	Linear (dmft)	Logistic (dmft)	Linear (caries)	Logistic (caries)
n	$\diamond$	$\star$	$\star$	$\diamond$
N		$\star$		
$\bar{m}$	$\star$	$\diamond$	$\star$	$\star$
$\bar{M}$		$\star$		
$FluoridationStatus$			$\diamond$	
$MedicalCardStatus$				
$AgeGroup$		$\star$	$\star$	
$\frac{n}{N}$	$\star$	$\star$	$\diamond$	$\star$
$\frac{\bar{m}}{\bar{M}}$			$\diamond$	$\star$

$\diamond$ = Variable Selected in the Final Model (5% Significance)

$\star$ = Variable Significant in Univariate Model Only (20% Significance)

3.4.3 Concluding Remarks

Chapter 3 looked at calculating SE and DE estimates for an Irish national children's oral health sample. The presence of many DE's less than one in a cluster sample, using two differing calculation methods, was a result not encountered by the author before. As a result, an exploratory analysis was undertaken to try and ascertain an underlying cause. Comparisons to Chapter 2 showed no similarities identifying where DE's less than one occur, so cluster analysis methods were implemented to assess if some geographic cause could be attributed. When cluster analysis methods failed to attribute an underlying cause, linear model methods were then implemented. Linear models suggested that n, the number of clusters sampled showed a relationship with the DE, suggesting that as n decreased, so did the DE value. However, n could not be reduced too severely as there is a risk of SE (and as a result DE) increasing again if n is decreased too much. Furthermore, if n is reduced too severely one may obtain a biased sample. If some reduction in n were possible, and this optimum value could be calculated, this could introduce a large cost saving into future surveys, and this will be the focus of Chapter 4.

Chapter 4

Analysing the Effect of Sampling Fewer Clusters on Bias and SE of dmft Estimates

4.1 Outline

This chapter looks at the effect of reducing the number of clusters (schools) sampled (n) on the SE and DE estimates of the survey data analysed in Chapter 3. Section 4.3 details the methodology used to calculate the SE and DE estimates when reducing (n) the number of clusters sampled. Section 4.4 presents the results of the SE and DE estimates across all CCA's and all age groupings, while Section 4.5 gives some concluding remarks for the chapter.

4.2 Motivation

After identifying some variables in Chapter 3 which may help indicate when the design effect (DE) will be less than one, the next step involved investigating how changing these variable values affected the SE and DE estimates. The most promising variable which may indicate a CCA's DE was n, the number of clusters sampled. The linear models in Chapter 3 suggested that reducing the number of clusters would in turn reduce the DE of the CCA, however one cannot reduce the number of clusters too much as SE will eventually become large enough to give a DE greater than one, or greatly increase a DE already greater than one.

Furthermore, reducing the number of clusters sampled may introduce bias into the mean estimate, as the sample may no longer be representative of the population. This chapter aims to assess the effect of reducing the number of clusters (n) on the bias and SE of the mean dmft and proportion of caries free estimates in each CCA. Although related, the variables chosen are routinely used in oral health surveys and are key variables in assessing the effect of caries in a population (WHO, 2013).

4.3 Methods

The analysis conducted in this chapter was completed on the same dataset used in Chapter 3 and the description of this dataset can be found in Section 3.3.1. As variable estimates are presented at CCA level in Chapter 3 (Table 3.1 to Table 3.8), the results generated in this chapter will also be generated at CCA level to allow comparison to these original estimates. The reduction in n, the number of clusters, will be carried out at CCA level for all of the CCA's within the sample, as no definitive underlying characteristic could be found in Chapter 3 to identify the CCA's with DE's less than one.

The reduction in the number of clusters sampled (n) was implemented using a Jackknife procedure for each number of clusters omitted. In the case of $n = 1$ clusters removed, one of the sampled clusters (i.e. one school) was omitted within each CCA, and estimates of the mean and SE for both the dmft and caries free variable were obtained. This cluster was then replaced and another of the clusters was omitted, and the updated mean and SE estimates were recorded. The iterative process was completed omitting each of the clusters in turn and obtaining estimates of the mean and SE of the dmft and dmfs variables. In the case of $n = 2$ clusters being omitted, each pair of schools was omitted in turn with the means and SEs being recorded. This was repeated for $n = 3, 4, 5, \ldots,$ until five clusters remained within the CCA, or n reached a maximum value of ten clusters dropped.

Five clusters were left in each CCA to try and ensure reliable estimates of the SE. Although throughout the literature there is no explicit definition of a small number of clusters, Koop (1972) suggests that the TSL estimator requires $n \geq 5$ clusters to ensure the estimate is reliable. Although linearisation is not implemented (closed form estimators suffice), due to the two-stage cluster design used here, one needs to ensure the estimator is not biased downwards due to the small

cluster problem. Again, while no hard cut off is given, Cameron and Miller (2015) suggests this happens with small n, particularly in the case of $n = 2$. Furthermore, in the simulations conducted in Chapter 2, the estimator described as TSL (which in the case of Chapter 2 uses the same closed form expression as the design in this Chapter) showed problems with small n. Using information from the above three sources, and in the absence of an explicit definition of small, the cut off of $n \geq 5$ was chosen.

In the few cases where $n > 20$, the number of combinations became too large to obtain all replications, and thus only a random sample of the estimates were obtained. Once the number of combinations exceeded 420,000, resampling was used to obtain 420,000 random independent replications. This cut off was chosen to ensure the runtime was of a reasonable length, while also trying to maintain reliable estimates of the mean and SE of the dmft and caries free variables. It is noteworthy that in cases where resampling was required (n being large), the effect on the means and SE estimates is likely to be less than in cases where n is small (i.e. omitting 7 clusters out of 50 should have less of an effect than omitting 7 clusters out of 12).

Once all the mean and SE estimates of dmft and the proportion caries free were produced within each CCA for each value of the reduction in number of schools selected (n), the mean of these estimates was taken and compared to the original estimates. This allowed a percentage change in both the mean and SE to be obtained relative to the estimates from the original survey. This allowed the bias and increase in SE to be explored within each CCA, for each value of the reduction in number of schools selected, across the eight combinations of fluoridation status and age group. The results are shown in Table 4.1 to Table 4.16. These tables also present the average, max and 90th percentile of the percentage change in mean across all CCAs, for each value of the reduction in number of schools selected.

In some cases the number of clusters dropped could not reach 10, as there were less than 15 clusters within a CCA. This is particularly common for the Age 15 grouping as there are generally less secondary schools per CCA, due to bigger schools. However, this problem occurs in the other age groups too. To ensure a small number of schools were not the sole influence in obtaining estimates for a large reduction in the number of schools selected (i.e. selecting 7-10 schools less per CCA), the final estimate with $n = 5$ clusters remaining was continued across the CCA to give estimates for the remaining reduction in schools estimates. This is highlighted to the reader by shading the cell grey. Furthermore, the estimates

of the percentage change in SE for $n = 4, 3, 2$ clusters remaining are given in the table and are shown in brackets using small font. These estimates show the unreliable nature of the SE estimator when the number of clusters would become too small ($n < 5$).

The DE is represented in these tables using bold text. If the CCA had a DE less than one in Chapter 3, this is represented by bolding the CCA name (i.e. CCA 2 appears in bold for Fluoridated Age 5 results). The new SE estimate for each reduction in the number of schools selected was then compared to the original SRS SE estimate. If the new SE estimate was lower than the original SRS SE estimate (indicating a DE less than one) this percentage estimate is bolded in the table. Thus for CCA 2 in the fluoridated Age 5 group, although the SE increases by on average 4% when the number of schools is reduced by one, this new estimate of SE is still less than the SRS SE estimate for the full sample from CCA 2, and thus the result appears in bold in the table.

The effect of reducing the number of clusters selected on SE and bias are presented in Table 4.1 to Tables 4.16 for the dmft variable. The results of the caries free analysis lead to almost identical results and hence are presented in Appendix G (Table G.1 to Table G.16).

4.4 Results

4.4.1 Tables of Percentage Change in SE and Bias by Reduction in Number of Fluoridated Schools Selected

The fluoridated age 5 results (Table 4.1 and Table 4.2) have an empty row for CCA 4 as no observations were sampled for this particular combination of age, fluoridation status and CCA. This result was seen previously in Chapter 3. Thus, the estimates for column averages, percentiles and maximums are obtained from 29 CCA estimates. Of the 29 CCAs, only 11 of these geographical regions had $n \geq 15$ clusters to allow observed estimates of the change in SE and bias at each reduction in the number of schools selected (i.e. from one to ten). The remaining CCA's relied on continuing the estimate of SE and bias change across the row, and this is observed by the shaded cells within the table.

Table 4.1: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	4	9	15	21	29	38	49	49 (63)	49 (80)	49 (93)
CCA 2	**4**	**8**	13	18	25	33	33 (43)	33 (56)	33 (65)	33
CCA 3	4	8	12	17	22	28	33	33 (39)	33 (42)	33 (35)
CCA 4										
CCA 5	**4**	**9**	**15**	**21**	29	38	49	63	63 (82)	63 (107)
CCA 6	5	11	18	26	36	48	48 (65)	48(86)	48 (110)	48
CCA 7	4	8	12	17	22	27	33	40	48	48 (55)
CCA 8	4	8	12	17	23	29	37	37 (45)	37(51)	37 (45)
CCA 9	**3**	**5**	**8**	12	16	20	25	30	36	42
CCA 10	**4**	**8**	**12**	**18**	24	31	39	49	61	75
CCA 11	**6**	**14**	**23**	**36**	**53**	53 (78)	53 (117)	53 (172)	53	53
CCA 12	2	5	7	8	9	9 (7)	9 (1)	9 (-7)	9	9
CCA 13	**4**	**8**	13	18	24	31	39	49	61	75
CCA 14	1	3	5	6	8*	9*	11*	13*	15*	17*
CCA 15	1	3	4	6	7*	9*	11*	12*	14*	16*
CCA 16	1	2	4	5*	6*	8*	9*	10*	12*	13*
CCA 17	**3**	**5**	**9**	**12**	16	21	27	34	42	42 (51)
CCA 18	**5**	**10**	**16**	**22**	**22 (28)**	**22(31)**	**22 (19)**	22	22	22
CCA 19	4	8	12	17	22	28	35	42	50	58
CCA 20	5	10	15	21	28	36	44	44 (54)	44(66)	44(76)
CCA 21	**4**	**8**	**13**	**18**	**24**	32	40	51	64	79
CCA 22	**5**	**11**	**18**	**27**	**38**	**51**	**51 (66)**	51 (79)	51 (79)	51
CCA 23	3	6	10	14	18	23	28	34	41	49
CCA 24	3	7	10	14	19	23	28	32	37	41
CCA 25	4	9	14	21	28	36	46	59	75	75 (95)
CCA 26	4	7	12	17	22	29	37	46	58	58 (72)
CCA 27	**4**	**9**	14	21	28	37	48	48 (61)	48(77)	48(84)
CCA 28	**4**	8	13	19	26	34	34(43)	34 (50)	34 (43)	34
CCA 29	2	5	7	10	12	15	18	22*	25*	29*
CCA 30	3	7	11	16	21	27	35	42	50	50 (56)
Col Avg	4	8	12	17	23	28	34	38	42	45
Col Max	6	14	23	36	53	53	53	63	75	79
Col 90th Percentile	5	10	16	23	30	40	49	51	61	75
Num CCAs	29	29	29	29	29	29	29	29	29	29

- **NUMBER** = DE< 1 vs original SRS SE even with NUMBER % increase in SE
- **CCA X** = CCA X has DE < 1 in original data
- NUM = $n < 5$ clusters left in CCA so estimate for n=5 (NUM) used to fill row
- (NUM) = actual estimate of SE with $n < 5$ clusters
- * = resampling used rather than taking all replications

Reducing the number of clusters selected by two across all CCA's shows on average an 8% increase in SE, while the 90th percentile for reducing the number of clusters by two was 10%. As the number of clusters sampled is reduced, naturally the SE estimates increase. While the percentage change in SE may not seem large, the reader needs to be aware that each of the cell entries is a mean percentage change per CCA for a particular reduction in the number of schools made up of all the combinations of dropping this number of schools. Thus, while the entry in the top left of Table 4.2 suggests sampling one less school in CCA 1 will lead to a 4% increase in SE, this is an average of all the combinations of sampling one school less. Sometimes the percentage SE change will be greater than 4% and sometimes it will be less. The $*$ symbol in the table represents estimates where the total number of combinations were too large to use the complete Jackknife resampling procedure, and thus 420,000 independent replications were used to obtain the particular cell estimate.

Table 4.2: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	0	1	1	1	1	1 (2)	1 (3)	1 (5)
CCA 2	**1**	**2**	4	5	7	10	10 (12)	10 (16)	10 (20)	10
CCA 3	0	0	-1	-1	-2	-2	-3	-3 (-5)	-3 (-8)	-3 (-13)
CCA 4										
CCA 5	**0**	**0**	**1**	**1**	1	2	3	4	4 (5)	4 (8)
CCA 6	0	0	1	1	2	3	3 (4)	3 (6)	3 (11)	3
CCA 7	0	0	0	0	0	0	0	0	0	0 (1)
CCA 8	0	1	1	2	3	4	6	6 (8)	6 (12)	6 (18)
CCA 9	**0**	**0**	**0**	-1	-1	-1	-2	-2	-3	-4
CCA 10	**0**	**0**	**0**	**0**	0	1	1	1	1	2
CCA 11	**0**	**1**	**1**	**2**	**3**	3 (5)	3 (8)	3 (12)	3	3
CCA 12	0	0	0	0	-1	-1 (-1)	-1 (-2)	-1 (-6)	-1	-1
CCA 13	**0**	**0**	1	1	1	1	2	2	3	4
CCA 14	0	0	0	0	0*	0*	0*	0*	0*	0*
CCA 15	0	0	0	0	0*	0*	0*	0*	0*	0*
CCA 16	0	0	0	0*	0*	0*	0*	0*	0*	0*
CCA 17	**0**	**0**	**-1**	**-1**	-1	-2	-2	-3	-4	-4 (-5)
CCA 18	**0**	**0**	**0**	**0**	**0 (0)**	**0 (-1)**	**0 (-3)**	0	0	0
CCA 19	0	1	1	1	2	2	3	4	5	6
CCA 20	0	0	0	1	1	1	2	2 (3)	2 (4)	2 (5)
CCA 21	**0**	**0**	**0**	**0**	**0**	0	0	0	0	-1
CCA 22	**0**	**0**	**0**	**1**	**1**	**1**	**1 (1)**	1 (0)	1 (0)	1
CCA 23	**0**	0	0	0	0	0	0	-1	-1	-1
CCA 24	0	0	0	0	0	1	1	1	1	1
CCA 25	0	1	1	2	3	4	5	7	10	10 (14)
CCA 26	0	0	1	1	1	2	2	3	4	4 (5)
CCA 27	**0**	**0**	0	0	1	1	1	1 (1)	1 (2)	1 (2)
CCA 28	**0**	1	2	3	4	5 (8)	5 (12)	5 (19)	5	5
CCA 29	0	0	0	0	0	0	0	0*	0*	0*
CCA 30	0	0	0	0	1	1	0	0	-1	-1 (-2)
Col Avg	0	0	0	1	1	1	1	2	2	2
Col Max	1	2	4	5	7	10	10	10	10	10
Col 90th Percentile	0	1	1	2	3	4	5	5	5	6
Num CCAs	29	29	29	29	29	29	29	29	29	29

The results of the bias estimates show that the dmft estimate will remain close to unbiased even if the number of schools sampled is reduced by as many as 10, as the bias is, on average, only 2%. However, again, these are mean estimates and the

maximum bias estimate is as much as 10%. This estimate of 10% again represents a mean, and thus, depending on the random resample, this value could increase or decrease. It is noteworthy that omitting 10 clusters is a dramatic change to make, and is unlikely to be used in practice. The estimates are produced to assess the potential impact, however it is much more likely the increase in SE will rule out such a large reduction in the number of clusters sampled.

Changing focus to the estimates presented in small text in brackets, (i.e the estimates when $n < 5$), the justification to exclude these estimates seems validated. The estimates become, in general, quite large and as a result these would never be considered as a potential reduction in the number of clusters selected. Furthermore, one notices the small cluster problem with some of these estimates for $n < 5$, particularly in the case of $n = 2$. This, seemingly counter intuitive result suggests that sampling 10 clusters less will give a smaller SE than the original SE estimate for the CCA. However, while the estimate of the SE is small, the bias is large and thus, this smaller resample may not capture the full information of the CCA and is ruled out on this reasoning. Furthermore, this small SE estimate could be attributed to the small cluster problem and may indicate the potential need for alternative SE estimation methods, and these are described in Cameron and Miller (2015). However these methods are beyond the requirement of this Chapter, which looks at using routinely available SE estimators, and would rule out sampling two clusters on the basis of the estimator being biased.

The SE and bias estimates for the eight year olds are given in Table 4.3 and Table 4.4.

Table 4.3: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	**4**	8	13	18	24	32	40	50	50 (63)	50 (78)
CCA 2	**4**	**9**	**13**	**19**	**19 (24)**	**19 (27)**	**19 (19)**	19	19	19
CCA 3	4	9	14	20	27	36	36 (47)	36 (61)	36 (70)	36
CCA 4	**4**	**8**	**13**	18	24	30	37	45	55	66
CCA 5	4	8	13	18	24	29	35	35 (41)	35 (45)	35 (44)
CCA 6	4	7	12	17	22	29	36	45	56	56 (69)
CCA 7	3	7	11	16	21	26	32	39	47	47 (55)
CCA 8	5	10	16	24	33	44	44 (58)	44 (77)	44 (106)	44
CCA 9	**4**	**8**	**13**	**18**	**25**	**32**	41	52	66	84
CCA 10	2	5	8	10	13	16	18	18 (19)	18 (18)	18 (16)
CCA 11	**5**	**11**	17	26	35	35 (46)	35 (58)	35 (64)	35	35
CCA 12	5	11	18	26	26 (36)	26 (46)	26 (45)	26	26	26
CCA 13	**4**	**8**	**12**	18	23	30	38	47	58	72
CCA 14	4	7	12	16	21	27	33	39	39 (45)	39 (47)
CCA 15	**4**	**7**	**12**	16	22	27	34	40	46	46 (51)
CCA 16	**5**	**12**	**19**	**28**	38	50	50 (65)	50 (81)	50 (94)	50
CCA 17	5	12	19	28	38	51	51 (65)	51 (80)	51 (83)	51
CCA 18	(-4)	(-16)								
CCA 19	2	3	5	7	9	11	13	16	19	21
CCA 20	5	10	16	23	30	39	48	48 (58)	48 (63)	48 (45)
CCA 21	4	9	14	20	26	32	37	37 (40)	37 (39)	37 (32)
CCA 22	1	2	4	5	5 (6)	5 (7)	5 (1)	5	5	5
CCA 23	3	6	10	14	19	23	29	34	41	48
CCA 24	4	9	14	20	27	34	43	43 (54)	43 (66)	43 (73)
CCA 25	**3**	**6**	**9**	13	17	21	26	32	39	46
CCA 26	**4**	8	12	17	23	30	39	49	62	62 (79)
CCA 27	5	10	16	23	31	41	52	52 (65)	52 (79)	52 (81)
CCA 28	**4**	**8**	**12**	**18**	23	30	30 (36)	30 (38)	30 (25)	30
CCA 29	2	5	8	11	14	17	21	24	29*	33*
CCA 30	**4**	**8**	**13**	**18**	**23**	**27**	**27 (30)**	**27 (28)**	**27 (15)**	27
Col Avg	4	8	13	18	23	29	34	37	40	42
Col Max	5	12	19	28	38	51	52	52	66	84
Col 90th Percentile	5	11	18	26	33	41	49	50	56	63
Num CCAs	29	29	29	29	29	29	29	29	29	29

The SE results for the fluoridated 8 year olds are comparable to those of the fluoridated five year olds. CCA 18 has no estimates that contribute to the column averages as the number of clusters is $n = 4$. Thus, it is not possible to sample a reduced number of clusters and still have $n = 5$ remaining.

Observing Table 4.3, the CCA's which initially had a DE less than one (bold values), still have SE values less than the SRS SE estimate of the original samples, even after reducing the number of clusters sampled. In some cases this only happens for a reduction of one school sampled, however in other cases 3 or more schools can be dropped and the SE estimate is still less than the SRS SE estimate from the full sample. This result appears throughout the other SE tables for varying age groups.

Table 4.4: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	**0**	0	0	0	0	1	1	1	1 (1)	1 (1)
CCA 2	**0**	**-1**	**-1**	**-2**	**-2 (-3)**	**-2 (-4)**	**-2 (-8)**	-2	-2	-2
CCA 3	1	1	2	3	5	7	7 (10)	7(14)	7 (19)	7
CCA 4	**0**	**0**	**1**	1	1	2	2	3	4	5
CCA 5	0	0	0	0	0	0	1	1 (0)	1 (0)	1 (-1)
CCA 6	0	0	0	0	0	0	0	0	0	0 (0)
CCA 7	1	1	2	2	3	4	5	7	8	8 (11)
CCA 8	0	0	0	1	1	1	1(2)	1 (4)	1 (9)	1
CCA 9	**0**	**0**	**0**	**1**	**1**	**1**	1	2	2	3
CCA 10	0	0	0	0	0	0	-1	-1 (-1)	-1(-3)	-1 (-6)
CCA 11	**1**	**2**	3	4	5	5 (7)	5 (9)	5 (11)	5	5
CCA 12	0	0	0	1	1 (1)	1 (1)	1 (0)	1	1	1
CCA 13	**0**	**1**	**1**	2	2	3	4	5	6	8
CCA 14	0	0	1	1	1	2	3	3	3 (5)	3 (6)
CCA 15	**0**	**0**	**0**	1	1	1	1	2	2	2 (2)
CCA 16	**0**	**0**	**0**	**0**	1	1	1(2)	1(4)	1(6)	1
CCA 17	0	1	1	1	2	2	2 (3)	2 (4)	2 (5)	2
CCA 18	(-3)	(-9)								
CCA 19	0	0	-1	-1	-1	-2	-2	-3	-4	-5
CCA 20	0	1	1	2	3	4	5	5 (7)	5 (9)	5 (10)
CCA 21	0	-1	-1	-2	-3	-5	-6	-6 (-9)	-6 (-13)	-6 (-20)
CCA 22	-1	-2	-3	-5	-5 (-8)	-5 (-12)	-5 (-18)	-5	-5	-5
CCA 23	0	0	1	1	1	1	2	2	3	4
CCA 24	0	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-2)
CCA 25	**0**	**0**	**1**	1	2	2	3	4	5	7
CCA 26	**0**	0	1	1	2	2	3	4	5	5 (6)
CCA 27	0	0	0	1	1	1	2	2 (3)	2 (4)	2 (7)
CCA 28	**0**	**-1**	**-1**	**-1**	-2	-2	-2 (-3)	-2 (35)	-2 (-7)	-2
CCA 29	0	0	0	0	0	0	0	0	0*	0*
CCA 30	**0**	**-1**	**-1**	**-2**	**-3**	**-4**	**-4 (-6)**	**-4 (-9)**	**-4 (-15)**	-4
Col Avg	0	0	0	0	1	1	1	1	1	2
Col Max	1	2	3	4	5	7	7	7	8	8
Col 90th Percentile	0	1	1	2	3	4	5	5	5	7
Num CCAs	29	29	29	29	29	29	29	29	29	29

Once again the mean bias estimates are low regardless of the number of clusters dropped. The percentage change in bias is of little concern if reducing the sample sample by 5 clusters or less, as the mean bias does not exceed 1% while the 90th

percentile bias is still less than 3%. The results of the fluoridated age 12 estimates are shown in Table 4.5 and Table 4.6.

Table 4.5: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	4	9	14	19	26	33	40	40 (48)	40 (56)	40(58)
CCA 2	4	8	12	16	21	25	25 (28)	25 (27)	25(12)	25
CCA 3	4	9	15	21	28	36	36 (46)	36 (57)	36 (74)	36
CCA 4	5	11	17	24	32	41	52	66	66 (82)	66 (102)
CCA 5	4	10	15	22	31	41	53	53 (70)	53 (94)	53 (140)
CCA 6	5	10	15	22	29	38	48	60	75	75 (93)
CCA 7	**4**	**9**	15	22	29	39	50	64	64(80)	64 (98)
CCA 8	4	8	12	17	24	31	40	40 (51)	40 (64)	40 (73)
CCA 9	2	5	8	11	14	17	21	25	29	34
CCA 10	7	16	27	41	41 (61)	41 (89)	41 (128)	41	41	41
CCA 11	2	5	7	10	12	15	18	22	22 (25)	22 (28)
CCA 12	6	14	23	35	50	50 (69)	50 (93)	50 (117)	50	50
CCA 13	5	11	18	26	35	46	59	59 (75)	59 (91)	59 (94)
CCA 14	**4**	**8**	12	17	23	30	38	47	58	58 (71)
CCA 15	**4**	**8**	**13**	19	26	34	44	44 (56)	44 (71)	44 (86)
CCA 16	4	8	13	18	23	29	29 (36)	29 (42)	29 (41)	29
CCA 17	**5**	11	17	25	35	46	61	61 (80)	61(105)	61 (135)
CCA 18	**(11)**	**(26)**	**(46)**							
CCA 19	3	5	8	11	15	18	22	27	32	38
CCA 20	3	5	9	12	16	21	26	32	39	47
CCA 21	4	8	12	17	23	28	34	34 (39)	34 (39)	34 (28)
CCA 22	**6**	**13**	**24**	**40**	**40 (65)**	40 (109)	40 (173)	40	40	40
CCA 23	3	6	9	12	16	20	25	30	36	44
CCA 24	**3**	6	10	13	18	22	27	33	39	46
CCA 25	3	5	9	12	16	20	24	30	36	43
CCA 26	**5**	10	16	23	31	40	40(50)	40 (61)	40(61)	40
CCA 27	3	6	9	13	17	17 (21)	17 (27)	17 (41)	17	17
CCA 28	5	11	18	25	34	34 (45)	34(56)	34 (55)	34	34
CCA 29	**3**	**6**	**9**	**12**	**16**	**20**	25	30	36	43
CCA 30	**3**	**8**	**13**	**19**	**28**	**28 (41)**	**28 (57)**	**28 (69)**	28	28
Col Avg	4	8	14	20	26	31	36	39	42	43
Col Max	7	16	27	41	50	50	61	66	75	75
Col 90th Percentile	5	11	19	27	36	42	52	60	61	61
Num CCAs	29	29	29	29	29	29	29	29	29	29

The percentage change in SE results for the fluoridated 12 year olds are similar to those of the previous two age groupings. Once again the number of clusters in CCA 18 is too small to contribute to the column estimates. Despite reducing the

number of clusters sampled by one or two, CCA's with a DE less than one in the original data still retain a DE less than one. For example looking at CCA 15, the survey SE estimate has increased by 4% reducing the number of clusters sampled by one, 8% reducing the number of clusters sampled by two and 13% reducing the number of clusters sampled by three, yet each of these revised survey SE estimates are still less than the original data SRS SE estimates. Thus, despite sampling fewer clusters the survey variance is still less than the original SRS SE estimate.

Table 4.6: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	0	0	0	1	1	1 (1)	1(1)	1(1)
CCA 2	0	0	-1	-1	-1	-2	-2 (-3)	-2 (-5)	-2 (-8)	-2
CCA 3	0	0	0	0	0	1	1 (1)	1 (1)	1 (2)	1
CCA 4	0	0	0	1	1	1	2	3	3(4)	3(7)
CCA 5	0	0	0	0	0	0	0	0 (0)	0 (0)	0(0)
CCA 6	0	0	0	1	1	1	2	3	3	3 (5)
CCA 7	**0**	**0**	0	0	1	1	1	2	2 (2)	2 (4)
CCA 8	0	0	-1	-1	-1	-2	-2	-2 (-3)	-2 (-5)	-2 (-7)
CCA 9	0	0	0	0	0	-1	-1	-1	-1	-2
CCA 10	1	2	3	5	5 (7)	5 (11)	5 (18)	5	5	5
CCA 11	0	1	1	2	2	3	4	5	5 (7)	5 (9)
CCA 12	0	0	1	1	2	2 (2)	2 (4)	2(6)	2	2
CCA 13	0	0	0	0	1	1	1	1(2)	1 (4)	1 (6)
CCA 14	**0**	**0**	0	-1	-1	-1	-1	-2	-2	-2 (-3)
CCA 15	**0**	**0**	**0**	0	1	1	1	1 (2)	1 (3)	1 (5)
CCA 16	0	1	1	2	3	4	4 (6)	4 (8)	4 (11)	4
CCA 17	**0**	0	0	0	0	-1	-1	-1 (-1)	-1 (-1)	-1 (-1)
CCA 18	**(0)**	**(-1)**	**(-2)**							
CCA 19	0	0	1	1	1	2	2	3	3	4
CCA 20	0	0	-1	-1	-1	-2	-2	-3	-3	-4
CCA 21	0	0	0	0	1	1	1	1 (1)	1 (1)	1 (2)
CCA 22	**0**	**0**	**1**	**1**	**1 (2)**	1 (3)	1 (6)	1	1	1
CCA 23	0	1	1	1	2	2	3	4	5	6
CCA 24	**0**	0	0	0	0	0	-1	-1	-1	-1
CCA 25	0	0	0	-1	-1	-1	-1	-1	-1	-1
CCA 26	**0**	0	1	1	1	2	2 (3)	2(4)	2 (6)	2
CCA 27	0	0	1	1	1	1 (2)	1 (3)	1(5)	1	1
CCA 28	0	0	1	1	2	2 (2)	2 (4)	2 (7)	2	2
CCA 29	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	0
CCA 30	**0**	**0**	**0**	**0**	**0**	**0**(1)	**0**(1)	**0**(1)	0	0
Col Avg	0	0	0	1	1	1	1	1	1	1
Col Max	1	2	3	5	5	5	5	5	5	6
Col 90th Percentile	0	1	1	1	2	3	3	4	4	4
Num CCAs	29	29	29	29	29	29	29	29	29	29

The bias estimates for the fluoridated age 12 grouping were lower than the previous two age groups. The mean bias of 1% when reducing the number of clusters

sampled by 10 should not cause any worry, thus leaving the SE table to decide how much of a reduction in clusters could be implemented. This will depend on the level of increased SE the survey designer would be willing to tolerate in a new design. The results for the 15 year old age grouping are shown in Table 4.7 and Table 4.8.

Table 4.7: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	4	9	14	20	27	34	43	43 (52)	43 (61)	43 (57)
CCA 2	**7**	14	23	35	49	49 (67)	49 (93)	49 (127)	49	49
CCA 3	7	15	25	36	50	50 (66)	50 (86)	50 (108)	50	50
CCA 4	**6**	**6 (14)**	**6 (24)**	6(40)	6	6	6	6	6	6
CCA 5	5	11	18	26	37	49	49 (66)	49(85)	49 (97)	49
CCA 6	6	12	20	29	39	50	50 (63)	50 (74)	50 (73)	50
CCA 7	6	12	18	26	34	34 (42)	34 (49)	34 (46)	34	34
CCA 8	4	8	13	18	25	32	41	52	52 (64)	52 (74)
CCA 9	3	7	11	14	14(18)	14(20)	14(9)	14	14	14
CCA 10	8	17	29	29(45)	29 (67)	29 (99)	29	29	29	29
CCA 11	3	7	12	18	18 (25)	18 (34)	18 (42)	18	18	18
CCA 12	7	16	16 (28)	16 (47)	16 (78)	16	16	16	16	16
CCA 13	9	19	19 (32)	19 (46)	19 (58)	19	19	19	19	19
CCA 14	**5**	12	21	21 (33)	21 (49)	21 (69)	21	21	21	21
CCA 15	6	13	21	31	31 (42)	31 (56)	31 (69)	31	31	31
CCA 16	6	12	18	25	33	41	48	48 (55)	48 (57)	48 (46)
CCA 17	4	10	16	23	33	33 (45)	33 (64)	33 (100)	33	33
CCA 18	6	11	11 (14)	11 (14)	11 (7)	11	11	11	11	11
CCA 19	3	5	7	7 (7)	7 (4)	7 (-4)	7	7	7	7
CCA 20	5	11	18	26	36	36 (48)	36 (63)	36 (79)	36	36
CCA 21	**4**	10	17	17 (30)	17 (50)	17 (77)	17	17	17	17
CCA 22	6	13	13 (19)	13 (22)	13 (13)	13	13	13	13	13
CCA 23	1	1 (0)	1(-6)	1 (-20)	1	1	1	1	1	1
CCA 24	7	15	24	35	47	47 (62)	47 (80)	47 (97)	47	47
CCA 25	**5**	**11**	**11 (18)**	**11 (26)**	**11 (26)**	11	11	11	11	11
CCA 26	8	17	27	27 (40)	27(54)	27 (66)	27	27	27	27
CCA 27	10	10 (22)	10 (35)	10 (45)	10	10	10	10	10	10
CCA 28	**(28)**									
CCA 29	3	7	12	19	19 (27)	19 (37)	19 (49)	19	19	19
CCA 30	10	10 (22)	10 (34)	10 (44)	10	10	10	10	10	10
Col Avg	6	11	16	20	24	25	26	27	27	27
Col Max	10	19	29	36	50	50	50	52	52	52
Col 90th Percentile	8	16	24	31	40	49	49	49	49	49
Num CCAs	29	29	29	29	29	29	29	29	29	29

The results of the fluoridated age 15 year olds are presented for consistency with Chapter 3. However, the estimates may not be as reliable as the other age group-

ings due to the reduced number of clusters sampled within each CCA. Within a CCA, there are usually less secondary schools than primary schools, however the schools are usually larger in size. Thus observing Table 4.8, no new information is gained once seven schools have been omitted as the estimates are just continued across the row, and thus, this table could be truncated at this point. The results are left to see the percentage change in SE when $n \leq 5$.

Due to the reduced number of schools, the mean SE estimates are higher at almost all values for the reduction in the number of schools. However, if a survey designer was willing to put up with an increase in SE of 20% at the 90th percentile then two schools less per CCA could be sampled. Even such a small reduction in the number of schools would be a large cost saving for the survey design.

Table 4.8: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	0	0	1	1	1	1 (1)	1 (2)	1 (2)
CCA 2	**0**	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-4)	-1	-1
CCA 3	0	0	0	0	1	1 (1)	1 (1)	1 (2)	1	1
CCA 4	**0**	**0 (0)**	**0 (1)**	0(1)	0	0	0 (1)	0 (1)	0 (2)	0
CCA 5	0	0	0	0	0	0	0 (0)	0 (0)	0 (1)	0
CCA 6	0	0	0	1	1	1	1 (2)	1 (3)	1 (4)	1
CCA 7	0	0	0	-1	-1	-1 (-2)	-1 (-3)	-1 (-5)	-1	-1
CCA 8	0	0	0	0	0	0	1	1	1 (1)	1 (2)
CCA 9	0	0	0	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1
CCA 10	1	1	2	2 (4)	2 (8)	2 (17)	2	2	2	2
CCA 11	0	0	0	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1
CCA 12	0	0	0 (0)	0 (1)	0 (2)	0	0	0	0	0
CCA 13	0	1	1 (1)	1 (2)	1 (5)	1	1	1	1	1
CCA 14	**0**	0	0	0 (-1)	0 (-1)	0 (-2)	0	0	0	0
CCA 15	0	0	0	0	0 (0)	0 (0)	0(1)	0	0	0
CCA 16	0	0	0	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-2)
CCA 17	0	0	0	0	0	0 (0)	0 (0)	0 (1)	0	0
CCA 18	1	1	1 (2)	1 (3)	1 (3)	1	1	1	1	1
CCA 19	0	0	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1	-1
CCA 20	0	0	0	0	0	0 (0)	0 (1)	0 (1)	0	0
CCA 21	**0**	0	0	0 (-1)	0 (-1)	0 (-3)	0	0	0	0
CCA 22	0	0	0 (0)	0 (0)	0 (-1)	0	0	0	0	0
CCA 23	-1	-1 (-2)	-1 (-4)	-1 (-7)	-1	-1	-1	-1	-1	-1
CCA 24	0	0	0	0	1	1 (1)	1 (1)	1 (2)	1	1
CCA 25	**0**	**0**	**0 (0)**	**0 (0)**	**0 (0)**	0	0	0	0	0
CCA 26	0	0	0	0 (0)	0 (0)	0 (1)	0	0	0	0
CCA 27	0	0 (0)	0 (1)	0 (2)	0	0	0	0	0	0
CCA 28	**(-1)**									
CCA 29	0	0	0	0	0 (0)	0 (-1)	0 (-1)	0	0	0
CCA 30	0	0 (0)	0 (-1)	0 (-2)	0	0	0	0	0	0
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	1	1	2	2	2	2	2	2	2	2
Col 90th Percentile	0	0	0	1	1	1	1	1	1	1
Num CCAs	29	29	29	29	29	29	29	29	29	29

The bias estimates for the fluoridated age 15 results are very small with each of the column averages showing a bias of 0%. This needs to be considered with the change in SE for the survey designer to make a decision. The results of the non-fluoridated observations are presented in the next eight tables, beginning with the non-fluoridated age 5 observations in Table 4.9 and Table 4.10.

4.4.2 Tables of Percentage Change in SE and Bias by Reduction in Number of Non-Fluoridated Schools Selected

Table 4.9: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	2	3	5	7	9	10	12	13	14	15
CCA 10	3	7	11	15	20	25	31	38	45	45 (52)
CCA 11	**4**	**9**	**15**	**21**	**28**	35	45	55	55 (67)	55 (81)
CCA 12	3	6	9	12	16	20	25	30	36	42
CCA 13	3	7	10	14	18	22	26	30	33	36
CCA 14	0	1	1	2*	2*	3*	4*	4*	5*	5*
CCA 15	1	1	2	2	3*	4*	4*	5*	6*	6*
CCA 16	1	3	4	6*	8*	9*	11*	12*	14*	16*
CCA 17	**2**	**4**	**7**	**9**	**12**	15	18	21	24	24 (28)
CCA 18	2	5	8	11	14	17	21	25*	29*	33*
CCA 19	**3**	**6**	**10**	**14**	**18**	**24**	**30**	**38**	**47**	59
CCA 20	-1	-2	-2	-2	-1	1	4	9	9 (16)	9 (26)
CCA 21	1	3	4	6	7	9	10	12	14	16
CCA 22	3	5	8	11	14	18	21	25	30	34
CCA 23	3	6	10	13	17	21	26	30	35	35 (39)
CCA 24	3	5	8	11	15	18	22	27	32	37
CCA 25	**4**	**9**	**14**	21	28	37	47	61	61 (79)	61 (103)
CCA 26	4	4 (9)	4 (12)	4 (10)	4	4	4	4	4	4
CCA 27	2	5	7	10	13	16	19	23*	27*	31*
CCA 28	4	8	13	19	25	33	42	53	67	85
CCA 29	5	10	16	23	32	42	42 (55)	42 (72)	42 (90)	42
CCA 30	1	1	1	1	1	0	-2	-4	-4 (-9)	-4 (-17)
Col Avg	2	5	8	10	14	17	21	25	28	31
Col Max	5	10	16	23	32	42	47	61	67	85
Col 90th Percentile	4	9	14	20	27	35	42	52	54	59
Num CCAs	22	22	22	22	22	22	22	22	22	22

As in Chapter 3, the non-fluoridated results show no estimates for CCA 1 to CCA 8, as the populations are entirely fluoridated. Throughout the non-fluoridated results more CCA's require estimates to be continued across the table as there are not enough clusters to drop. This is partly due to the fact that non-fluoridated schools are less common in certain CCA's due to a predominantly fluoridated

water supply. An example of this is CCA 26, the South Lee region, which would only feature non-fluoridated children commuting to schools in the city. However, dropping these observations could lead to a bias within the data, and thus these observations should be retained to capture the full diversity of the CCA.

The number of CCA's with DE's less than one are reduced compared to the fluoridated aged 5 grouping. Estimates of the percentage increase in SE are lower at all values for the reduction in the number of schools, again, compared to the fluoridated age 5 results.

Table 4.10: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	0	0	-1	-1	-1	-2	-2	-3	-3	-4
CCA 10	0	1	1	2	3	4	5	6	7	7 (9)
CCA 11	**0**	**1**	**2**	**3**	**4**	5	6	9	9 (12)	9 (16)
CCA 12	0	0	0	0	0	0	1	1	1	1
CCA 13	0	1	1	2	3	4	4	5	7	8
CCA 14	0	0	0	0*	0*	0*	0*	0*	0*	0*
CCA 15	0	0	0	0	0*	*0	0*	0*	0*	0*
CCA 16	0	0	0	0*	0*	0*	0*	0*	0*	0*
CCA 17	**1**	**2**	**3**	**5**	**6**	8	10	13	17	17 (21)
CCA 18	0	0	0	0	0	0	0	1*	1*	1*
CCA 19	**0**	**0**	**0**	**0**	**-1**	**-1**	**-1**	**-1**	**-2**	-2
CCA 20	0	0	0	0	0	0	-1	-1	-1 (-1)	-1 (-1)
CCA 21	-1	-1	-2	-2	-3	-4	-5	-6	-8	-9
CCA 22	0	1	1	1	2	2	3	3	4	5
CCA 23	0	0	0	0	0	0	0	0	-1	-1 (-1)
CCA 24	0	0	0	0	0	0	0	0	0	0
CCA 25	**0**	**0**	**0**	0	0	0	-1	-1	-1 (-1)	-1 (-1)
CCA 26	0	0 (-1)	0 (-2)	0 (-4)	0	0	0	0	0	0
CCA 27	0	0	0	0	0	0	0	0*	0*	0*
CCA 28	0	0	0	0	0	0	1	1	1	2
CCA 29	0	1	1	1	2	3	3 (4)	3 (5)	3 (7)	3
CCA 30	0	-1	-1	-1	-2	-3	-5	-7	-7 (-10)	-7 (-14)
Col Avg	0	0	0	0	1	1	1	1	1	1
Col Max	1	2	3	5	6	8	10	13	17	17
Col 90th Percentile	0	1	1	2	3	4	4	6	7	8
Num CCAs	22	22	22	22	22	22	22	22	22	22

The bias estimates for the non-fluoridated age 5 grouping are similar to those of the fluoridated aged 5 results. The risk of a large bias estimate is increased compared to the fluoridated results, noting that one of the CCA's maximum average bias can be as much as 17%, but only when dropping a large number of schools (9 or 10).

Table 4.11: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	3	6	9	12	16	20	24	28	33	38
CCA 10	**5**	**11**	**17**	**25**	**33**	**43**	43 (55)	43 (69)	43 (83)	43
CCA 11	**6**	**13**	**22**	**22 (34)**	**22 (50)**	**22 (73)**	22	22	22	22
CCA 12	3	6	9	13	17	21	25	30	35	35 (39)
CCA 13	6	12	20	29	29 (40)	29 (51)	29 (49)	29	29	29
CCA 14	3	6	9	13	17	21	25	30	36	42
CCA 15	**4**	**8**	**13**	19	26	35	45	45 (57)	45 (71)	45 (81)
CCA 16	**4**	**8**	**12**	**17**	**21**	25	29	29 (32)	29 (30)	29 **(18)**
CCA 17	3	6	10	13	17	22	27	32	37	37 (42)
CCA 18	**3**	**5**	**8**	**11**	**15**	**19**	**23**	27	32	38
CCA 19	-1	-3	-5	-7	-9	-12	-16	-21	-26	-33
CCA 20	6	12	19	26	26 (34)	26 (40)	26 (24)	26	26	26
CCA 21	7	14	21	21 (30)	21 (39)	21 (41)	21	21	21	21
CCA 22	**3**	**5**	**8**	**11**	**13**	**16**	**18**	**18 (19)**	**18 (19)**	**18 (17)**
CCA 23	2	5	7	10	13	16	19	23	26	30
CCA 24	3	7	11	16	22	28	35	44	44 (53)	44 (63)
CCA 25	**4**	**8**	**13**	**18**	25	33	43	56	56 (73)	56 (100)
CCA 26	**8**	**19**	**32**	**32 (49)**	**32 (74)**	**32 (122)**	32	32	32	32
CCA 27	**3**	**6**	**9**	12	16	20	24	29	34*	40*
CCA 28	7	7 (17)	7 (31)	7 (53)	7	7	7	7	7	7
CCA 29	(1)	(2)	(-13)							
CCA 30	4	4 (5)	4 (2)	4 (-8)	4	4	4	4	4	4
Col Avg	4	8	12	16	18	21	24	26	28	29
Col Max	8	19	32	32	33	43	45	56	56	56
Col 90th Percentile	7	13	21	26	29	33	43	44	44	44
Num CCAs	21	21	21	21	21	21	21	21	21	21

CCA 29 has no percentage change in SE estimates for the non-fluoridated age 8 year olds due to an insufficient number of clusters to produce reliable SE estimates. Comparing the non-fluoridated results to the fluoridated equivalent, one notices more of the values have been imputed across the CCA rows. This is due to non-fluoridated schools being less common in particular CCAs.

The SE estimates are initially similar to those of the fluoridated results, however

the SE estimates remain less variable in the non-fluoridated results. Again DE's less than one are present, and the SE estimates remain less than the original sample SRS SE for a substantial percentage increase in SE in most such CCA's.

Table 4.12: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	0	0	1	1	1	1	2	2	2	3
CCA 10	**0**	**1**	**1**	**2**	**3**	**4**	4 (6)	4 (9)	4 (16)	4
CCA 11	**1**	**3**	**6**	**6 (11)**	**6 (19)**	**6 (38)**	6	6	6	6
CCA 12	0	0	0	0	0	0	0	-1	-1	-1 (-1)
CCA 13	0	0	0	0	0 (1)	0 (2)	0 (7)	0	0	0
CCA 14	0	0	1	1	1	1	2	2	3	3
CCA 15	**0**	**0**	**0**	0	0	-1	-1	-1 (-1)	-1 (-1)	-1 (0)
CCA 16	**0**	**0**	**0**	**0**	**0**	-1	-1	-1 (-1)	-1 (-3)	-1 **(-6)**
CCA 17	1	1	2	3	4	5	6	8	9	9 (11)
CCA 18	**0**	**0**	**0**	**0**	**0**	**0**	**0**	-1	-1	-1
CCA 19	-1	-2	-4	-5	-7	-10	-12	-15	-19	-23
CCA 20	0	1	2	3	3 (5)	3 (8)	3 (14)	3	3	3
CCA 21	1	2	4	4 (6)	4 (10)	4 (17)	4	4	4	4
CCA 22	**1**	**1**	**2**	**3**	**4**	**6**	**8**	**8 (10)**	**8 (12)**	**8 (16)**
CCA 23	0	0	0	0	1	1	1	1	1	2
CCA 24	0	0	0	0	1	1	1	1	1(2)	1 (2)
CCA 25	**0**	**-1**	**-1**	**-1**	-2	-2	-2	-3	-3 (-2)	-3 (-1)
CCA 26	**1**	**3**	**6**	**6 (9)**	**6 (16)**	**6 (32)**	6	6	6	6
CCA 27	**0**	**0**	**0**	0	0	0	0	1	1*	1*
CCA 28	2	2 (6)	2 (12)	2 (24)	2	2	2	2	2	2
CCA 29	(1)	(3)	(8)							
CCA 30	-2	-2 (-4)	-2 (-8)	-2 (-14)	-2	-2	-2	-2	-2	-2
Col Avg	0	1	1	1	1	1	1	1	1	1
Col Max	2	3	6	6	6	6	8	8	9	9
Col 90th Percentile	1	2	4	4	4	6	6	6	6	6
Num CCAs	21	21	21	21	21	21	21	21	21	21

Once more the bias estimates are similar to the fluoridated age 8 results. However, like the non-fluoridated age 5 results, the bias estimates are more variable as the reduction in number of schools increases, relative to the fluoridated estimates in the same age grouping.

Table 4.13: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	4	8	13	19	25	33	41	41 (52)	41 (65)	41 (79)
CCA 10	3	6	9	13	17	17 (22)	17 (29)	17 (38)	17	17
CCA 11	3	7	11	15	20	25	31	37	37 (43)	37 (48)
CCA 12	4	8	12	17	23	29	36	45	55	67
CCA 13	**5**	**5 (11)**	**5 (17)**	**5 (23)**	5	5	5	5	5	5
CCA 14	3	5	8	12	15	19	23	28	33	38
CCA 15	**3**	**7**	**12**	**16**	**22**	**28**	**35**	**44**	55	55 (68)
CCA 16	4	8	12	16	21	26	31	31 (35)	31 (39)	31 (29)
CCA 17	**3**	**6**	**9**	**12**	**16**	**20**	**24**	**29**	34	40
CCA 18	3	7	11	15	19	24	30	36	44	52
CCA 19	3	7	11	15	21	27	27 (37)	27 (46)	27 (41)	27
CCA 20	7	16	26	26	26	26	26	26	26	26
CCA 21	10	22	22 (36)	22 (45)	22 (28)	22	22	22	22	22
CCA 22	8	17	17 (27)	17 (34)	17 (27)	17	17	17	17	17
CCA 23	3	7	11	16	21	26	33	39	47	47 (54)
CCA 24	**3**	**7**	**10**	**15**	**19**	25	31	38	46	55
CCA 25	**3**	**7**	**11**	15	20	25	32	39	47	57
CCA 26	10	10 (22)	10 (35)	10 (34)	10	10	10	10	10	10
CCA 27	2	4	7	9	12	15	18	21*	25*	28*
CCA 28	**(27)**	(72)								
CCA 29	(-7)									
CCA 30	10	10 (23)	10 (43)	10 (80)	10	10	10	10	10	10
Col Avg	5	9	12	15	18	21	25	28	31	34
Col Max	10	22	26	26	26	33	41	45	55	67
Col 90th Percentile	10	16	18	19	23	28	35	42	48	56
Num CCAs	20	20	20	20	20	20	20	20	20	20

Comparing the non-fluoridated age 12 year olds with the fluoridated, the non-fluoridated results have slightly higher increase in mean SE for the initial reduction in the number of schools selected. However, as the number of schools selected is reduced, the non-fluoridated results then show a smaller SE compared with the fluoridated results. This may be influenced by the imputing of estimates for the higher values of reduction in number of schools selected, and in general, with standard CCA sizes of 15 to 20, dropping 7 to 10 schools may be too extreme. However, in situations where DE's less than one are initially present, the SE estimate still remains smaller than the original SRS SE estimate for a reduction of one or two schools.

Table 4.14: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	0	0	0	0	0	0	1	1 (1)	1 (1)	1 (3)
CCA 10	1	2	3	4	5	5 (7)	5 (9)	5 (11)	5	5
CCA 11	0	0	0	0	1	1	1	1	1 (2)	1 (2)
CCA 12	0	0	0	0	0	0	1	1	1	1
CCA 13	**-1**	**-1 (-2)**	**-1 (-4)**	**-1 (-7)**	-1	-1	-1	-1	-1	-1
CCA 14	0	0	0	0	0	0	0	0	0	0
CCA 15	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	1	1 (2)
CCA 16	0	0	-1	-1	-1	-2	-3	-3 (-4)	-3 (-5)	-3 (-8)
CCA 17	**0**	**0**	**0**	**0**	**0**	**-1**	**-1**	**-1**	-1	-1
CCA 18	0	0	1	1	1	1	2	2	3	4
CCA 19	0	0	-1	-1	-1	-2	-2 (-3)	-2 (-4)	-2 (-3)	-2
CCA 20	0	0	0	0 (0)	0 (-1)	0 (-2)	0	0	0	0
CCA 21	0	-1	-1 (-2)	-1 (-4)	-1 (-10)	-1	-1	-1	-1	-1
CCA 22	0	-1	-1 (-1)	-1 (-2)	-1 (-5)	-1	-1	-1	-1	-1
CCA 23	0	0	0	0	-1	-1	-1	-1	-2	-2 (-2)
CCA 24	**0**	**0**	**0**	**0**	**0**	0	0	0	1	1
CCA 25	**0**	**0**	**1**	1	1	1	2	2	3	3
CCA 26	0	0 (0)	0 (0)	0 (0)	0	0	0	0	0	0
CCA 27	0	0	0	0	0	0	0	0*	0*	0*
CCA 28	**(1)**	(6)								
CCA 29	(-1)									
CCA 30	-1	-1 (-1)	-1 (-2)	-1 (-2)	-1	-1	-1	-1	-1	-1
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	1	2	3	4	5	5	5	5	5	5
Col 90th Percentile	0	0	1	1	1	1	2	2	3	3
Num CCAs	20	20	20	20	20	20	20	20	20	20

The bias estimates for the 12 year old non-fluoridated results all have a mean of 0%. This suggests the number of clusters to drop should be made on the basis of the change in SE, however again one must note there are quite a few imputed values.

Table 4.15: Percentage Change in Mean SE of Estimate of Mean dmft by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	(4)	(7)	(6)							
CCA 10	9	9 (18)	9 (25)	9 (27)	9	9	9	9	9	9
CCA 11	9	20	20 (32)	20 (44)	20 (47)	20	20	20	20	20
CCA 12	6	14	24	24 (39)	24 (62)	24 (85)	24	24	24	24
CCA 13	**9**	20	32	32 (48)	32 (69)	32 (95)	32	32	32	32
CCA 14	6	13	21	21 (32)	21 (43)	21 (50)	21	21	21	21
CCA 15	**12**	25	42	42 (61)	42 (86)	42 (120)	42	42	42	42
CCA 16	3	7	12	17	24	24 (32)	24 (42)	24 (53)	24	24
CCA 17	4	8	12	16	16 (20)	16 (24)	16 (28)	16	16	16
CCA 18	5	10	17	23	31	31 (38)	31 (45)	31 (43)	31	31
CCA 19	**15**	**15 (37)**	**15 (71)**	15 (122)	15	15	15	15	15	15
CCA 20	(10)	(20)	(24)							
CCA 21										
CCA 22	10	10 (24)	10 (41)	10 (63)	10	10	10	10	10	10
CCA 23	7	15	23	23 (31)	23 (37)	23 (31)	23	23	23	23
CCA 24	8	16	26	37	49	64	64 (82)	64 (104)	64 (129)	64
CCA 25	(1)	(-16)								
CCA 26	1	1 (1)	1 (-2)	1 (-13)	1	1	1	1	1	1
CCA 27	5	5 (13)	5 (23)	5 (29)	5	5	5	5	5	5
CCA 28	(-6)									
CCA 29	**(31)**									
CCA 30	(-9)									
Col Avg	7	13	18	20	22	23	23	23	23	23
Col Max	15	25	42	42	49	64	64	64	64	64
Col 90th Percentile	11	20	30	35	38	38	38	38	38	38
Num CCAs	15	15	15	15	15	15	15	15	15	15

The SE results for the 15 year old non-fluoridated observations are only present for 15 CCA's. There is an increased number of CCA's that did not have enough schools to facilitate the reduction in number of schools and still have 5 schools remaining, thus the table contains blank rows. Furthermore, as with the fluoridated 15 year olds, there are a lot of CCA's that require imputation of the results across each row. Thus, as with the fluoridated age 15 observations, reducing the number of schools sampled should be exercised with caution.

The initial percentage increase in SE by reducing the number of schools sampled are larger for the age 15 observations compared with the younger age groups. Also, the CCAs with a DE less than one only allow one to two schools less to be sampled before the survey SE becomes larger than the original SRS SE.

Table 4.16: Percentage Bias of Mean dmft ($\bar{y}_r$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	(1)	(4)	(8)							
CCA 10	-1	-1 (-2)	-1 (-3)	-1 (-8)	-1	-1	-1	-1	-1	-1
CCA 11	0	0	0 (0)	0 (-1)	0 (-1)	0	0	0	0	0
CCA 12	0	1	1	1 (2)	1 (3)	1 (7)	1	1	1	1
CCA 13	**0**	0	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1	-1
CCA 14	0	0	0	0 (-1)	0 (-1)	0 (-1)	0	0	0	0
CCA 15	**0**	0	0	0 (-1)	0 (-1)	0 (0)	0	0	0	0
CCA 16	0	0	0	0	1	1 (1)	1 (2)	1 (3)	1	1
CCA 17	0	-1	-2	-2	-2 (-3)	-2 (-4)	-2 (-5)	-2	-2	-2
CCA 18	0	1	1	2	3	3 (4)	3 (5)	3 (7)	3	3
CCA 19	**0**	**0 (0)**	**0 (-1)**	0 (-1)	0	0	0	0	0	0
CCA 20	(1)	(2)	(3)							
CCA 21										
CCA 22	1	1 (3)	1 (6)	1 (14)	1	1	1	1	1	1
CCA 23	0	0	0	0 (-1)	0 (-2)	0 (-4)	0	0	0	0
CCA 24	0	0	0	0	1	1	1 (1)	1 (2)	1 (3)	1
CCA 25	(2)	(7)								
CCA 26	-2	-2 (-4)	-2 (-8)	-2 (-14)	-2	-2	-2	-2	-2	-2
CCA 27	0	0 (0)	0 (-1)	0 (-2)	0	0	0	0	0	0
CCA 28	(0)									
CCA 29	**(5)**									
CCA 30	(4)									
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	1	1	1	2	3	3	3	3	3	3
Col 90th Percentile	0	1	1	1	1	1	1	1	1	1
Num CCAs	15	15	15	15	15	15	15	15	15	15

The bias estimates for the non-fluoridated age 15 results are identical to the fluoridated age 15 results and thus, the choice in number of clusters less sampled should be based on the SE. Again, as mentioned in the fluoridated age 15 results, the smaller numbers of clusters within the CCA's cause the user to be cautious with this approach and the estimates are presented more for consistency.

4.4.3 Summary Figures of Percentage Change in SE and Bias by Reduction in Number of Schools Selected

The results in the preceding tables are summarized in the following figures. The figures can be used as a guide for conducting a similar survey, and allow the designer to examine the expected average, and 90th percentile, increase in SE at a particular age group. Using the tables, the designer could choose a particular increase in SE and bias they would tolerate in their survey, and this figure allows them to estimate the number of clusters they could reduce from the sample.

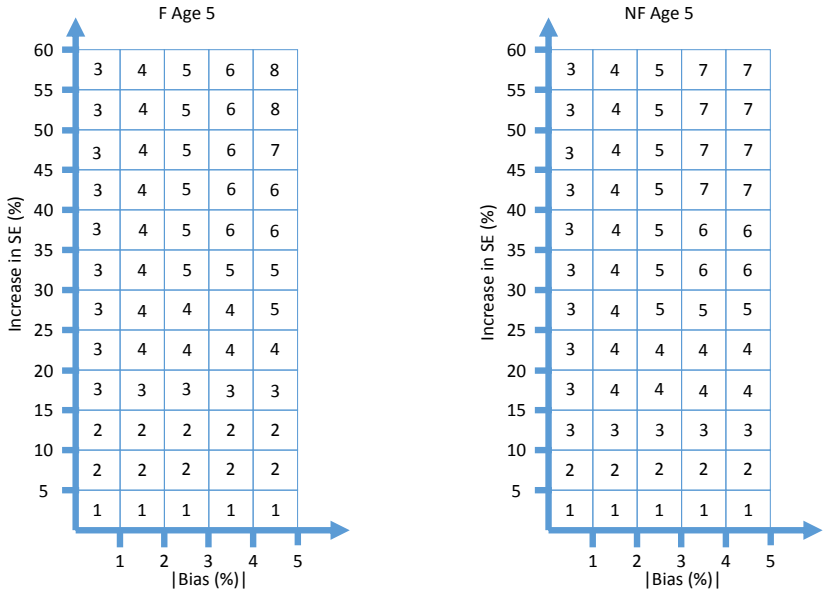


Figure 4.1: Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 5.

Figure 4.1 shows the potential reduction in the number of clusters for a given change in the 90^{th} percentile SE and bias estimates. The 90^{th} percentile is chosen to ensure the increase in SE should not, in general, be more than the specified percentage for a given reduction in the number of clusters selected. Thus, if the survey designer is willing to tolerate at most a 5% increase in SE, then they can reduce the number of clusters selected per CCA by one. However, if they are willing to tolerate a 10% increase in SE then they could sample two schools less per CCA. Both of these estimates hold with less than a 1% change in bias. If however the survey designer is willing to tolerate a 25% increase in SE, the level of bias now becomes important. Thus if the designer is willing to tolerate, at most, a 25% increase in SE and a 2% increase in bias, they can select four schools less per CCA. However if the designer wants to keep bias less than 1% they can

only allow the reduction in schools sampled to be three.

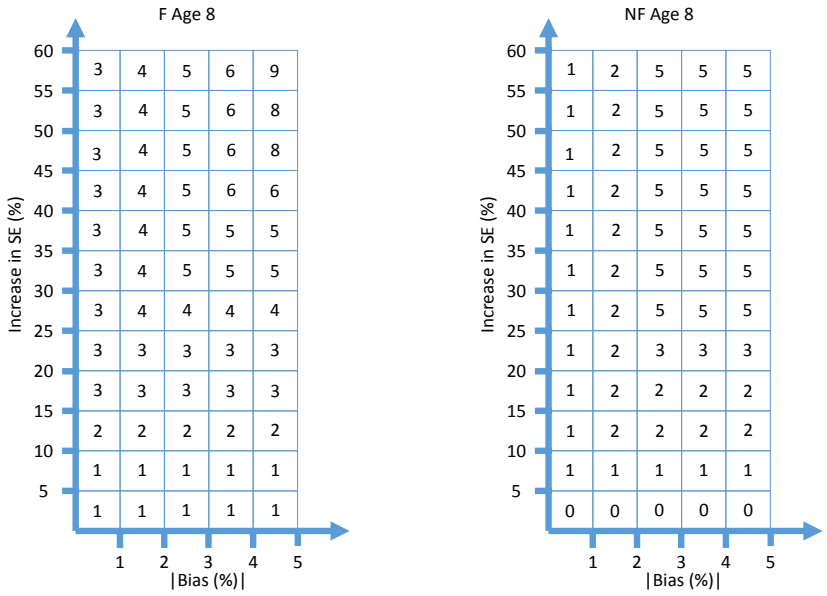


Figure 4.2: Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 8.

The results for the fluoridated age 8 group are similar to those of the fluoridated age 5 group. The results of the non-fluoridated age 8 results show that the reduction in clusters sampled must be less than those of the non-fluoridated 5 year olds at the same levels of bias and SE. This difference can be attributed to the bias associated with the respective estimates.

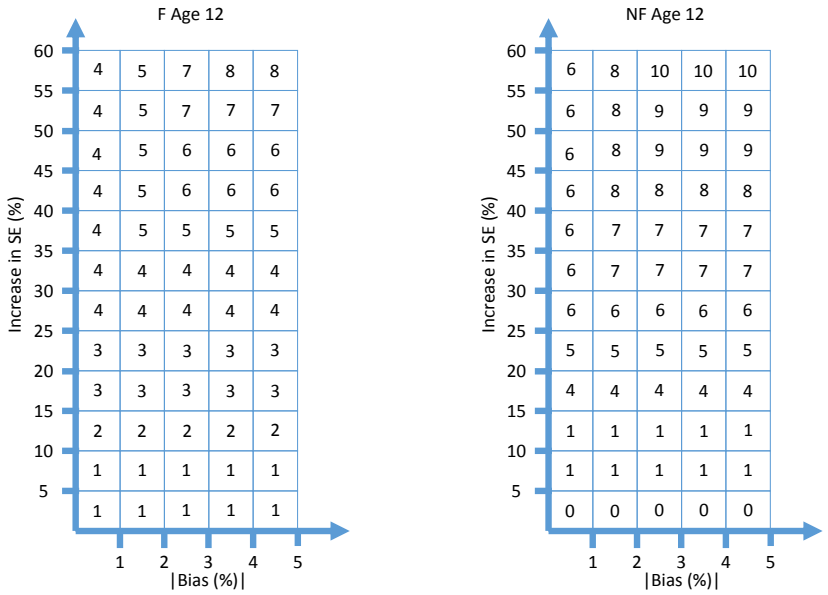


Figure 4.3: Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 12.

The results of the fluoridated 12 year olds are similar to those of the 5 and 8 year olds. The non-fluoridated results show a dramatic increase in the number of clusters that can be reduced compared to the results of the 8 year olds. This is mainly due to the difference in bias between the two sets of estimates.

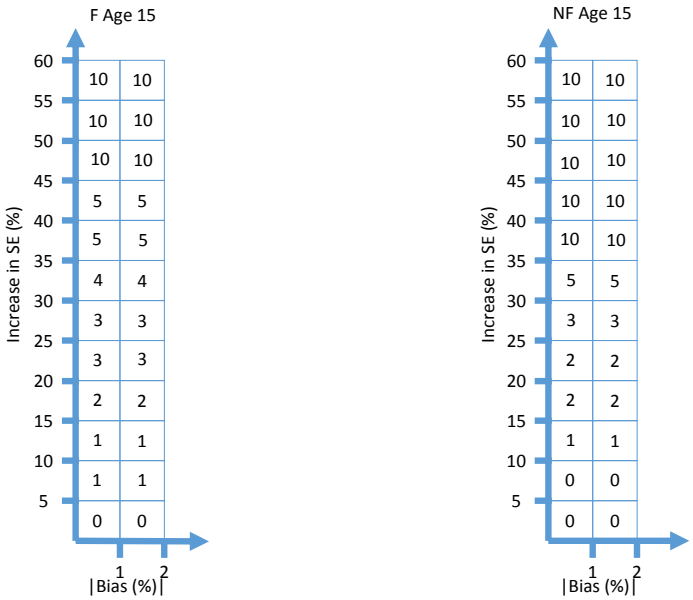


Figure 4.4: Potential Reduction in the Number of Schools for a given Bias and Increase in SE (90^{th} Percentile) for Children aged 15.

The results of the 15 year old are only presented for bias of 1% and 2% as the

estimates are all less than 2% bias. However, as aforementioned, there are less estimates forming these values and they should be taken with caution.

It should be noted that the above estimates are all at the 90th percentile of the CCA estimates, and as a result these are very conservative estimates of the reduction in the number of schools. Of course each of the estimates is compiled from the mean of a large number of replications, so there is some variability associated with the estimates as the survey designer will only be able to draw one replication, not all. However the underlying data is one replication of a sampling design. As the 90th percentile estimates are conservative, Figure 4.5 presents estimates for each value of the reduction in the number of clusters sampled, based on the mean of the CCA estimates, not the 90 percentile, providing more liberal estimates.

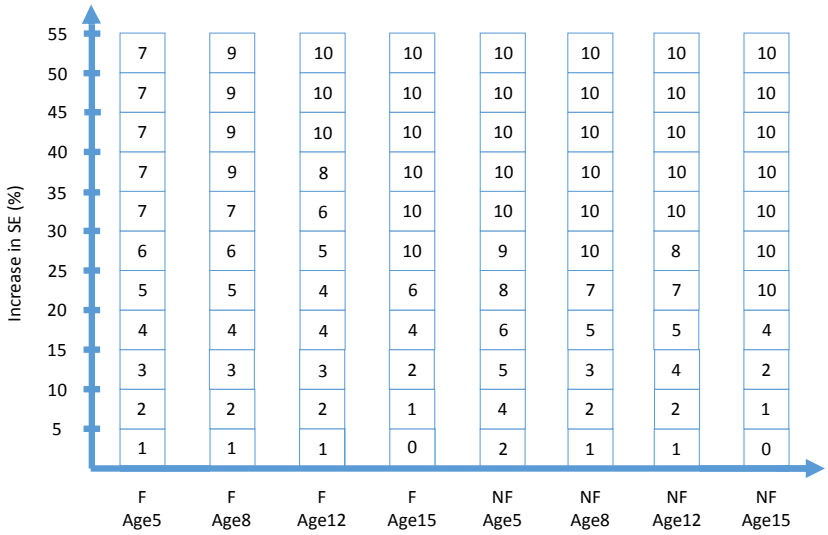


Figure 4.5: Change in Mean SE by Reduction in Number of Clusters Sampled at 1% Bias.

Looking at the mean estimates, the reduction in the number of clusters selected may be a little extreme, and thus the author advises that the 90th percentile estimates may be a better guide. Furthermore, while the results are presented up to a SE increase of 55% to 60% in the previous figures, in reality such a large increase in SE would be too extreme and probably not be advisable, the results are just shown in the exploratory sense. As many of the CCA's in cluster surveys may not have a large number of schools sampled, dropping 10 clusters per CCA may not be feasible and in such cases the results are presented for consistency and as an exploratory guide.

4.5 Concluding Remarks

Chapter 4 looks at the effect of reducing the number of clusters sampled on the SE and DE of each CCA, relative to the original SE estimates in the survey data in Chapter 3. Using a systematic resampling approach, the effect of reducing the number of clusters sampled by $n = 1, 2, \ldots, 10$ was explored. As the underlying cause of CCA's with DE's less than one could not be attributed to one particular reason, the estimates of percentage change in SE (Figure 4.1 to Figure 4.5) are compiled across all CCA's, not just those with DE's less than one. The results showed that a small reduction in the number of clusters sampled had minimal effect on the SE estimates (i.e. for $n = 1, 2$) and had no effect on the bias of the estimate. Given the scale of many cluster surveys at national level, this could result in a large cost reduction in survey planning, provided the survey planner can tolerate the small increase in SE.

For example, most oral health surveys involve a clinical examination usually carried out in the school. This requires both a dentist and a recorder logging the examination results to attend the school. Further costs include the cost of travel to the school, the dental consumables used conducting examinations along with bringing the required furniture (chair, lighting, etc.) to carry out the examination. Moreover, reducing the number of schools visited reduces the imposition on schools, and this in turn enhances the good will of the schools for future studies. Given each of the aforementioned steps apply to each school, a small reduction in the number of schools visited can have large cost and time saving implications for surveys.

Chapter 5

Discussion, Future Research and Conclusion

5.1 Chapter Two

Throughout the literature review (Section 1.3.3) there are instances of comparisons of the standard variance estimation methods for cluster samples carried out, including comparisons in the presence of unusual characteristics in the underlying data (skewed distributions, small sample sizes, large sampling fraction). However, most of the literature reviewed makes no attempt to define small or large, and furthermore none of the literature reviewed looks at these conditions occurring simultaneously. The simulation results in Chapter 2 compared the two typical variance estimators (Taylor series linearization and the delete-one jackknife) on the simulated population, showing that the estimates obtained using both the TSL and JK1 methods are almost identical, when the sampling fraction is small and the number of second-stage elements sampled is large. Thus, as mentioned by Rust (1985), deciding between them should be made on the basis of the available software or familiarity of the analyst with one method over the other. However, care must be taken when the sampling fraction is large or the number of second-stage elements is small. Under a normal distribution, divergence between the two methods occurred when the sampling fraction became larger than 75% when there was 20 second-stage elements sampled ($m_i = 20$), or when the sampling fraction becomes larger than 30% when the number of second-stage elements sampled was $m_i = 5$. Changing to a Poisson($\lambda_i \sim U(3.5, 6.5)$) distribution, which is both discrete and slightly skewed and keeping the same number of second-stage sampling

elements caused divergence between the estimates to occur for smaller sampling fractions (approx 50% for $m_i = 20$ and 20% for $m_i = 5$). Increasing the level of skewness ($\lambda_i \sim U(0.5, 1.5)$) had little effect on the sampling fraction where divergence began, but the difference in estimates between the two methods became larger. Divergence in estimates under the lognormal distribution, despite being skewed, occurs at similar sampling fractions to the normal distribution.

The simulation results showed TSL to be the more conservative estimate of the two SE methods in almost all cases, except when the number of clusters was small, for a skewed distribution. Thus, if one was unsure about a variable's distribution or sampling fraction, provided the number of clusters sampled is greater than five (Koop, 1972), familiarity with a particular variance estimation method should be avoided, and one should use the gold standard method, TSL. The large sampling fraction affects the JK1 estimate far more than it affects the TSL estimate, possibly due to the JK1's initial development for a with-replacement design. From the literature TSL is seen as the gold standard, and the simulation highlights where divergence occurs under different distributional assumptions, given variables that would be typical of an oral health survey.

Both methods perform poorly on skewed distributions, particularly when the number of second-stage elements is small (JK1), or the number of sampled clusters is small (TSL). Thus, in highly skewed populations, neither method may be appropriate to estimate the SE, and alternative, less commonly used methods may need to be used or developed. Thus, when planning a sampling design, the number of second-stage sampling elements should be reasonably large to ensure accurate parameter estimates and reliable confidence intervals, leading to true inferences. The simulations show far greater stability in the SE estimates, particularly using TSL, when the number of second-stage sampling elements are large, even for large sampling fractions. Thus, the survey design should keep a sufficient number of elements at the final stage of sampling, and also be aware if the sampling fraction of the population is large. Moreover, if using TSL, the number of clusters should be kept greater than or equal to five to ensure reliable estimates. If the survey designer knows that the second stage sample sizes will be small (often unavoidable with small class sizes in rural areas) other sampling methods, rather than cluster sampling may be more appropriate. Estimates of where divergence begins between TSL and JK1 are obtained in specific situations and presented in Section 2.4. While, as stated by Murray (1998), an overall guide may never be possible assessing when methods are inappropriate (i.e. a "small" second-stage sample size or a "large" sampling fraction), Section 2.4 gives

scenarios of when the methods differ in the context of a school setting as the parameters were chosen to model this design setup. Thus depending on class sizes, and the variable of interest (which determines distribution shape), one may be able to tell appropriate sample sizes needed, or appropriate sampling fractions, using the results in Section 2.4 (e.g. for a skewed variable such as DMFT with class sizes of $m_i = 20$, if the sampling fraction exceeds 50%, the JK1 method should not be used, and thus 50% is a large sampling fraction. However for a normal distribution this is not the case, and a large sampling fraction may be in excess of 90%).

When analysing the results in Chapter 2, one notices the presence of design effects less than one for the JK1 results. In the context of cluster sampling, these characteristics would be seen as unusual (Cochran, 2007; Lohr, 2009; Wolter, 2007), unless there is large variation within clusters and small variation between clusters (which was not present in the simulation due to specified conditions). Thus when using the JK1 estimator on data with small second stage samples and large sampling fractions, design effects less than one can be obtained, even on data that have similar characteristics per cluster. However, the only occurrence of a design effect less than one for the TSL method is when the number of clusters sampled is less than five ($n < 5$) and in this case the TSL estimator is considered to be inappropriate.

Chapter 2 highlights the need to thoroughly examine the sampling scheme of the data before survey analysis is carried out, to avoid applying the methods of survey variance estimation incorrectly. The increased availability and ease of implementation of survey sampling methods has been dramatic in the last few decades and although more and more people are implementing cluster methods, there is need to ensure the estimates they produce are appropriate. In most surveys, this will not be an issue as the sampling fraction will be small, and the number of second-stage elements will be large, however, in the case of rare characteristics, or smaller settings this is more of an issue (i.e. rural schools). Thus, in large samples with small sampling fractions both methods give almost identical estimates, and the choice between them should fall to considerations outside accuracy. However, when smaller samples are present, influenced by the clustering unit, one must be careful with the type of design chosen and the estimator used to ensure reliable results.

The simulations in Chapter 2 have examined the effect of an increasing sampling fraction on multiple distributions, numbers of clusters and numbers of second

stage sampling elements. The simulations show that at large sampling fractions the estimates under the two methods diverge, and the definition of this 'large' sampling fraction varies based on the choice of the underlying distribution. Estimates of the sampling fraction where divergence occurs can be obtained by looking at the results section (e.g. under a normal distribution, divergence between the two methods occurred when the sampling fraction became larger than 75% when there was 20 second-stage elements sampled ($m_i = 20$), or when the sampling fraction becomes larger than 30% when the number of second-stage elements sampled was $m_i = 5$), and may be used as a guide for surveys with similar characteristics. Furthermore Chapter 2 shows that small numbers of clusters and small second stage sample sizes effect the estimates under both methods, as was described in the literature review, even in the presence of increasing sampling fractions. It is also noteworthy that design effects less than one have appeared, however for the TSL method, these occur only when $n < 5$ and there is doubt whether the variance estimation method is accurate. The design effects less than one appear more frequently for the JK1 method, suggesting it becomes inappropriate when the sampling fraction is large, and an estimate of what is large can be obtained from the results of the simulation.

Chapter 2 looked at the most commonly used distributions likely to be encountered in an oral health setting, examining the normal, lognormal and two versions of the Poisson distributions. However, a similar analysis could be conducted on other distributions to see if the estimators still give comparable estimates. Furthermore, different distributional parameter estimates using the same distributions as Chapter 2 could also be examined. Moreover, the design used and parameters estimated in this simulation were quite simple, and as a result the TSL SE could be estimated by a formula, not requiring linearisation to be implemented. The design could be made more complex, or the statistic of interest could be more complex and this could highlight some further differences between the TSL and JK1 estimates. Further sources of future research based on Chapter 2 include examining other less commonly used estimators, such as the adapted BRR method for more general designs. However, these are not routinely available in standard software packages.

5.2 Chapter Three

Chapter 3 looked at using the variance estimation methods compared in Chapter 2 to calculate SE estimates of a national oral health survey in Ireland. The results of Chapter 3 show the presence of DE's less than one in a cluster sample design, which is a very unusual result that has not been encountered by the author in the literature before. As a result no attempt to identify their occurrence has been seen. Chapter 3 used the ratio estimator to calculate the survey variance estimates, as the underlying cluster sizes (M_i) vary quite a bit. The variance of this estimator is, in general, the smaller of the two estimators described. However, even using the larger unbiased estimator, there are still many DE's less than one present. This suggests that the presence of DE's less than one are due to something underlying in the data or design.

The presence of DE's less than one within a CCA indicate that, in the case of this survey measuring children's dmft score, cluster sampling is a more efficient method than simple random sampling within this CCA. This suggests that the usual adjustment to increase the sample size when using cluster sampling is not required in this situation, and in fact the opposite, that sampling less observations may give the same precision as a SRS. A potential explanation may be children's dmft scores are not influenced by their cluster unit (school), lending to large variation within the cluster, however the overall average dmft estimates are close. This would be one scenario where the DE estimate would be less than one, however this would be very unusual in a cluster sample, as in general clusters have similar characteristics. If one was able to identify when this scenario occurs, this could have large cost saving effects for future surveys as the number of children sampled could potentially be reduced.

As no explanation for design effects less than one in a cluster setting had been encountered in the literature by the author before, Chapter 3 took an exploratory approach to ascertain some characteristic(s) of the data\design which might identify the occurrence of a design effect less than one. The initial exploration examined a result from Chapter 2. Earlier in the simulation conducted in Chapter 2, there are DE's less than one present, however these were extremely rare when using the ratio (TSL) estimator. The ratio estimator only gave design effects less than one when the number of clusters (n) was less than five, however this situation never occurs in the data in Chapter 3. Thus, further exploration was undertaken to identify when a design effect less than one would be present. An-

other idea explored in Chapter 3 suggested that the geographic location of the schools, and thus the pupils attending may influence the DE estimates. However, after implementing two forms of cluster analysis, no obviously interpretable geographic groupings were produced. Chapter 3 also focused on identifying if there was some underlying, measurable variable which may influence if DE's less than one would be present. Linear model methods suggested that some variable(s) may be useful in detecting/influencing a DE less than one.

Both simple linear regression and generalised linear model methods were implemented on a range of characteristics of each CCA, to ascertain if any characteristic could identify when a design effect less than one would be present within a CCA. Manual backward selection techniques were implemented to identify the variables of use within each model. Regression and GLM methods were carried out on the design effect estimates of the dmft and caries free (a binary representation of dmft) variables per CCA. The potential identifying variable which appeared most frequently, and is also a variable that can be decided in the sample design, was the number of clusters sampled per CCA (n). The regression results suggested that for this data, one can reduce the design effect by reducing the number of clusters sampled per CCA. This reduction cannot be too extreme, as sampling only one school would not be representative of the CCA, and would in turn increase the DE. This suggested that Chapter 4 should look at reducing the number of clusters that one could sample given DE results similar to that observed in Chapter 3.

However, while the linear model methods indicated the number of clusters sampled (n) as the variable most useful in identifying the presence of a design effect less than one, it is noteworthy that the R-squared value of the linear models is quite low, suggesting other variables, possibly not observed in this particular study may help improve the model fit, and this is an area for future research. Furthermore, this analysis was only conducted on one dataset and it would be useful to take some other oral health studies completed both in Ireland and globally to ascertain if design effects less than one are also present in these iterations.

Moreover, further work could look at the magnitude of the design effect estimates. While one is chosen as a hard cut off for the design effect, many of the estimates under the ratio estimator are close to one, and potentially the effect is not going to strongly influence the cost and time required to sample the children. One final limitation of the results in Chapter 3 is that the estimates of N and M_i are taken to be all available schools and children within the CCA, as the fluoridation status

of the child is not known until after sampling. Thus the estimates of N and M_i are larger than the true number of schools and children available, leading to larger variance estimates than if the true N and M_i estimates were known. However, as explained in Section 3.3.5, this is the only feasible option for estimation, other than omitting the FPC.

5.3 Chapter Four

The methodology used in Chapter 4 looked at systematically reducing the number of clusters sampled by $1, 2, \ldots, 10$ or until the maximum number of clusters was reached, retaining a minimum of five clusters per CCA to ensure reliable estimates using the ratio estimator. The analysis looked at all possible combinations of resamples within a CCA, and random samples were only drawn in few cases, where the number of resamples became too large for a reasonable computing runtime. Once the data was divided by age group and fluoridation status, a cut off of 24 hours per CCA was chosen as the upper limit of runtime. The SE of the results were examined and compared to the initial estimate of SE. Furthermore, the new SE was compared to the original SRS SE to allow examination of the DE. For each replication of SE, the bias relative to the original estimate was also observed to assess if the estimate remained accurate.

Based on the results of Chapter 4, Figure 4.1 to Figure 4.5 are compiled allowing comparison of the reduction in number of clusters versus the percentage change in bias and SE. While the actual number of clusters reduced will be chosen by the survey designer, the author notes that reducing by one or two clusters per CCA makes no difference in bias and a very small percentage increase in SE. Although this reduction in number of clusters is small per CCA, even if only one school is reduced across all CCA's, by age grouping and fluoridation status this would be a saving of 208 schools. For this survey to be conducted, each school received a dentist to undertake the examination and a recorder to record the results of the examination. Further savings include a reduction in dental consumables, a reduction in travel time and costs, which could be substantial depending on where the clinicians and school were based. Moreover, if less schools are sampled this reduces the burden on schools, allowing assessments like this to be conducted, and if the assessments are less frequent, this may encourage schools to participate in future iterations of the surveys.

The interpretation of the results will require some rules and subjectivity depend-

ing on the number of clusters in the CCA. Initially it is important to ensure there are sufficient clusters to avoid the small cluster issue and, as discussed in Section 4.3 (paragraph three), the author suggests that five clusters are left per CCA. This should ensure that the bias of the estimate is at a minimum, and that a representative sample is drawn. While the 90^{th} percentile estimates are available for the change in SE, one must remember that these are compiled from the mean of all resamples, and if a designer is to reduce the number of clusters they will not obtain all these resamples. Thus, there is a chance that the designer could obtain a large increase in SE or a biased sample. However even without reducing the number of clusters, this is a possibility as the underlying data is just one sample of the population available at the time the survey was conducted. The designer should consider that the increase in SE is tolerable, or that the saving in costs is worth the increase in SE. Finally the designer must ensure that the increase in SE will not lead to a clinically significant effect/difference being missed due to a larger CI.

Oral health literature suggests that a difference of 0.5 in the dmft for estimates of the 5, 12 and 15 year old would be a clinically significant finding, while a difference in dmft of 0.2 at age 8 would be clinically significant (James et al., 2018). Based on the overall dmft estimates at differing age groups, if the bias increases by 5% this will cause the dmft estimate to change by no more than 32% of a clinically significant difference. This results hold across all age groups and fluoridation status', but the percentage change is substantially lower for the younger age groupings and fluoridated age groups. Looking at the SE estimates, allowing the SE to increase by 5% caused the CI to change by at most 34% of a clinically significant difference. However, this is the largest estimate across all fluoridation status' and age groupings. As with bias, the SE change was smaller for younger age groupings and fluoridated groups. Thus, the author suggests that the survey designer should look at these results for each fluoridation status and age group, rather than an overall survey adjustment.

The results show that the increase in SE is quite small when a small number of clusters are omitted and this could lead to a significant cost saving in the planning of a design. The analysis in Chapter 4 was only conducted for the dmft and caries free variables, and while these are commonly used, further work could examine the effect on other variables of interest (e.g. dmfs, BMI, etc.). Furthermore, this resampling was only carried out on one survey dataset, and an area of future research would involve assessing the affect on other datasets to ensure the increases in SE are comparable. Several regional and national oral

health studies have been conducted in Ireland (Whelton et al., 2004; Parnell et al., 2007; Whelton, O'Mullane and Cronin, 2001*a*,*b*,*c*), and globally (Do and Spencer, 2016; Lu et al., 2018; Vernazza et al., 2016), and it would be of interest to see if design effects less than one are also present in these surveys, and if so to examine the effect of reducing the number of clusters sampled on the SE estimates in these surveys. If the results are similar for past oral health surveys in Ireland, one may be able to reduce the sample size inflation for clustered data on future iterations of oral health studies which would have a large financial impact. Moreover it would be of interest to assess if other variables observe similar design effect estimates in surveys outside of oral health with clustering at school level.

In conclusion, this thesis compared the most routinely implemented variance estimation methods using simulation on samples of data with varying characteristics including distributional assumptions and sample size restrictions frequently encountered in an oral health setting. These methods showed diverging estimates when the sampling fraction was large, and the number of observations, particularly at second stage, was small, with estimates of small and large being presented under multiple distributions. Furthermore this thesis analysed a national oral health survey, exploring cluster sampling efficiency versus simple random sampling, encountering unseen results meriting a thorough exploratory investigation. This investigation suggested that reducing the number of clusters sampled could reduce the design effect of the data. Finally this thesis explored the effect of reducing the number of clusters sampled on the bias, standard error and design effect estimates of oral health variables. The results showed that a small number of clusters per CCA can be dropped, with no impact on the bias of the estimate and a very small increase on the SE of the estimate, showing potential for some large cost savings if results of future surveys are comparable.

References

Binder, David A. 1983. "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review/Revue Internationale de Statistique* pp. 279–292.

Brillinger, David R. 1970. "Approximate estimation of the standard errors of complex statistics based on sample surveys." *Change* 61(61.7):2.

Bruch, Christian, Ralf Münnich and Stefan Zins. 2011. "Variance estimation for complex surveys." *Deliverable D3* 1.

Cameron, A Colin and Douglas L Miller. 2015. "A practitioner's guide to cluster-robust inference." *Journal of Human Resources* 50(2):317–372.

Carle, Adam C. 2009. "Fitting multilevel models in complex survey data with design weights: Recommendations." *BMC medical research methodology* 9(1):49.

Chowdhury, Sadeq R. 2013. "A comparison of Taylor linearization and balanced repeated replication methods for variance estimation in medical expenditure panel survey." *Rockville, MD: Agency for Healthcare Research and Quality* .

Cochran, William G. 2007. *Sampling techniques*. John Wiley & Sons.

Cornfield, Jerome. 1978. "Symposium on CHD prevention trials: Design issues in testing life style intervention randomization by group: A formal analysis." *American journal of epidemiology* 108(2):100–102.

Do, Loc G and A John Spencer. 2016. *Oral health of Australian children: the National Child Oral Health Study 2012–14*. University of Adelaide Press.

EasyGuides. 2018. "Assessing Clustering Tendency.".
URL: *https://www.datanovia.com/en/lessons/assessing-clustering-tendency/*

Eldridge, Sandra and Sally Kerry. 2012. *A practical guide to cluster randomised trials in health services research*. Vol. 120 John Wiley & Sons.

Eurostat. 2002. *Monographs of Official Statistics: Variance Estimation Methods in the European Union*. Luxembourg: Office for Official Publications of the European Communities.

Hair, Joseph F. 2006. *Multivariate data analysis*. Vol. 7.

Hammer, Heather, Hee-Choon Shin and Lorraine E Porcellini. 2003. A Comparison of Taylor Series and JK1 Resampling Methods for Variance Estimation. In *Proceedings of the Hawaii International Conference on Statistics*. pp. 1–9.

Hannigan, Ailish and Christopher D. Lynch. 2013. "Statistical methodology in oral and dental research: Pitfalls and recommendations." *Journal of Dentistry* 41(5):385–392.
URL: *http://www.sciencedirect.com/science/article/pii/S0300571213000584*

Hartigan, John A and Manchek A Wong. 1979. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1):100–108.

James, Patrice, Mairead Harding, Tara Beecher, Carmel Parnell, Deirdre Browne, Marie Tuohy, Dympna Kavanagh, Denis O'Mullane, Helena Guiney, Michael Cronin et al. 2018. "Fluoride And Caring for Children's Teeth (FACCT): Clinical Fieldwork Protocol." *HRB Open Research* 1.

Kassambara, Alboukadel. 2017. *Practical guide to cluster analysis in R: Unsupervised machine learning*. Vol. 1 STHDA.

Killip, Shersten, Ziyad Mahfoud and Kevin Pearce. 2004. "What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers." *Annals of Family Medicine* 2(3):204–208. 0020204[PII] 15209195[pmid] Ann Fam Med.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1466680/ http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1466680/pdf/0020204.pdf*

Kish, Leslie and Martin R Frankel. 1970. "Balanced repeated replications for standard errors." *Journal of the American Statistical Association* 65(331):1071–1094.

Kish, Leslie and Martin Richard Frankel. 1974. "Inference from complex samples." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–37.

Koop, JC. 1972. "On the derivation of expected value and variance of ratios without the use of infinite series expansions." *Metrika* 19(1):156–170.

Kovar, JG, JNK Rao and CFJ Wu. 1988. "Bootstrap and other methods to measure errors in survey estimates." *Canadian Journal of Statistics* 16(S1):25–45.

Lee, Eun Sul and Ronald N Forthofer. 2005. *Analyzing complex survey data*. Vol. 71 Sage Publications.

Lemeshow, Stanley, David W Hosmer and David Hislop. 1980. "The effect of non-normality on estimating the variance of the combined ratio estimate in complex surveys: The effect of non-normality on estimating." *Communications in Statistics-Simulation and Computation* 9(4):371–387.

Lesaffre, Emmanuel, Jocelyne Feine, Brian Leroux and Dominique Declerck. 2009. *Statistical and methodological aspects of oral health research*. Wiley Online Library.

Lohr, Sharon. 2009. *Sampling: design and analysis*. Nelson Education.

Lu, Hai Xia, Dan Ying Tao, Edward Chin Man Lo, Rui Li, Xing Wang, Bao Jun Tai, Y Hu, Huan Cai Lin, Bo Wang, Yan Si et al. 2018. "The 4th National Oral Health Survey in the Mainland of China: Background and Methodology." *The Chinese journal of dental research: the official journal of the Scientific Section of the Chinese Stomatological Association (CSA)* 21(3):161–165.

Lumley, Thomas. 2011. *Complex surveys: a guide to analysis using R*. Vol. 565 John Wiley & Sons.

MacQueen, James. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1 Oakland, CA, USA pp. 281–297.

Murray, David M. 1998. *Design and analysis of group-randomized trials*. Vol. 29 Oxford University Press, USA.

Paben, Steven. 1999. Comparison of Variance Estimation Methods for the National Compensation Survey. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Parnell, C, E Connolly, M O'Farrell, M Cronin, E Flannery, H Whelton et al. 2007. "Oral health of 5 year old children in the North East, 2002.".

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Rao, P.S.R.S. 2000. *Sampling methodologies*. Boca Raton: Chapman & Hall / CRC.

Rust, Keith. 1985. "Variance estimation for complex estimators in sample surveys." *Journal of Official Statistics* 1(4):381–397.

Särndal, Carl-Erik, Bengt Swensson and Jan Wretman. 2003. *Model assisted survey sampling*. Springer Science & Business Media.

SAS Institute, Inc. 2016. "SAS/ACCESS 9.4 Interface to ADABAS: Reference.".

Spitznagel, Edward L. 1985. "Comparison of Variance Estimation Methods for Complex Sample Designs Under Extreme Conditions.".
URL: *www.sascommunity.org/sugi/SUGI85/Sugi-10-193%20Spitznagel.pdf*

Vernazza, CR, SL Rolland, B Chadwick and N Pitts. 2016. "Caries experience, the caries burden and associated factors in children in England, Wales and Northern Ireland 2013." *British dental journal* 221(6):315.

Wears, Robert L. 2002. "Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data." *Academic emergency medicine* 9(4):330–341.

Weisberg, Sanford. 2005. *Applied linear regression*. Vol. 528 John Wiley & Sons.

Whelton, H, DM O'Mullane and M Cronin. 2001*a*. "Children's Dental Health in the Mid-Western Health Board Region 1997." *Mid-Western Health Board, Limerick* .

Whelton, H, DM O'Mullane and M Cronin. 2001*b*. "Children's Dental Health in the North Western Health Board Region 1997." *North Western Health Board, Donegal* .

Whelton, H, DM O'Mullane and M Cronin. 2001*c*. "Children's Dental Health in the South Eastern Health Board Region 1998." *South Eastern Health Board, Waterford* .

Whelton, H, E Crowley, D O'Mullane, M Donaldson, V Kelleher and M Cronin. 2004. "Dental caries and enamel fluorosis among the fluoridated and non-fluoridated populations in the Republic of Ireland in 2002." *Community Dental Health* 21(1):37–44.

Whelton, Helen, D O'Mullane, M Harding, H Guiney, M Cronin, E Flannery

and Virginia Kelleher. 2006. "North South Survey of Children's Oral Health in Ireland 2002.".

WHO. 2013. *Oral Health Surveys: Basic Methods*. World Health Organization.

Wolter, Kirk. 2007. *Introduction to variance estimation*. Springer Science & Business Media.

Appendix A

Choosing Number of Replications

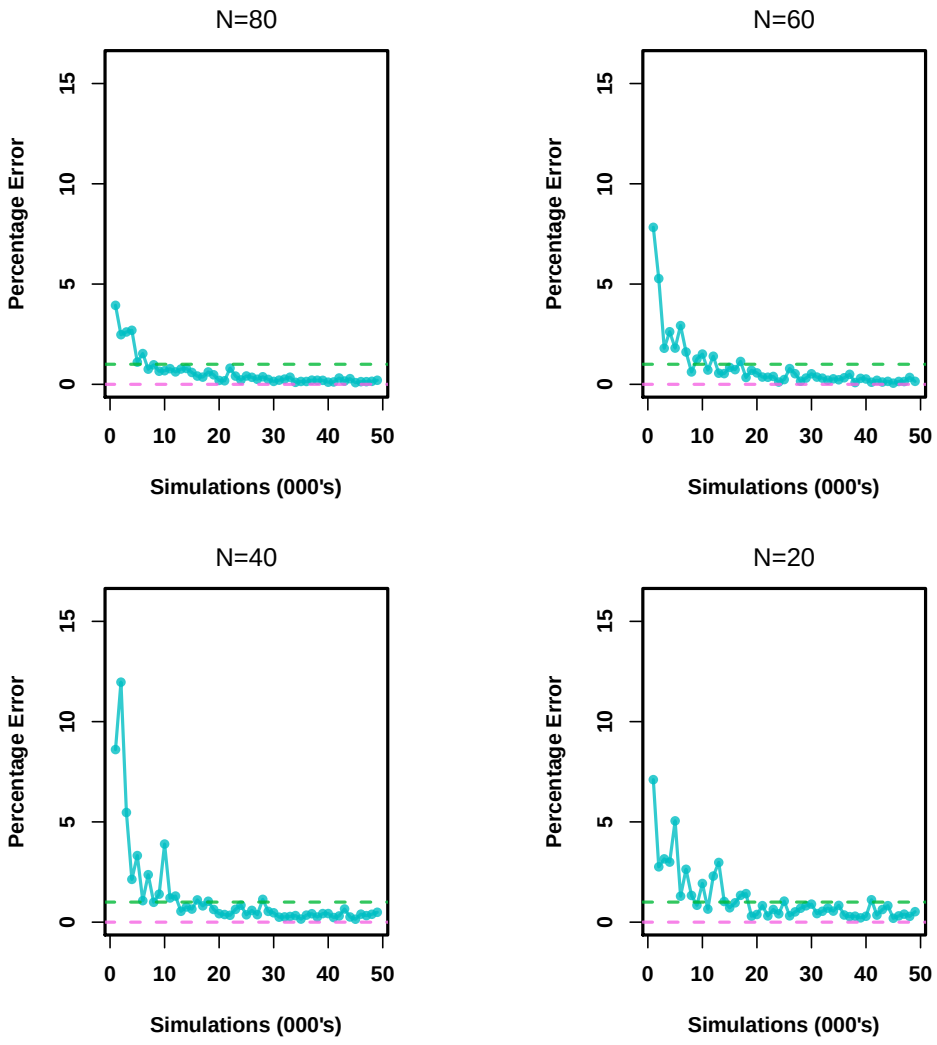


Figure A.1: Largest SE for Poisson($\lambda \sim U(3.5, 6.5)$) distribution across all m_i values for $N = 80, 60, 40$ and 20.

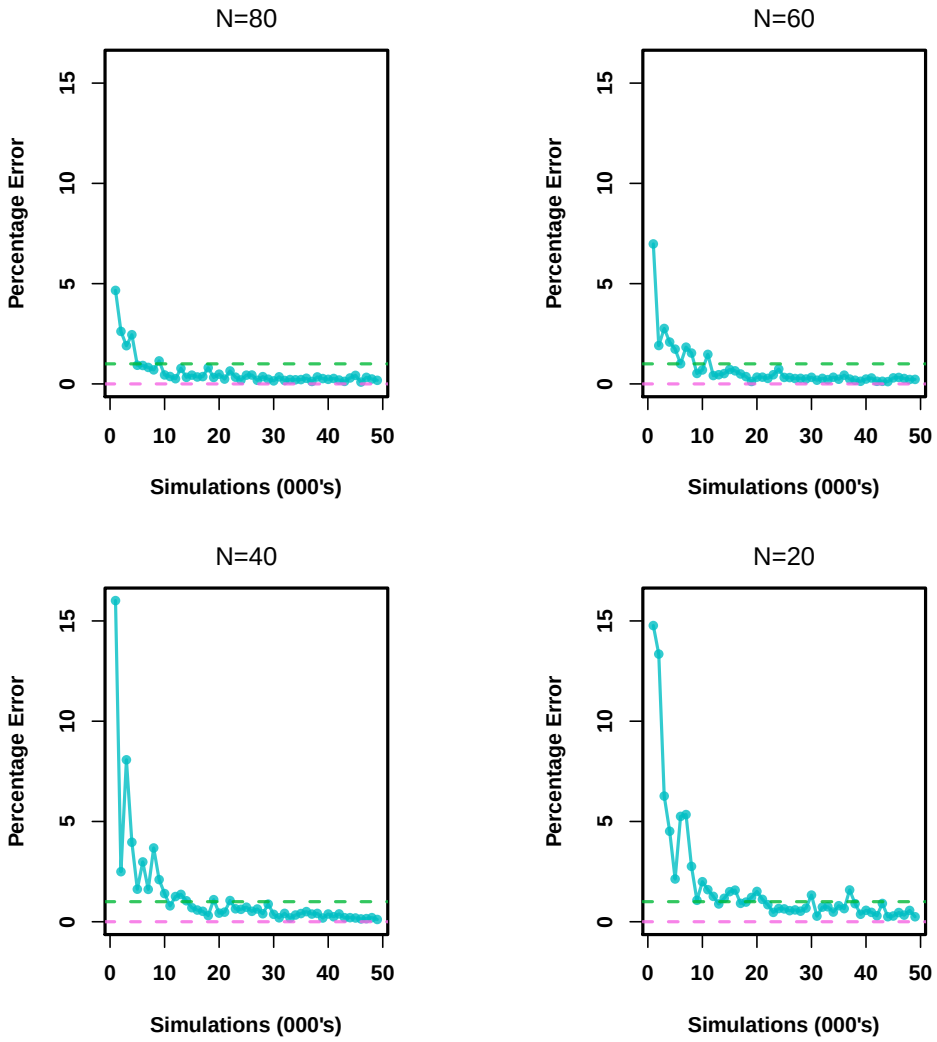


Figure A.2: Largest SE for Poisson($\lambda \sim U(0.5, 1.5)$) distribution across all m_i values for $N = 80, 60, 40$ and 20.

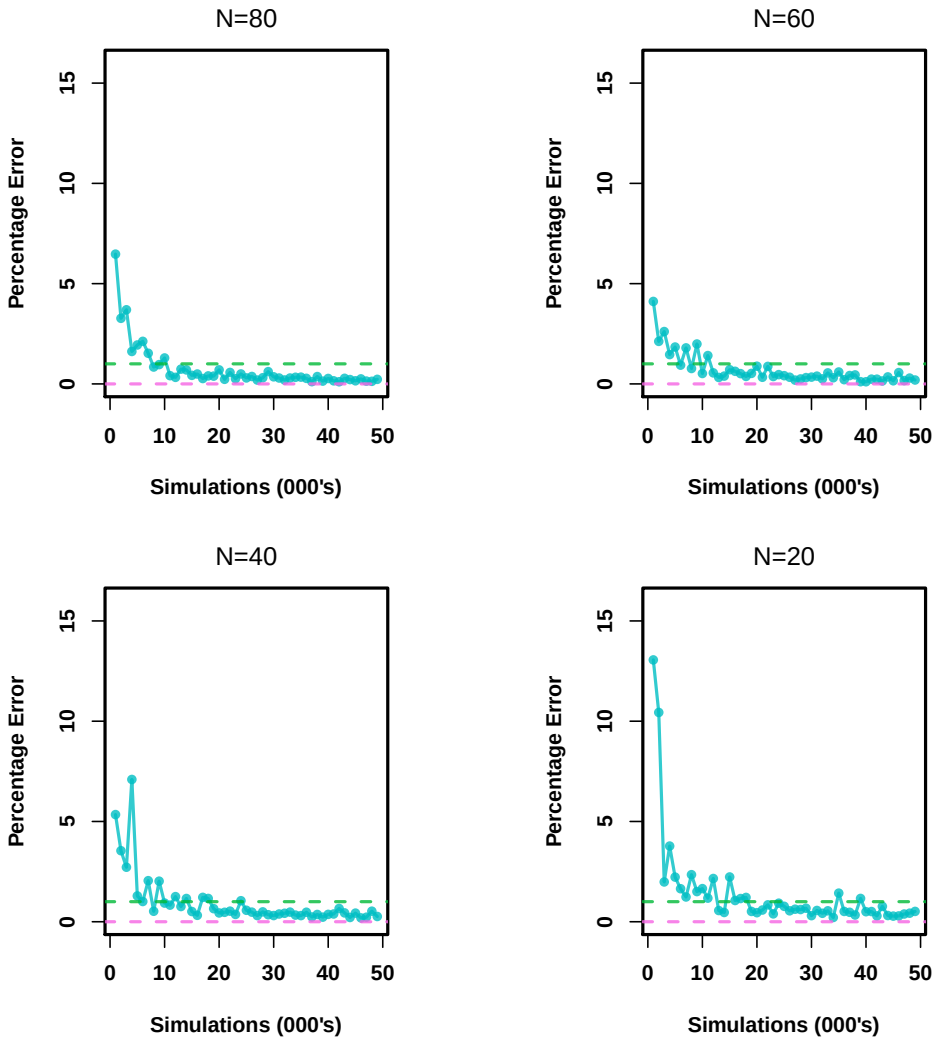


Figure A.3: Largest SE for lognormal distribution across all m_i values for $N = 80, 60, 40$ and 20.

Appendix B

SE Estimates for $N = 60, 40$ and 20

B.1 Normal Distribution

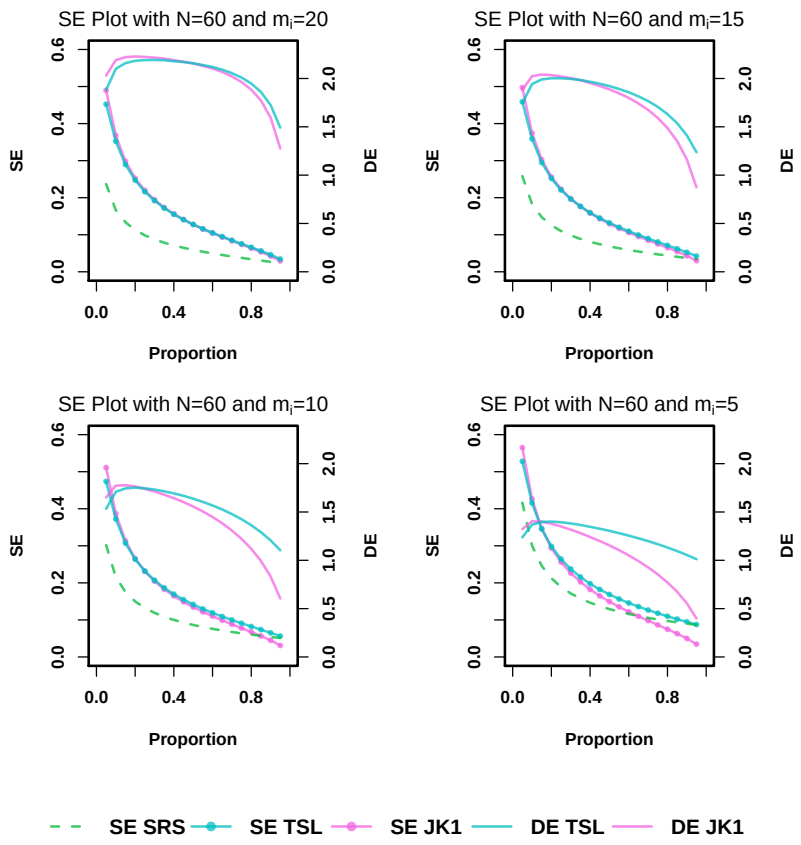


Figure B.1: SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes for N=60 clusters.

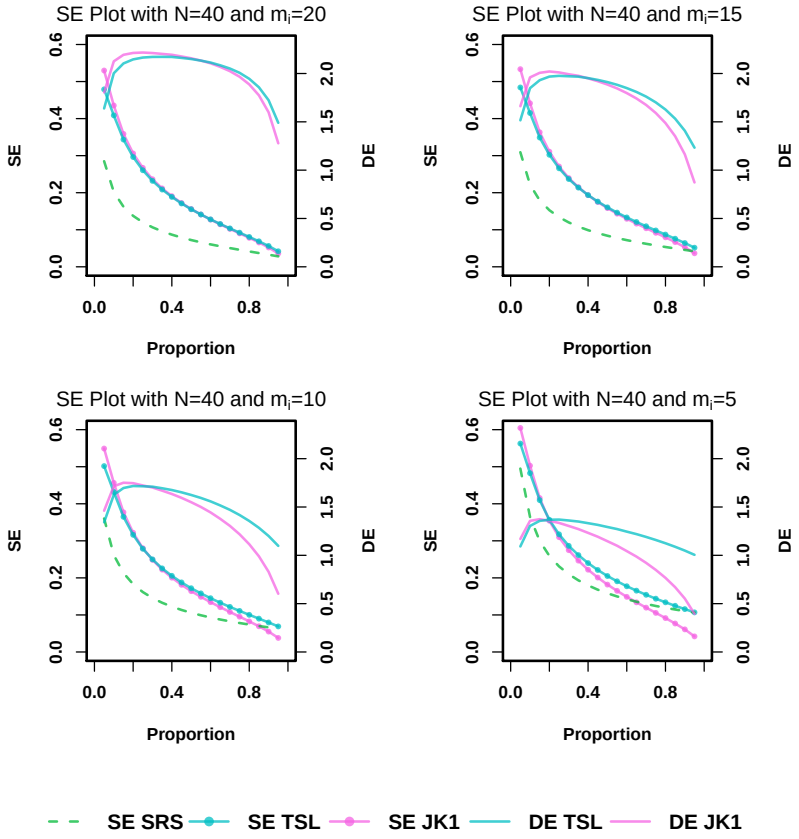


Figure B.2: SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes for N=40 clusters.

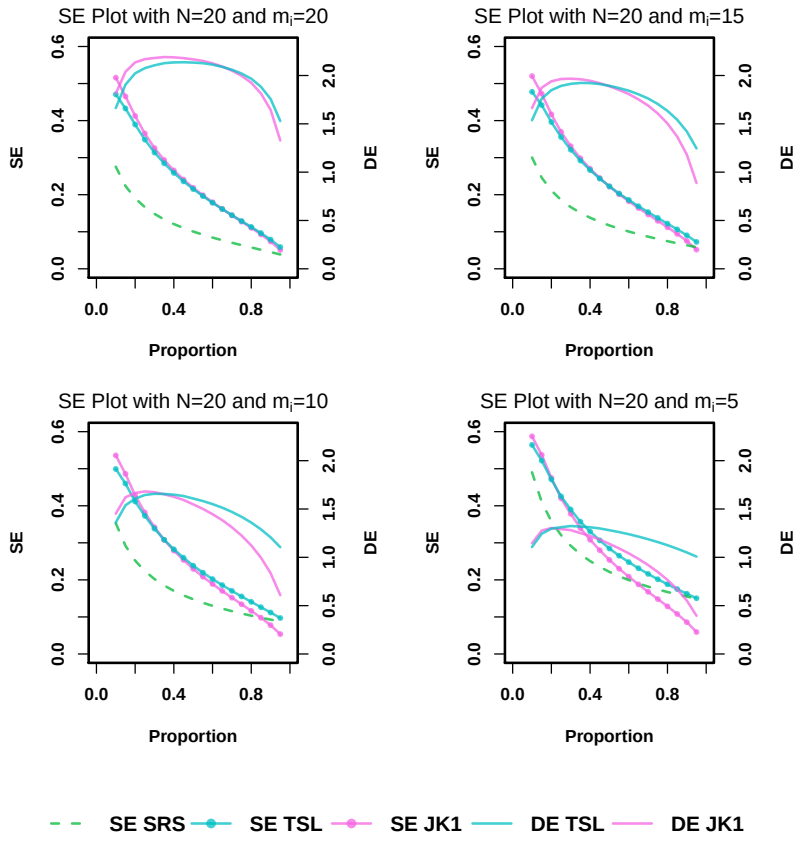


Figure B.3: SE and DE estimates under SRS, TSL and JK1 variance methods for a Normal ($\mu \sim U(3.5, 6.5), \sigma = 1.5$) distribution, with varying second-stage sampling sizes for N=20 clusters.

B.2 Poisson $\lambda \sim U(3.5, 6.5)$ Distribution

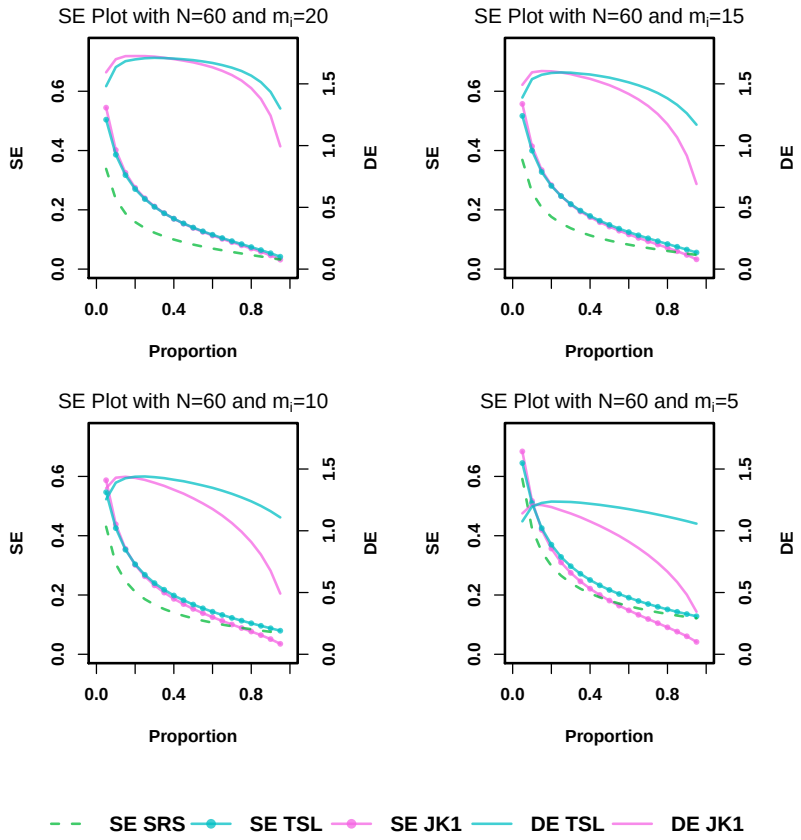


Figure B.4: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes for N=60 clusters.

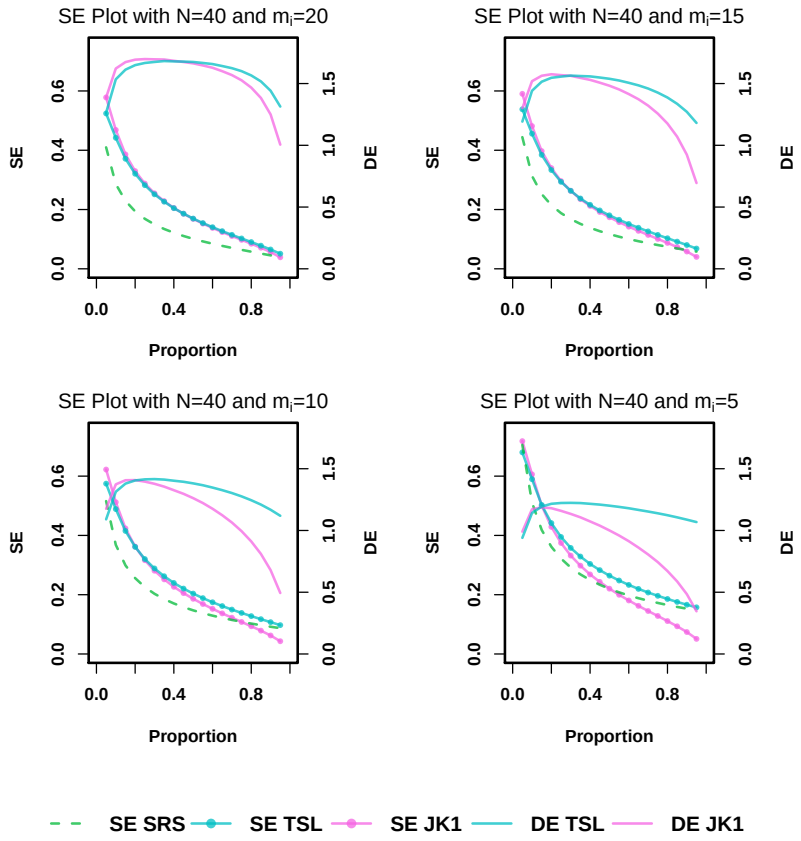


Figure B.5: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes for N=40 clusters.

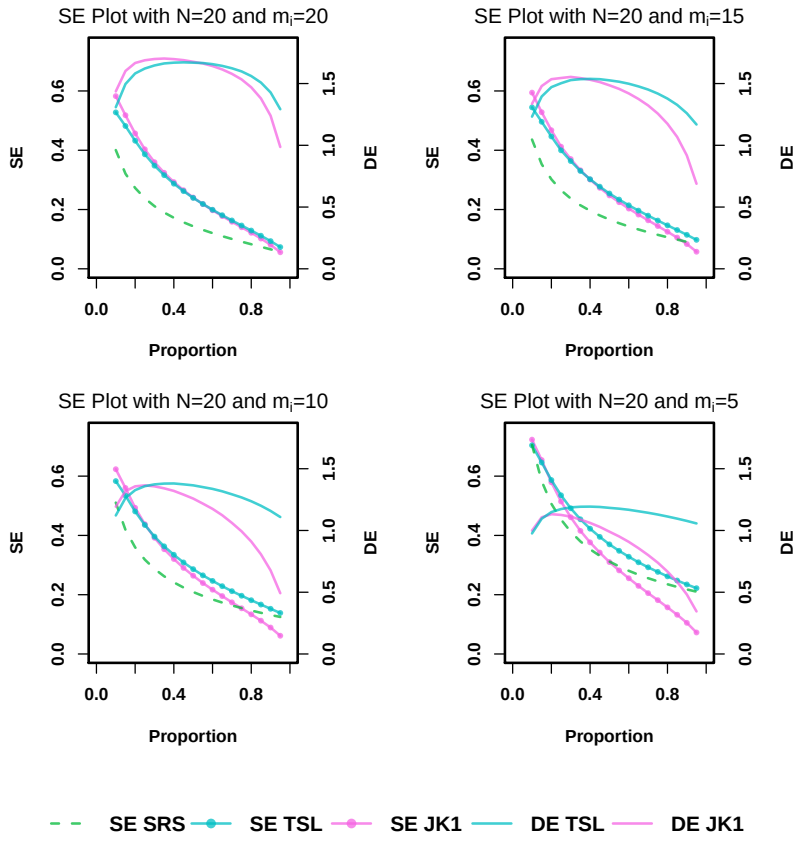


Figure B.6: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(3.5, 6.5)$) distribution, with varying second-stage sampling sizes for N=20 clusters.

B.3 Poisson $\lambda \sim U(0.5, 1.5)$ Distribution

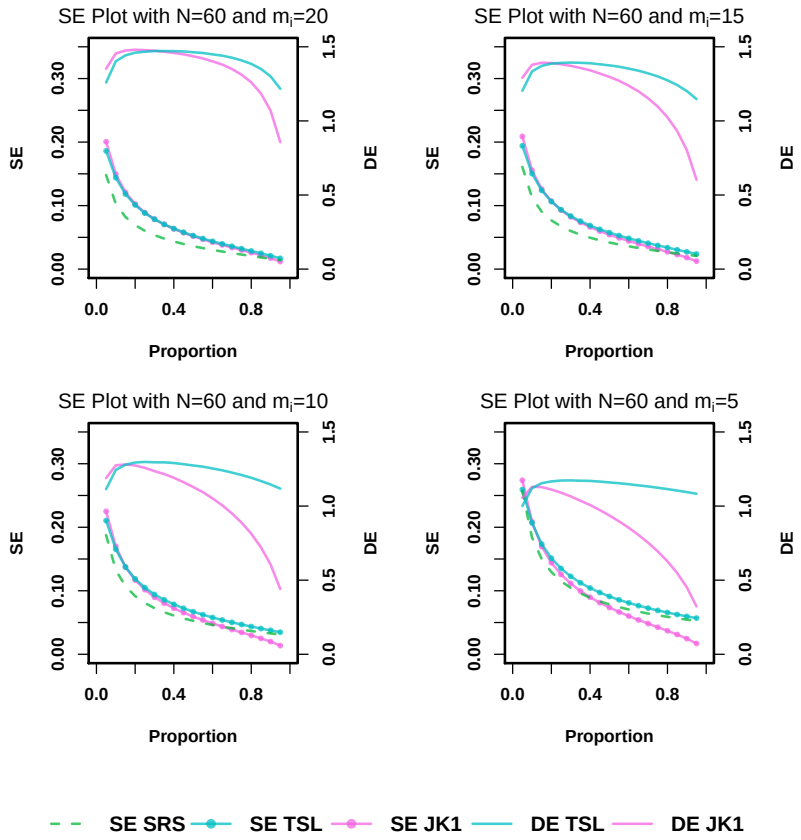


Figure B.7: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes for N=60 clusters.

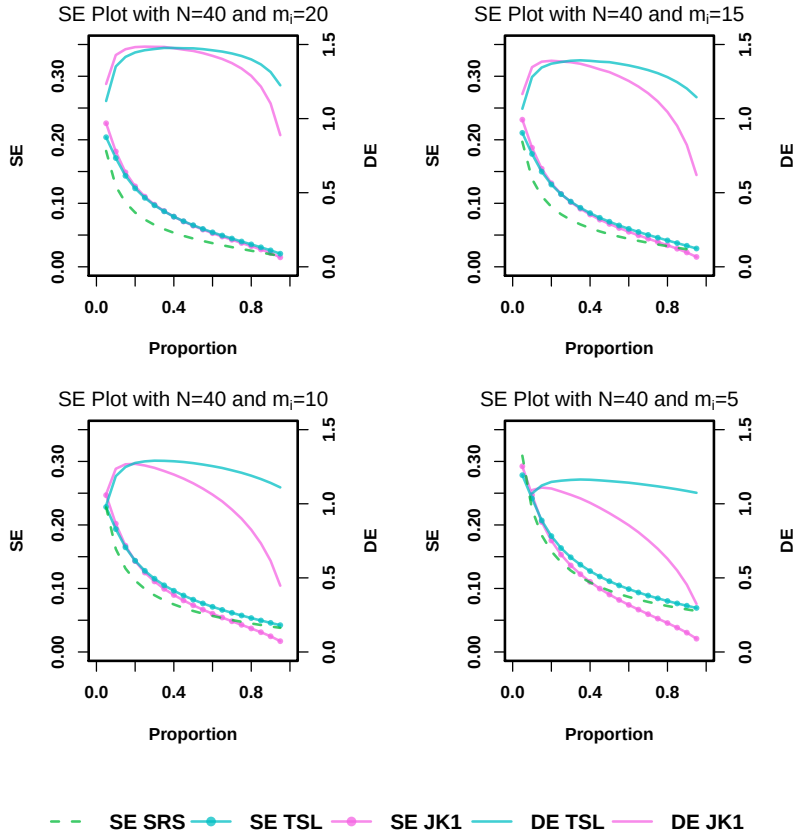


Figure B.8: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes for N=40 clusters.

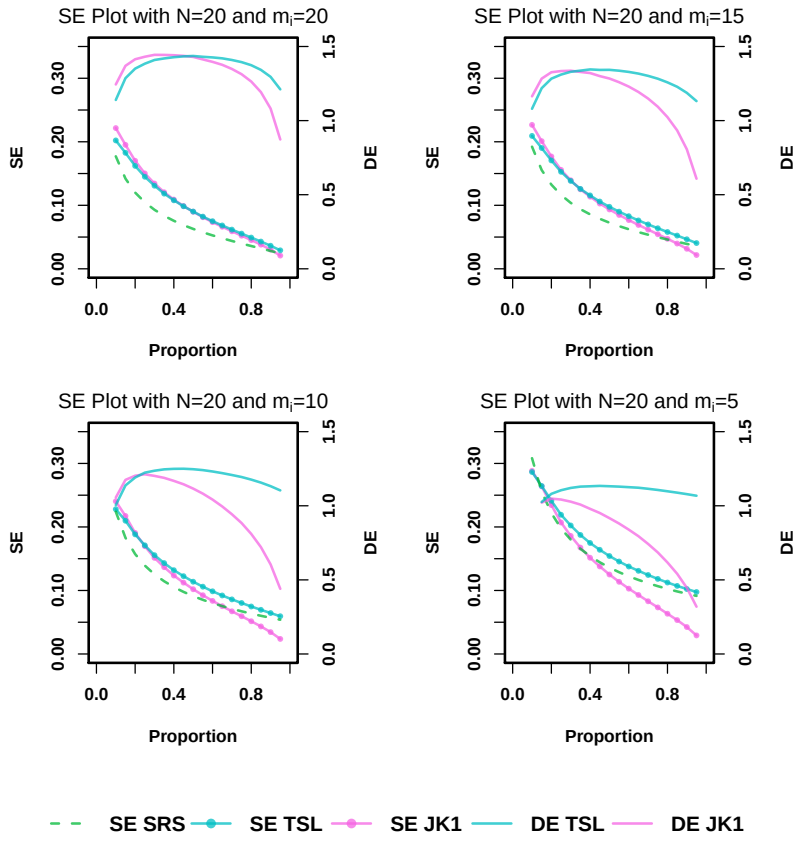


Figure B.9: SE and DE estimates under SRS, TSL and JK1 variance methods for a Poisson ($\lambda \sim U(0.5, 1.5)$) distribution, with varying second-stage sampling sizes for N=20 clusters.

B.4 Lognormal Distribution

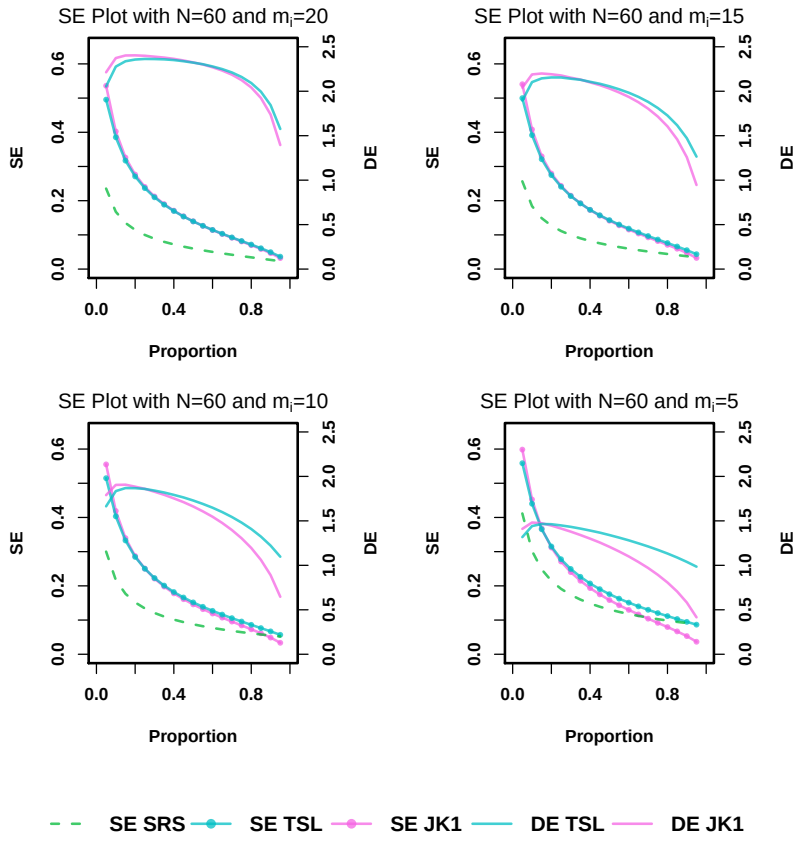


Figure B.10: SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes for N=60 clusters.

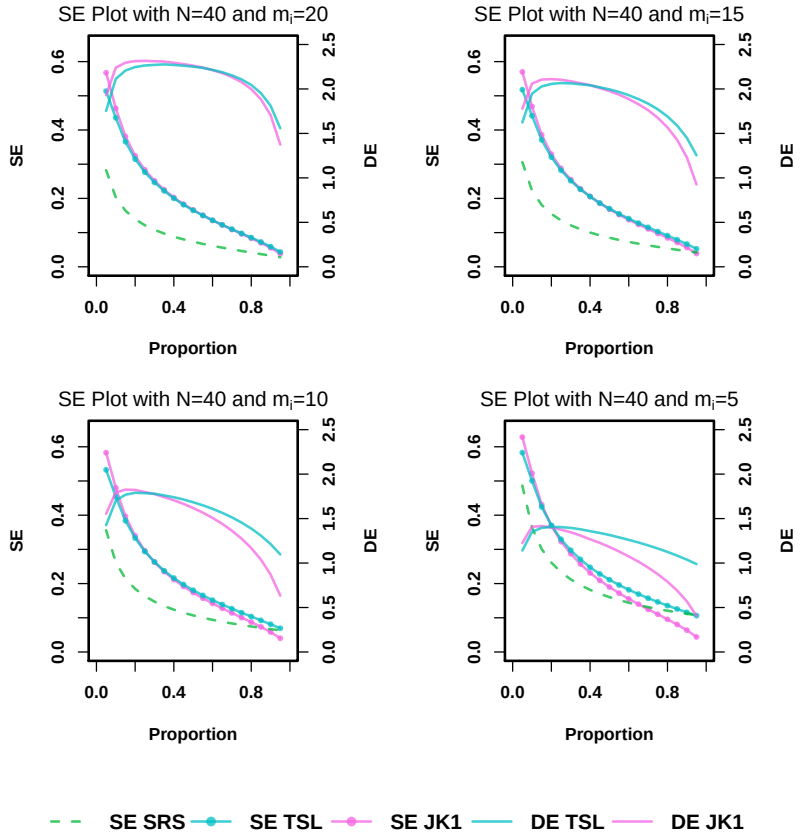


Figure B.11: SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes for N=40 clusters.

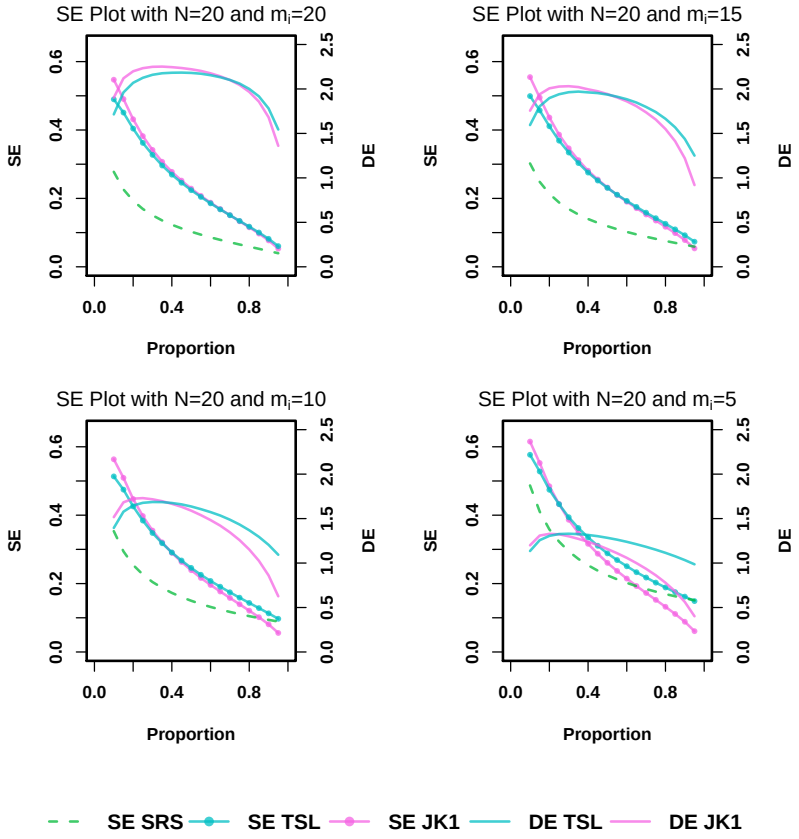


Figure B.12: SE and DE estimates under SRS, TSL and JK1 variance methods for a Lognormal ($\mu \sim U(1.1685, 1.8459), \sigma = 0.2936$) distribution, with varying second-stage sampling sizes for N=20 clusters.

Appendix C

Distributions of Oral Health Variables

C.1 Fluoridated dmft Observations

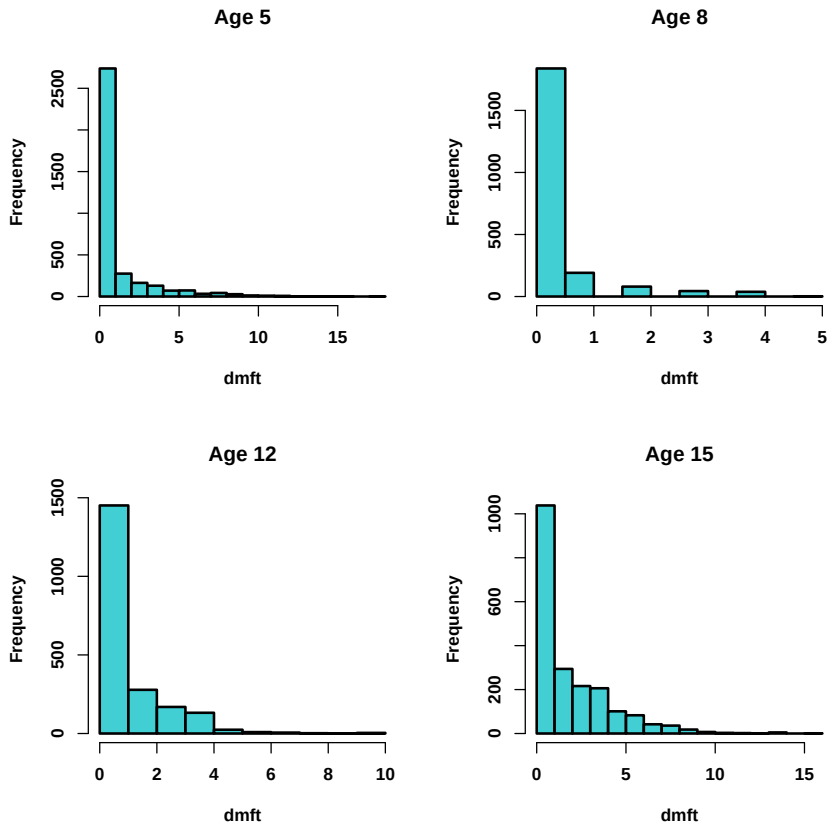


Figure C.1: Histogram of dmft values for fluoridated observations.

C.2 Non-Fluoridated dmft Observations

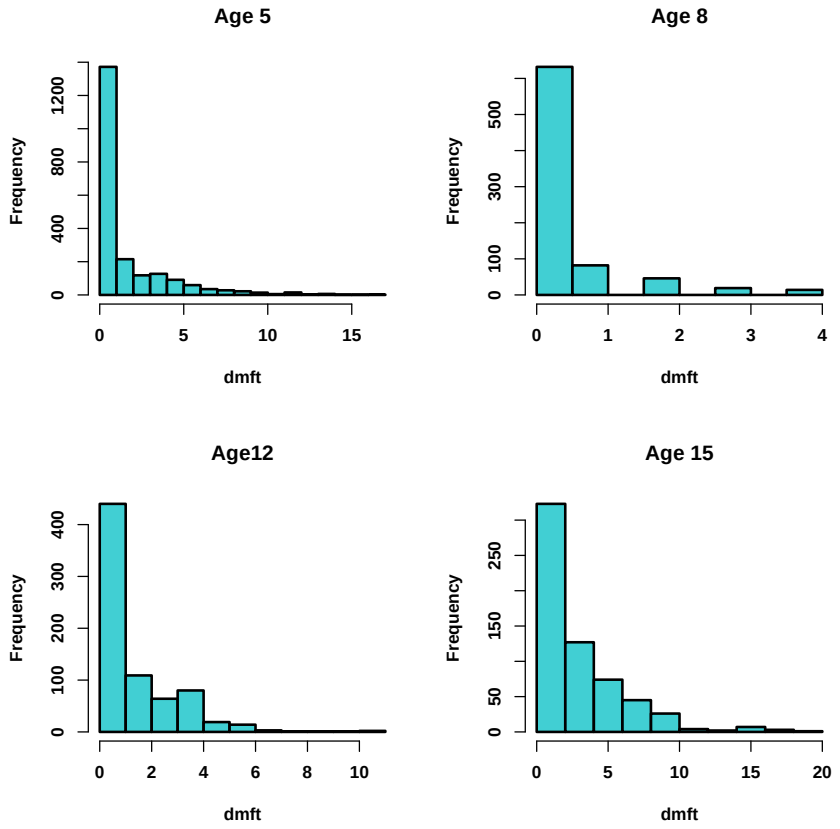


Figure C.2: Histogram of dmft values for non-fluoridated observations.

C.3 Fluoridated dmfs Observations

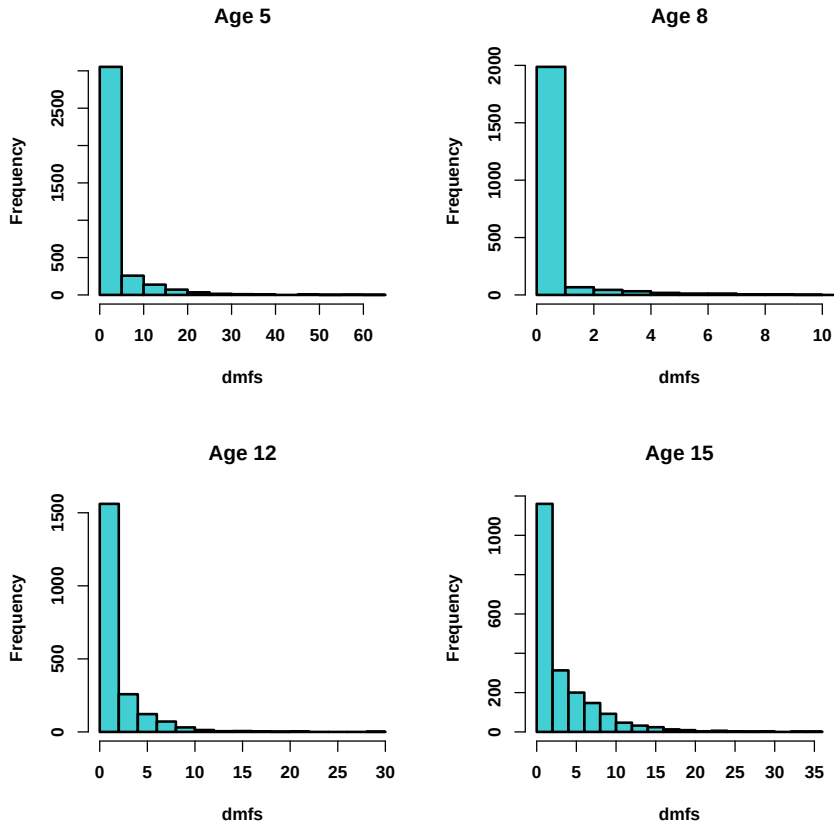


Figure C.3: Histogram of dmfs values for fluoridated observations.

C.4 Non-Fluoridated dmfs Observations

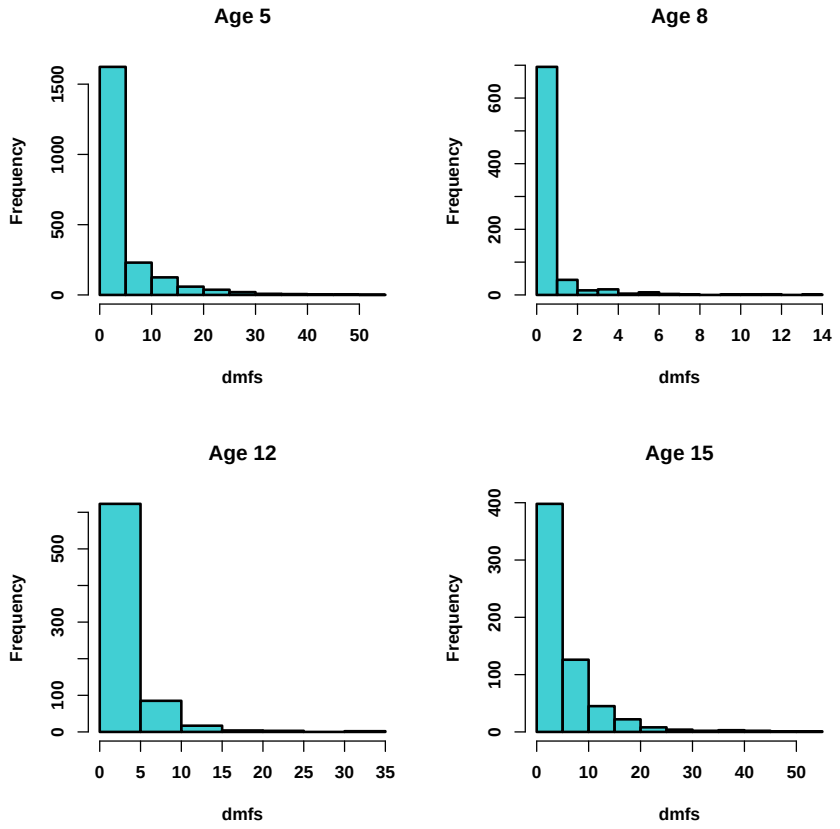


Figure C.4: Histogram of dmfs values for non-fluoridated observations.

C.5 Fluoridated BMI Observations

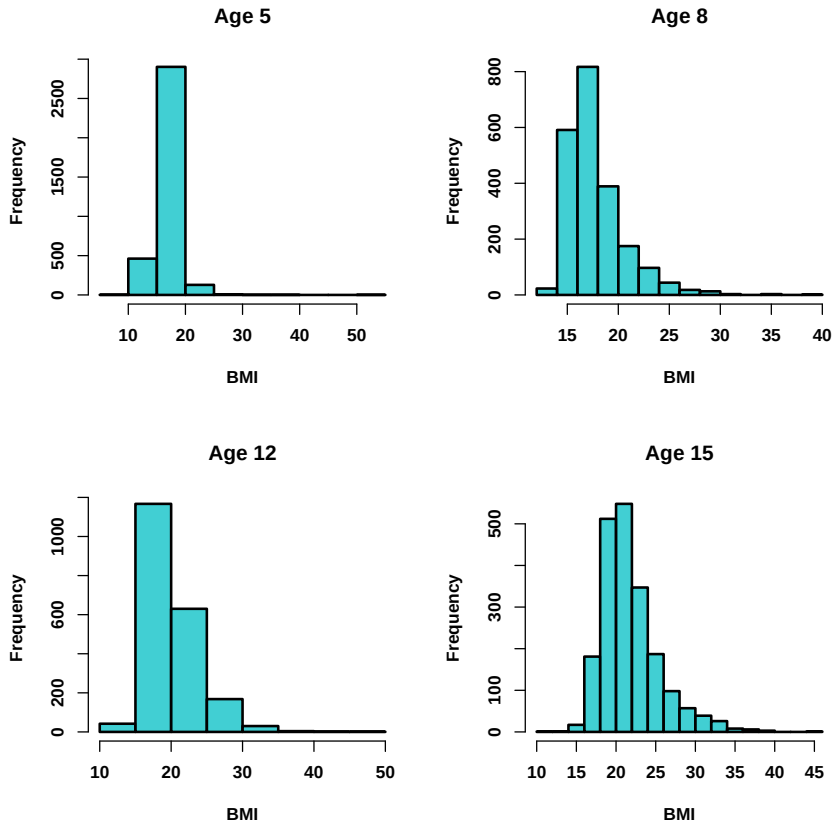


Figure C.5: Histogram of BMI values for fluoridated observations.

C.6 Non-Fluoridated BMI Observations

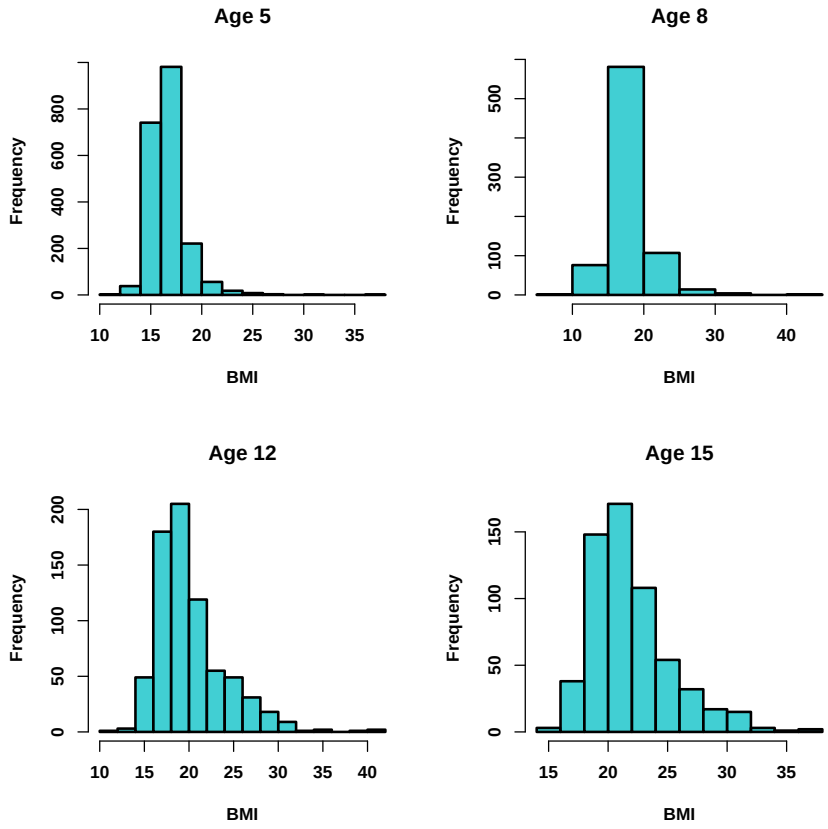


Figure C.6: Histogram of BMI values for non-fluoridated observations.

Appendix D

Caries Free Design Effect Results

Table D.1: Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 5

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	12	64	107	493	0.781	0.781	0.123	0.041	0.038	10.196	1.128
2	11	59	75	449	0.844	0.844	0.243	0.064	0.050	23.907	1.630
3	12	46	130	670	0.597	0.597	0.116	0.064	0.042	7.704	2.356
4	0	0	0	0	0	0	0	0	0	0	0
5	13	67	140	624	0.803	0.803	0.165	0.037	0.034	23.047	1.175
6	11	54	101	485	0.616	0.616	0.132	0.053	0.047	7.999	1.286
7	14	56	83	869	0.859	0.859	0.155	0.055	0.040	15.370	1.951
8	12	102	87	524	0.696	0.696	0.163	0.068	0.046	12.314	2.153
9	16	123	92	459	0.635	0.635	0.096	0.051	0.049	3.809	1.059
10	16	102	82	401	0.671	0.671	0.161	0.048	0.053	9.341	0.821
11	10	112	49	263	0.739	0.739	0.248	0.056	0.063	15.396	0.788
12	10	108	53	241	0.583	0.583	0.184	0.080	0.069	7.173	1.359
13	15	94	77	355	0.607	0.607	0.173	0.085	0.055	9.950	2.369
14	47	122	182	732	0.662	0.662	0.089	0.040	0.033	7.230	1.484
15	45	75	626	1,435	0.632	0.632	0.048	0.020	0.017	8.339	1.379
16	70	95	744	1,618	0.713	0.713	0.047	0.017	0.014	11.948	1.525
17	14	168	57	290	0.503	0.503	0.124	0.070	0.066	3.505	1.127
18	9	98	29	167	0.719	0.719	0.180	0.098	0.084	4.611	1.372
19	16	99	80	324	0.691	0.691	0.182	0.090	0.054	11.308	2.773
20	12	83	103	252	0.631	0.631	0.161	0.061	0.046	12.019	1.751
21	15	70	66	370	0.721	0.721	0.128	0.052	0.055	5.425	0.906
22	11	97	39	309	0.590	0.590	0.167	0.072	0.080	4.383	0.804
23	17	133	86	315	0.678	0.678	0.094	0.047	0.049	3.710	0.916
24	15	74	54	266	0.464	0.464	0.154	0.085	0.067	5.266	1.593
25	14	100	68	403	0.791	0.791	0.204	0.047	0.047	18.433	1.001
26	14	79	89	530	0.832	0.832	0.148	0.040	0.043	11.807	0.844
27	12	71	56	236	0.546	0.546	0.122	0.068	0.066	3.422	1.046
28	11	213	58	201	0.851	0.851	0.185	0.044	0.050	13.904	0.795
29	24	173	110	307	0.452	0.452	0.100	0.066	0.047	4.597	1.999
30	14	90	71	187	0.735	0.735	0.211	0.049	0.052	16.337	0.870

Table D.2: Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 8

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	13	67	114	515	0.851	0.851	0.122	0.037	0.032	14.645	1.363
2	9	60	76	343	0.902	0.902	0.148	0.028	0.037	16.176	0.600
3	11	47	123	560	0.797	0.797	0.213	0.040	0.036	33.912	1.227
4	15	38	90	488	0.905	0.905	0.112	0.035	0.032	12.093	1.161
5	12	72	130	540	0.853	0.853	0.162	0.035	0.030	28.301	1.349
6	14	61	130	526	0.805	0.805	0.134	0.042	0.033	16.211	1.592
7	14	60	87	725	0.938	0.938	0.230	0.030	0.027	72.604	1.214
8	11	107	81	480	0.754	0.754	0.160	0.066	0.044	13.134	2.207
9	15	129	81	326	0.817	0.817	0.152	0.039	0.044	11.998	0.771
10	12	107	62	309	0.865	0.865	0.204	0.050	0.045	20.886	1.280
11	10	115	65	254	0.894	0.894	0.269	0.034	0.041	44.026	0.692
12	9	112	66	227	0.717	0.717	0.147	0.067	0.053	7.864	1.627
13	16	101	79	295	0.919	0.919	0.185	0.035	0.031	34.359	1.201
14	13	126	39	221	0.808	0.808	0.168	0.073	0.072	5.383	1.019
15	14	77	63	323	0.908	0.908	0.170	0.036	0.039	18.870	0.822
16	11	99	34	320	0.749	0.749	0.221	0.093	0.079	7.841	1.377
17	11	168	58	257	0.830	0.830	0.191	0.055	0.050	14.867	1.205
18	4	104	19	54	0.833	0.833	0.150	0.120	0.085	3.095	1.980
19	17	106	86	352	0.784	0.784	0.151	0.045	0.041	13.271	1.164
20	12	88	77	235	0.778	0.778	0.214	0.082	0.044	23.286	3.435
21	12	74	60	292	0.799	0.799	0.181	0.055	0.050	13.409	1.225
22	9	98	47	261	0.581	0.581	0.137	0.142	0.067	4.220	4.514
23	17	141	61	295	0.904	0.904	0.151	0.044	0.041	13.774	1.185
24	12	81	53	147	0.527	0.527	0.107	0.076	0.065	2.666	1.334
25	16	104	82	380	0.960	0.960	0.195	0.021	0.024	68.945	0.783
26	14	79	88	526	0.896	0.896	0.145	0.033	0.032	20.432	1.072
27	12	75	47	175	0.800	0.800	0.135	0.068	0.054	6.239	1.580
28	11	224	68	240	0.893	0.893	0.181	0.028	0.039	21.343	0.526
29	22	181	77	312	0.750	0.750	0.105	0.076	0.044	5.789	3.032
30	11	93	50	149	0.918	0.918	0.174	0.031	0.042	17.145	0.528

Table D.3: Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 12

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	12	64	92	472	0.650	0.650	0.148	0.068	0.049	9.261	1.934
2	11	57	71	479	0.590	0.590	0.120	0.073	0.058	4.354	1.599
3	11	47	131	562	0.464	0.464	0.090	0.063	0.042	4.573	2.247
4	13	38	70	422	0.730	0.730	0.141	0.071	0.055	6.638	1.651
5	12	67	115	602	0.478	0.478	0.094	0.051	0.046	4.202	1.242
6	14	62	98	535	0.612	0.612	0.126	0.072	0.050	6.493	2.114
7	13	57	93	812	0.676	0.676	0.150	0.059	0.049	9.476	1.474
8	12	103	74	472	0.569	0.569	0.102	0.080	0.057	3.208	1.978
9	18	131	92	346	0.380	0.380	0.084	0.086	0.051	2.676	2.791
10	9	104	48	220	0.667	0.667	0.225	0.082	0.066	11.714	1.568
11	13	116	89	298	0.557	0.557	0.191	0.095	0.052	13.269	3.315
12	10	108	53	231	0.519	0.519	0.141	0.074	0.067	4.425	1.232
13	12	100	74	235	0.556	0.556	0.127	0.087	0.057	4.930	2.313
14	14	127	40	279	0.709	0.709	0.114	0.073	0.071	2.568	1.067
15	12	77	68	381	0.496	0.496	0.113	0.044	0.060	3.539	0.536
16	11	97	34	236	0.419	0.419	0.207	0.159	0.084	6.101	3.592
17	12	170	51	309	0.508	0.508	0.130	0.074	0.070	3.512	1.117
18	5	100	15	45	0.278	0.278	0.136	0.137	0.132	1.056	1.068
19	18	107	71	375	0.506	0.506	0.172	0.112	0.059	8.588	3.622
20	15	85	95	330	0.444	0.444	0.131	0.051	0.050	6.833	1.050
21	12	74	58	251	0.435	0.435	0.109	0.069	0.065	2.824	1.152
22	9	98	44	273	0.529	0.529	0.194	0.055	0.075	6.647	0.528
23	16	141	58	265	0.569	0.569	0.181	0.106	0.065	7.713	2.624
24	16	84	61	224	0.383	0.383	0.088	0.068	0.062	2.030	1.204
25	17	102	75	375	0.680	0.680	0.156	0.084	0.053	8.719	2.568
26	11	80	80	398	0.616	0.616	0.143	0.068	0.054	6.965	1.599
27	10	75	47	152	0.381	0.381	0.128	0.068	0.069	3.411	0.957
28	10	227	50	209	0.526	0.526	0.159	0.104	0.071	5.009	2.157
29	16	181	78	262	0.338	0.338	0.094	0.058	0.056	2.836	1.079
30	10	95	51	140	0.484	0.484	0.163	0.085	0.068	5.698	1.546

Table D.4: Mean and Variance Estimates of Caries Free for Fluoridated Observations Age 15

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
1	12	45	97	982	0.364	0.364	0.108	0.081	0.048	4.985	2.778
2	10	17	74	981	0.376	0.376	0.062	0.065	0.056	1.228	1.346
3	10	19	147	1,102	0.340	0.340	0.048	0.040	0.038	1.614	1.133
4	6	18	54	554	0.464	0.464	0.116	0.063	0.067	3.013	0.879
5	11	32	126	1,299	0.313	0.313	0.064	0.058	0.041	2.372	1.970
6	11	27	122	918	0.441	0.441	0.084	0.057	0.044	3.682	1.738
7	10	32	78	1,157	0.314	0.314	0.062	0.069	0.055	1.237	1.567
8	13	30	103	1,083	0.304	0.304	0.069	0.046	0.044	2.426	1.089
9	9	26	79	696	0.316	0.316	0.075	0.052	0.052	2.093	1.025
10	8	25	55	718	0.451	0.451	0.166	0.122	0.063	6.964	3.746
11	9	18	60	778	0.320	0.320	0.078	0.084	0.061	1.662	1.889
12	7	30	43	744	0.369	0.369	0.110	0.087	0.072	2.333	1.472
13	7	19	42	637	0.484	0.484	0.133	0.092	0.075	3.162	1.506
14	8	21	26	767	0.296	0.296	0.084	0.083	0.095	0.799	0.765
15	9	17	65	1,090	0.312	0.312	0.058	0.054	0.058	1.022	0.893
16	12	20	35	967	0.128	0.128	0.063	0.062	0.064	0.962	0.937
17	10	24	40	1,143	0.374	0.374	0.096	0.085	0.074	1.676	1.300
18	7	23	40	500	0.394	0.394	0.164	0.113	0.072	5.116	2.419
19	8	26	55	615	0.227	0.227	0.084	0.074	0.058	2.056	1.593
20	10	16	137	926	0.322	0.322	0.049	0.044	0.038	1.635	1.291
21	8	20	77	813	0.203	0.203	0.069	0.057	0.047	2.162	1.474
22	7	20	32	674	0.315	0.315	0.118	0.102	0.083	2.059	1.524
23	6	29	30	435	0.194	0.194	0.074	0.088	0.090	0.676	0.948
24	10	17	55	796	0.147	0.147	0.047	0.047	0.055	0.715	0.737
25	7	34	70	544	0.363	0.363	0.069	0.038	0.057	1.470	0.450
26	8	25	79	913	0.311	0.311	0.099	0.069	0.053	3.413	1.689
27	6	12	50	380	0.268	0.268	0.069	0.064	0.063	1.194	1.038
28	3	48	15	349	0.333	0.333	0.076	0.041	0.126	0.365	0.107
29	9	28	73	730	0.268	0.268	0.079	0.064	0.052	2.278	1.492
30	6	10	94	502	0.249	0.249	0.051	0.039	0.043	1.456	0.827

Table D.5: Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 5

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	19	123	63	460	0.467	0.467	0.125	0.130	0.062	4.069	4.411
10	14	102	40	238	0.721	0.721	0.223	0.114	0.075	8.724	2.275
11	13	112	51	199	0.796	0.796	0.197	0.080	0.069	8.287	1.346
12	19	108	68	272	0.565	0.565	0.121	0.073	0.060	4.137	1.483
13	15	94	34	266	0.731	0.731	0.237	0.107	0.081	8.676	1.761
14	88	122	534	1,166	0.617	0.617	0.050	0.030	0.017	8.157	2.920
15	55	75	284	1,167	0.527	0.527	0.076	0.066	0.027	8.015	6.079
16	77	95	337	1,413	0.615	0.615	0.053	0.035	0.024	4.782	2.083
17	14	168	39	233	0.598	0.598	0.294	0.142	0.080	13.576	3.151
18	24	98	100	303	0.489	0.489	0.107	0.081	0.048	4.952	2.885
19	15	99	32	275	0.368	0.368	0.104	0.106	0.089	1.383	1.435
20	13	83	67	174	0.442	0.442	0.179	0.096	0.059	9.143	2.621
21	17	70	54	288	0.499	0.499	0.093	0.145	0.062	2.232	5.419
22	19	97	57	422	0.587	0.587	0.158	0.097	0.066	5.794	2.194
23	14	133	53	201	0.390	0.390	0.114	0.115	0.068	2.827	2.889
24	20	74	48	273	0.245	0.245	0.086	0.075	0.068	1.591	1.213
25	13	100	44	209	0.636	0.636	0.092	0.098	0.072	1.611	1.824
26	6	79	24	145	0.575	0.575	0.105	0.097	0.103	1.039	0.878
27	24	71	78	340	0.493	0.493	0.102	0.077	0.055	3.470	1.964
28	15	213	56	208	0.441	0.441	0.102	0.076	0.067	2.364	1.300
29	11	173	15	140	0.121	0.121	0.060	0.075	0.118	0.263	0.404
30	13	90	36	144	0.606	0.606	0.167	0.134	0.066	6.477	4.134

Table D.6: Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 8

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	17	129	54	306	0.859	0.859	0.177	0.061	0.056	9.855	1.179
10	11	107	21	138	0.896	0.896	0.142	0.062	0.078	3.358	0.630
11	8	115	11	123	0.943	0.943	0.220	0.044	0.122	3.284	0.130
12	14	112	71	213	0.718	0.718	0.121	0.078	0.051	5.612	2.367
13	9	101	23	168	0.725	0.725	0.181	0.124	0.105	2.954	1.388
14	17	126	74	247	0.870	0.870	0.156	0.047	0.041	14.662	1.349
15	12	77	54	189	0.837	0.837	0.137	0.072	0.045	9.162	2.552
16	12	99	28	234	0.887	0.887	0.111	0.058	0.067	2.763	0.746
17	14	168	47	255	0.816	0.816	0.245	0.098	0.060	16.789	2.668
18	20	104	59	247	0.866	0.866	0.103	0.044	0.044	5.485	1.020
19	15	106	30	225	0.639	0.639	0.098	0.175	0.074	1.790	5.627
20	9	88	20	121	0.870	0.870	0.200	0.088	0.081	6.047	1.186
21	8	74	14	117	0.850	0.850	0.205	0.110	0.096	4.528	1.291
22	12	98	19	304	0.898	0.898	0.240	0.079	0.096	6.269	0.676
23	16	141	46	219	0.884	0.884	0.142	0.063	0.056	6.462	1.254
24	13	81	41	160	0.734	0.734	0.130	0.071	0.066	3.795	1.150
25	13	104	49	194	0.811	0.811	0.131	0.106	0.047	7.885	5.169
26	8	79	20	278	0.896	0.896	0.199	0.064	0.109	3.334	0.345
27	22	75	76	308	0.794	0.794	0.101	0.052	0.052	3.709	1.007
28	6	224	20	98	0.887	0.887	0.236	0.110	0.069	11.841	2.573
29	5	181	7	90	0.800	0.800	0.287	0.202	0.143	4.047	1.995
30	6	93	9	89	0.678	0.678	0.135	0.200	0.147	0.847	1.863

Table D.7: Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 12

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	12	131	34	177	0.478	0.478	0.169	0.155	0.086	3.830	3.257
10	10	104	14	221	0.577	0.577	0.291	0.192	0.138	4.444	1.919
11	13	116	27	223	0.643	0.643	0.170	0.122	0.092	3.441	1.761
12	16	108	60	247	0.423	0.423	0.101	0.090	0.063	2.618	2.075
13	6	100	13	85	0.346	0.346	0.132	0.114	0.140	0.895	0.663
14	19	127	54	282	0.674	0.674	0.123	0.088	0.067	3.393	1.734
15	14	77	50	275	0.344	0.344	0.103	0.091	0.068	2.294	1.762
16	12	97	35	212	0.352	0.352	0.109	0.112	0.085	1.656	1.769
17	20	170	43	351	0.417	0.417	0.101	0.128	0.077	1.732	2.782
18	18	100	72	287	0.566	0.566	0.153	0.092	0.058	7.025	2.570
19	11	107	25	191	0.296	0.296	0.101	0.103	0.091	1.241	1.283
20	8	85	16	156	0.301	0.301	0.127	0.148	0.128	0.978	1.324
21	7	74	9	112	0.196	0.196	0.113	0.140	0.175	0.420	0.636
22	7	98	9	162	0.457	0.457	0.196	0.210	0.175	1.246	1.440
23	14	141	43	189	0.469	0.469	0.115	0.117	0.076	2.272	2.359
24	17	84	31	240	0.473	0.473	0.120	0.098	0.088	1.864	1.236
25	16	102	53	248	0.624	0.624	0.141	0.069	0.067	4.394	1.063
26	6	80	25	144	0.564	0.564	0.187	0.158	0.101	3.454	2.461
27	23	75	93	301	0.304	0.304	0.064	0.060	0.046	1.903	1.652
28	4	227	6	69	0.659	0.659	0.231	0.174	0.211	1.201	0.679
29	3	181	6	91	0.125	0.125	0.124	0.119	0.167	0.558	0.508
30	6	95	16	69	0.469	0.469	0.169	0.188	0.127	1.772	2.177

Table D.8: Mean and Variance Estimates of Caries Free for Non Fluoridated Observations Age 15

CCA	n	N	$\sum m_i$	$\sum M_i$	$\bar{y}_{unb}$	$\bar{y}_r$	SE($\bar{y}_{unb}$)	SE($\bar{y}_r$)	SE($\bar{y}_{srs}$)	DE($\bar{y}_{unb}$)	DE($\bar{y}_r$)
9	5	26	16	476	0.501	0.501	0.248	0.182	0.119	4.321	2.329
10	6	25	17	572	0.253	0.253	0.098	0.124	0.123	0.639	1.026
11	7	18	11	607	0.250	0.250	0.144	0.137	0.122	1.400	1.275
12	8	30	43	756	0.308	0.308	0.124	0.103	0.070	3.121	2.149
13	8	19	16	697	0.166	0.166	0.082	0.088	0.100	0.669	0.779
14	8	21	82	767	0.314	0.314	0.076	0.059	0.046	2.733	1.667
15	8	17	15	994	0.416	0.416	0.140	0.144	0.126	1.242	1.312
16	10	20	42	830	0.364	0.364	0.114	0.105	0.071	2.567	2.187
17	9	24	39	994	0.312	0.312	0.128	0.130	0.068	3.547	3.690
18	10	23	79	631	0.336	0.336	0.149	0.120	0.050	8.817	5.767
19	6	26	14	416	0.135	0.135	0.062	0.084	0.132	0.222	0.401
20	5	16	6	384	0.294	0.294	0.244	0.217	0.166	2.154	1.698
21	2	20	4	251	0.221	0.221	0.221	0.244	0.250	0.782	0.954
22	6	20	24	492	0.343	0.343	0.142	0.088	0.094	2.262	0.873
23	8	29	44	640	0.553	0.553	0.139	0.124	0.073	3.661	2.903
24	11	17	33	869	0.235	0.235	0.089	0.085	0.080	1.223	1.119
25	4	34	25	237	0.560	0.560	0.312	0.228	0.087	12.974	6.912
26	6	25	28	705	0.228	0.228	0.073	0.104	0.089	0.675	1.349
27	6	12	60	380	0.244	0.244	0.116	0.110	0.048	5.745	5.168
28	3	48	3	340	0	0	0	0	0		
29	3	28	3	276	0.355	0.355	0.336	0.329	0.333	1.014	0.974
30	3	10	8	210	0.287	0.287	0.236	0.202	0.163	2.101	1.534

Appendix E

Cluster Analysis Results

E.1 Fluoridated Results

E.1.1 Table Representation of Results

Table E.1: CCA DE values and assigned cluster indicator.

	DE_age5	DE_age8	DE_age12	DE_age15	cluster
Dun Laoghaire	1.306	0.918	2.243	2.555	5
Wicklow	0.807	0.521	1.675	0.852	4
Coolock	1.510	1.599	1.517	1.739	4
Kilbarrack	0.865	0.725	1.349	0.521	4
Roselawn	0.618	1.405	1.068	1.089	4
Cornmarket	1.221	1.043	1.683	2.044	4
Crumlin	1.435	1.003	0.830	1.840	4
Kildare	7.166	1.728	1.733	1.695	1
Laois/Offaly	0.841	0.565	3.812	2.872	5
Longford/Westmeath	0.658	1.302	1.719	1.985	4
Clare	0.391	0.811	2.667	1.676	5
Limerick	1.834	1.343	0.905	1.484	4
North Tipp	0.832	0.744	2.255	1.073	5
Cavan/Monaghan	1.557	1.084	0.809	0.875	4
Louth	1.340	0.737	0.762	1.295	4
Meath	2.683	0.533	3.435	1.109	5
Donegal	0.746	1.340	0.897	0.949	4
Sligo/Leitrim	0.338	1.980	0.404	3.451	3
Carlow/Kilkenny	1.557	3.139	2.708	1.450	5
South Tipp	1.383	1.926	1.294	1.198	4
Waterford	0.636	0.995	1.646	0.835	4
Wexford	0.333	1.932	0.337	3.177	3
Kerry	0.922	1.409	1.108	9.683	2
North Cork	1.574	1.168	0.895	0.972	4
North Lee	1.248	0.816	2.483	0.488	5
South Lee	1.003	0.878	0.856	3.787	3
West Cork	0.787	1.665	1.033	1.527	4
Galway	0.923	0.704	3.627	0.368	5
Mayo	1.869	3.060	0.677	2.028	3
Roscommon	1.094	0.488	0.160	1.031	4

E.1.2 Results of k-means (Outliers Removed)

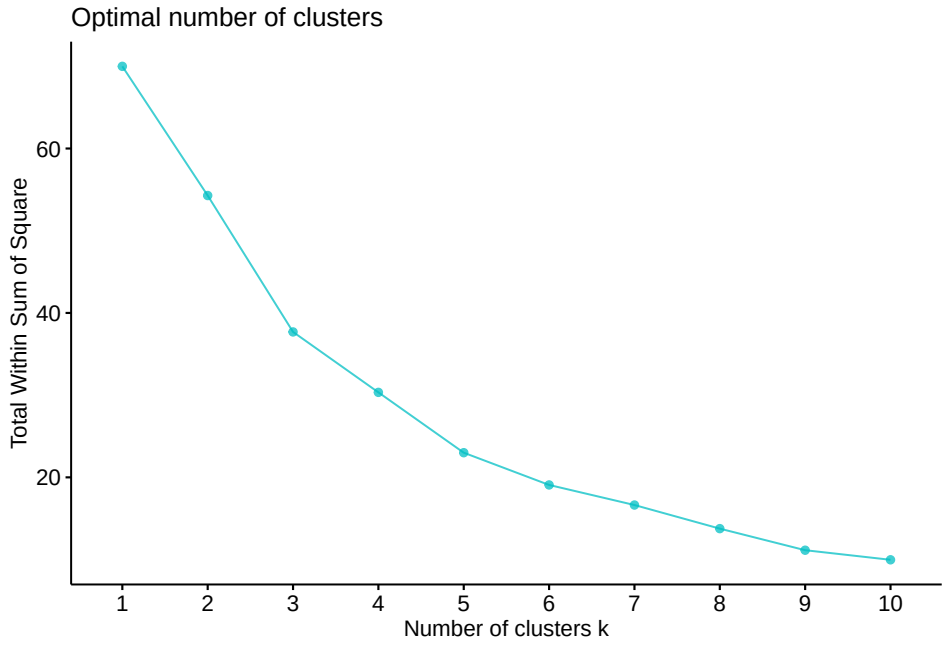


Figure E.1: Within Sum of Squares against Number of Clusters Used (Outliers Removed)

Observing Figure E.1, the number of clusters is again taken to be $k = 5$, however the choice is more difficult with outliers removed as the plot is closer to being linear. The resulting clusters formations are presented in the Figure E.2. While the removal of outliers has stopped the presence of a single CCA within a cluster, there are still some small clusters present. Furthermore, quite a lot of the CCA in the centre of the plot overlap, and the separation into distinct clusters is not obvious.

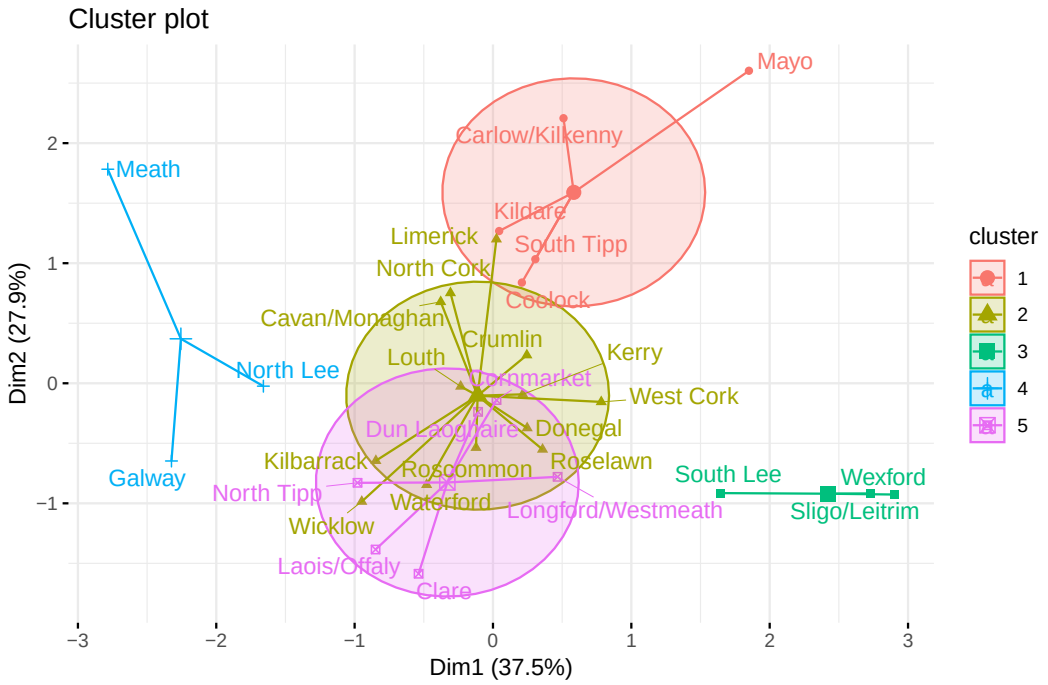


Figure E.2: Cluster Analysis of Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups (Outliers Removed)

E.1.3 Results of Agglomerative Clustering

Agglomerative Clustering starts with all clusters being individual and joins the closest two based on the chosen measure and continues this until all observations form one cluster. The higher the height of the fusion the less similar the objects are. This height is known as the cophenetic distance between two objects. Again, this was initially implemented on the raw dataset, not removing the extreme DE values. The results of the agglomerative analysis is shown in Figure E.3.

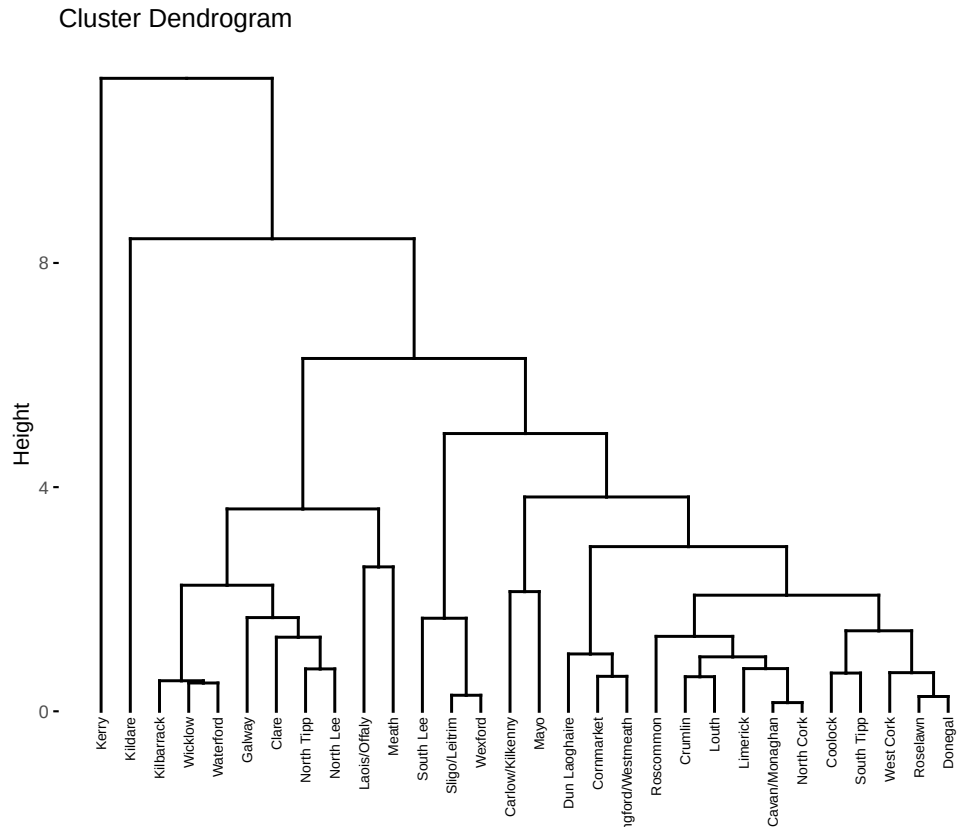


Figure E.3: Agglomerative Clustering Dendrogram (Ward Linkage)

One way to assess how well the cluster tree reflects your data is to compute the correlation between the cophenetic distances and the original distance data. If the clustering is valid, the correlation result will be strong ($\rho = 0.82$ in tree above Figure E.3). This was obtained using the Ward linkage method for clustering. The Ward linkage mininizes the total within-cluster variance. At each step the Ward method pairs the two clusters with minimum between cluster distance.

Another method of computing a dendrogram using a different distance measure is the average distance measure, and this is popular as it produces, in general, higher correlation between the original distances matrix and the cophenetic distance

matrix. The average linkage defines the distance between two clusters as the average distance between the elements in cluster 1 and the elements in cluster 2. The results of the average linkage method are shown in Figure E.4 and produce a correlation of 0.97 between the cophenetic distances and the original distance data.

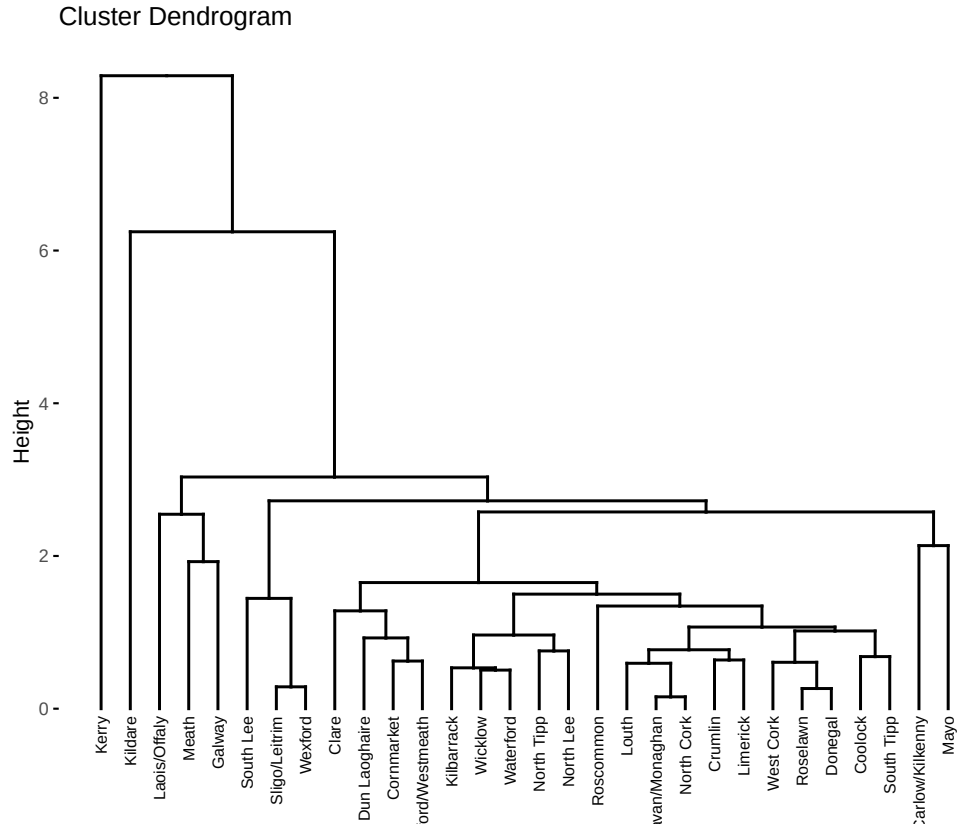


Figure E.4: Agglomerative Clustering Dendrogram (Average Distance Linkage)

While the Ward method is the default method in most settings, the average method shows much stronger results than the Ward method and thus the tree will be divided into groups based on the results of the average distance linkage. The tree is divided into $k = 5$ clusters for comparability to the earlier k-means cluster analysis results.

Figure E.5 and Figure E.6 show the resulting clusters from the agglomerative cluster method. Figure E.5 shows a coloured dendrogram according to the clusters, while Figure E.6 gives the clusters expressed in terms of their first two principal components.

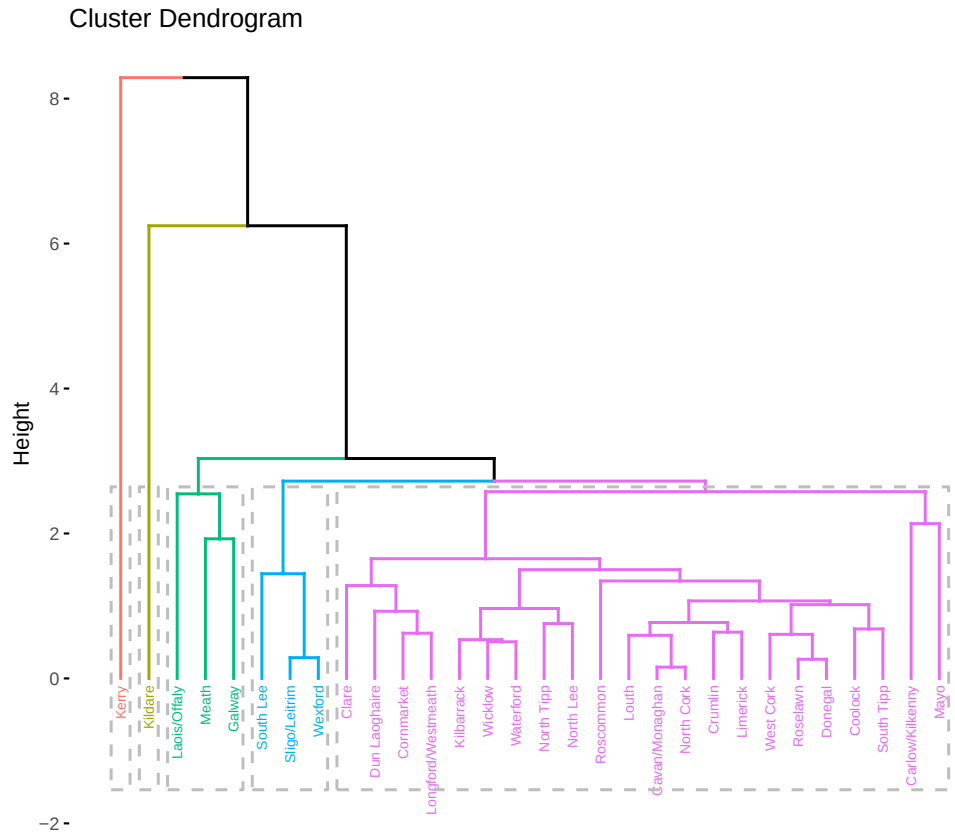


Figure E.5: Agglomerative Clustering Dendrogram (Average Distance Linkage) [k=5 Groups]

Again working with the raw data, Kerry and Kildare are assigned to individual clusters due to large DE present for one of their age groups. Two of the other clusters formed are quite small in size, with only three CCA's contained in each cluster. The majority of the CCA's ($n = 22$) fall into one large cluster. As in the case of k-means, removing the extreme DE values also yielded no useful groupings that might aid in identifying the occurrences of DE's less than one.

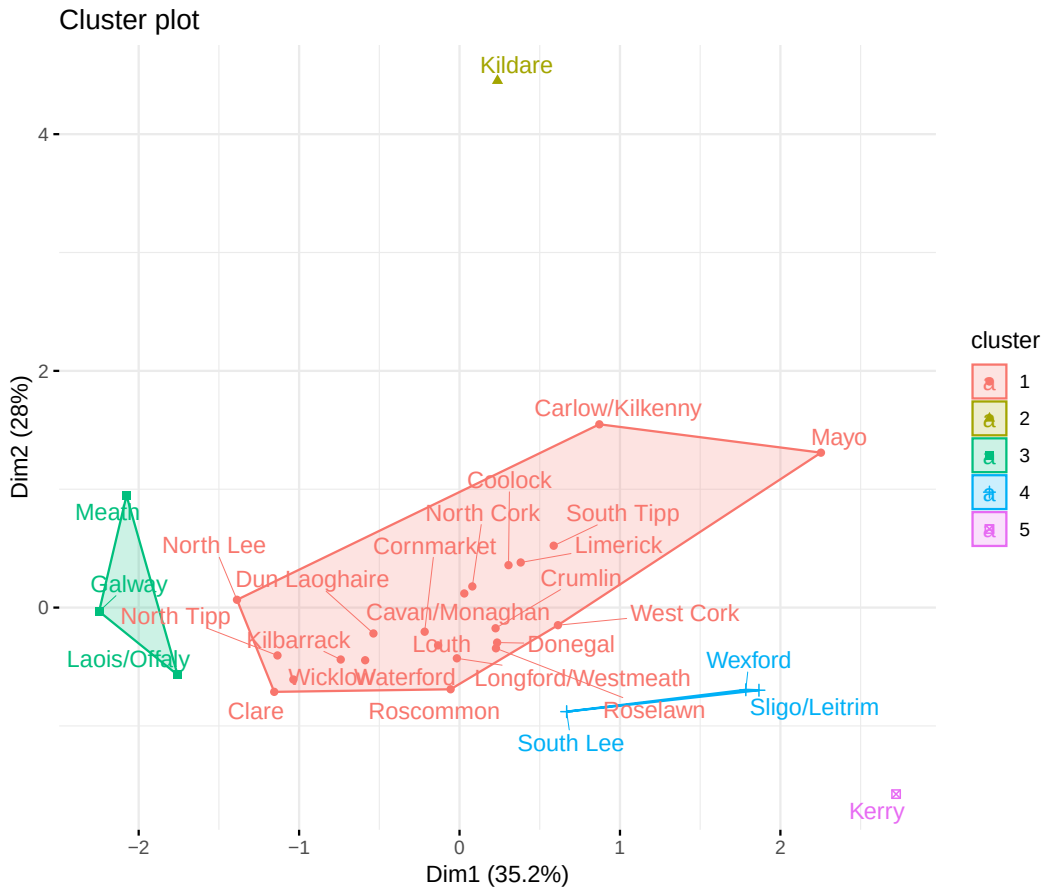


Figure E.6: Plot of Clusters by Principal Components One and Two for the Average Distance Linkage using Agglomerative Clustering

E.2 Non-Fluoridated Results

E.2.1 Results of k-means (Raw Data)

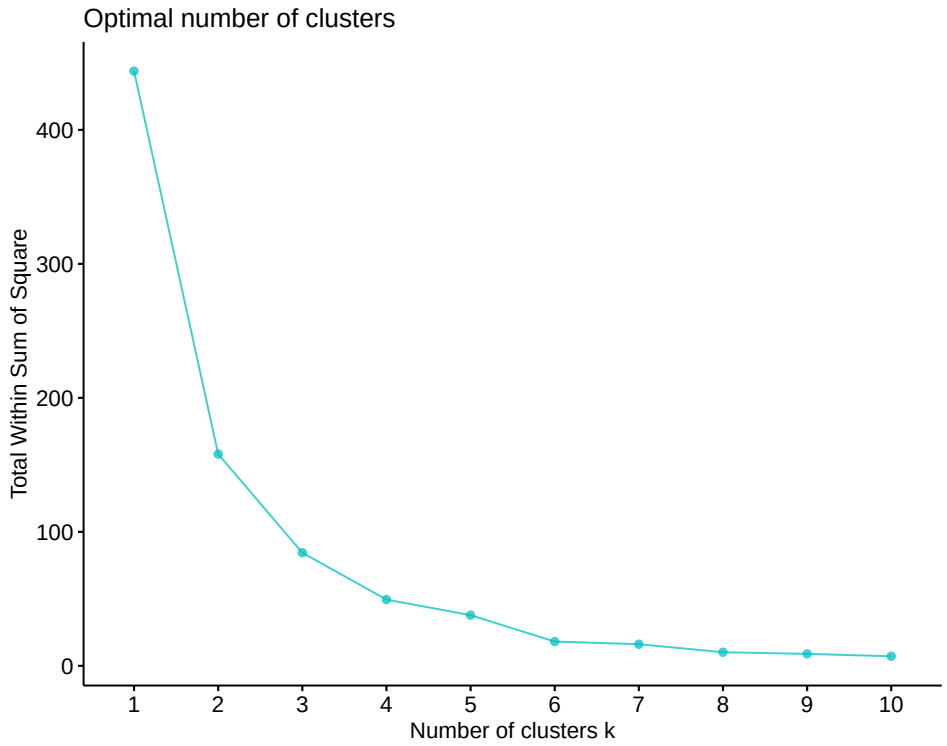


Figure E.7: Within Sum of Squares against Number of Clusters Used

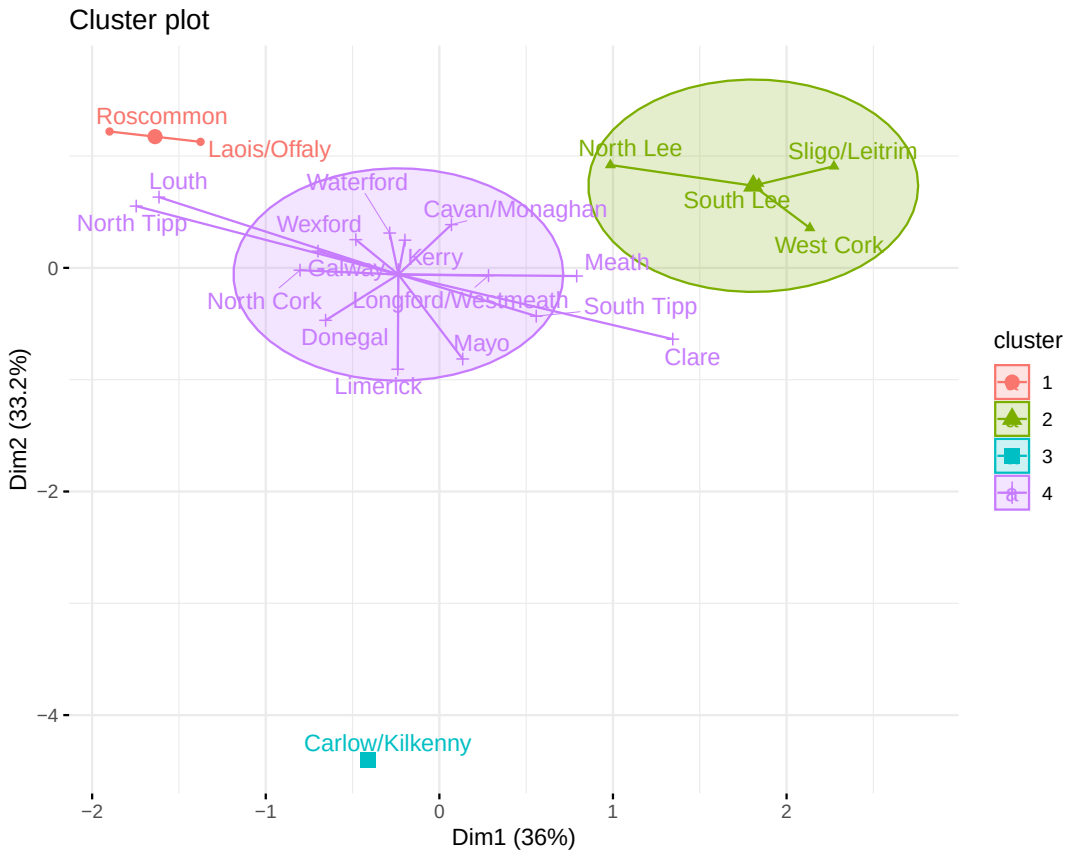


Figure E.8: Cluster Analysis of Non-Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups

E.2.2 Results of k-means (Outliers Removed)

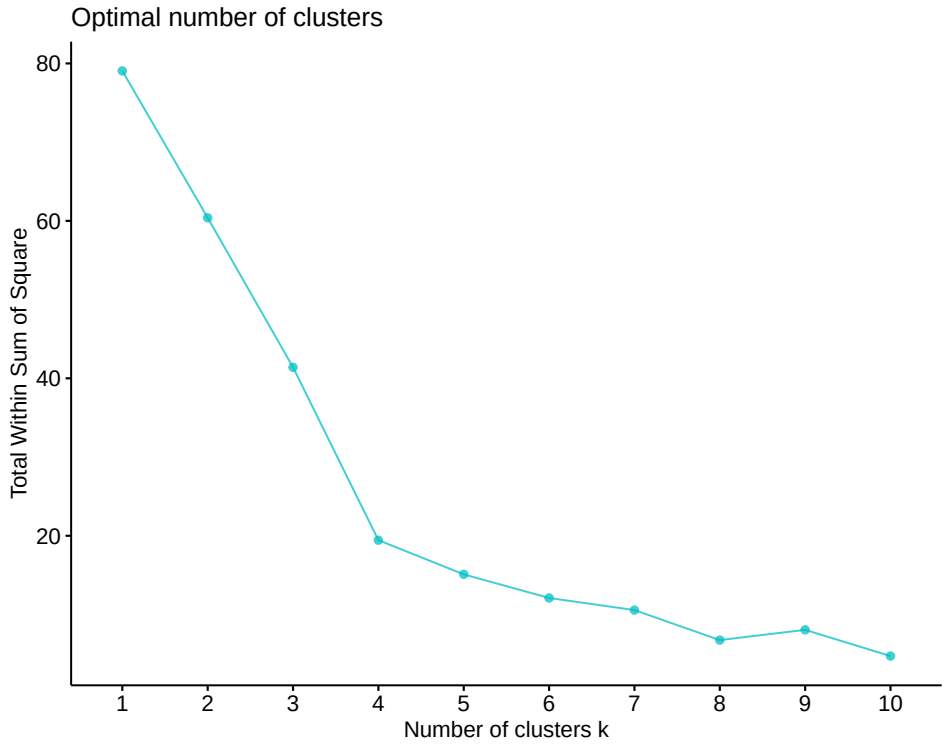


Figure E.9: Within Sum of Squares against Number of Clusters Used (Outliers Removed)

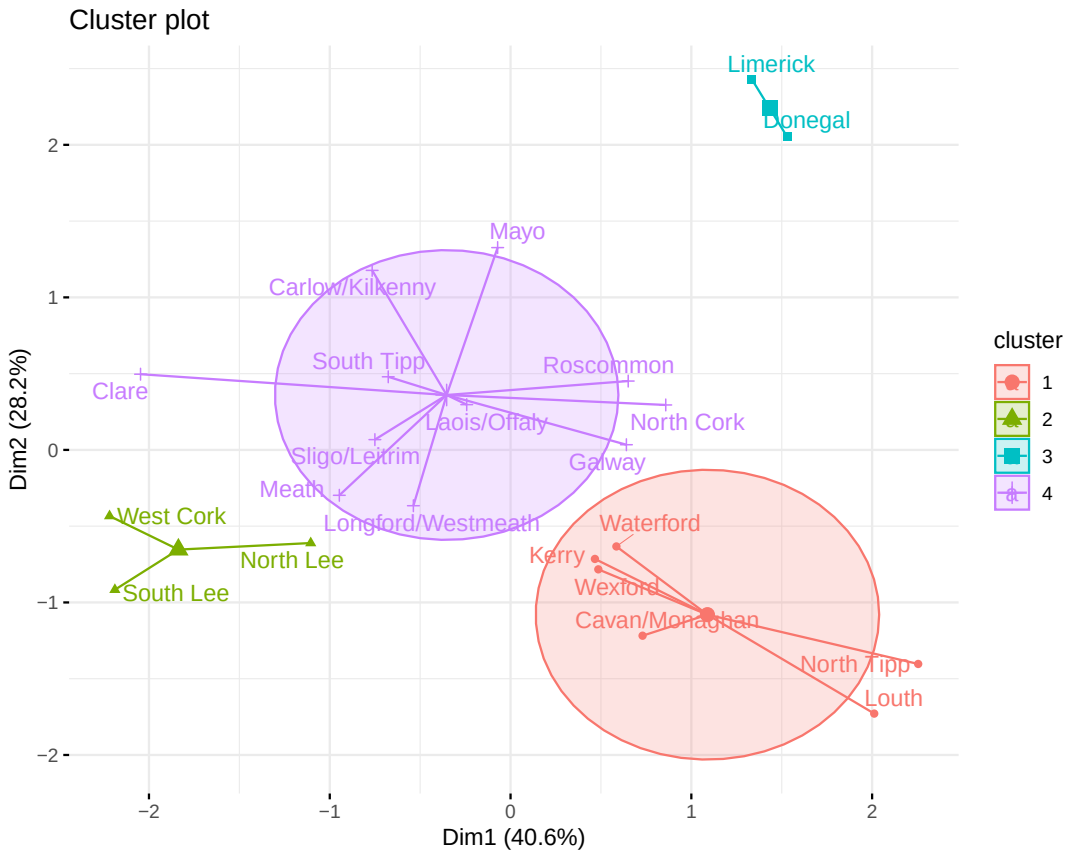


Figure E.10: Cluster Analysis of Non-Fluoridated dmft DE's under the Ratio Estimator Across all Age Groups (Outliers Removed)

E.2.3 Results of Agglomerative Clustering

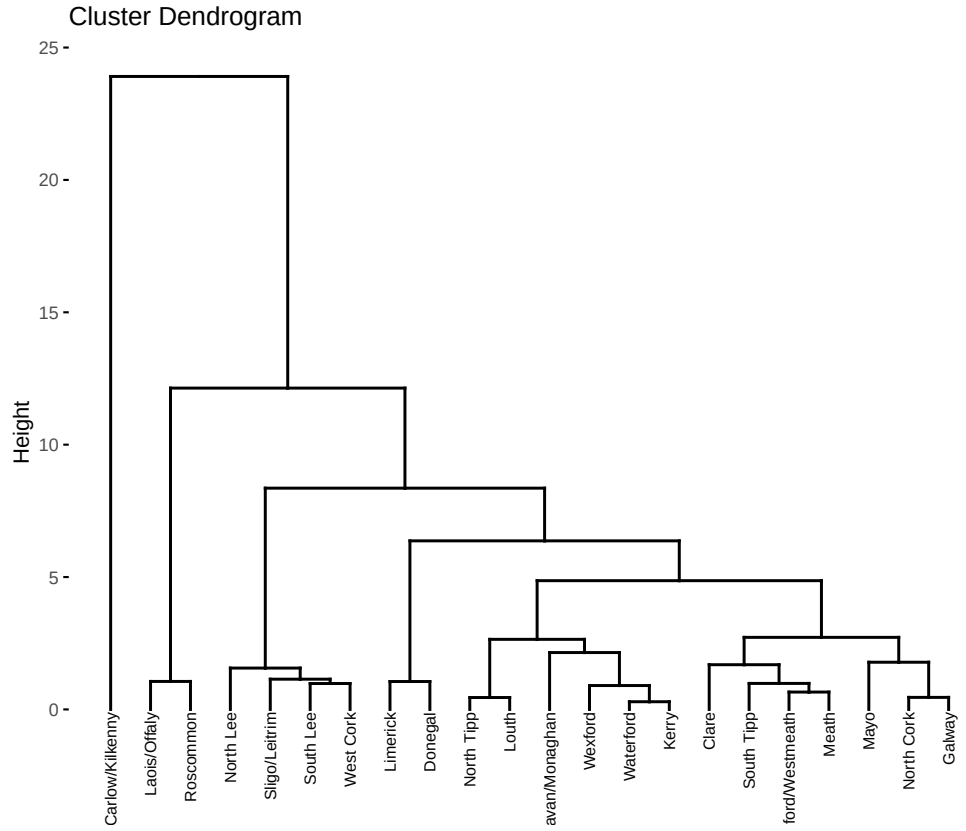


Figure E.11: Agglomerative Clustering Dendrogram (Ward Linkage)

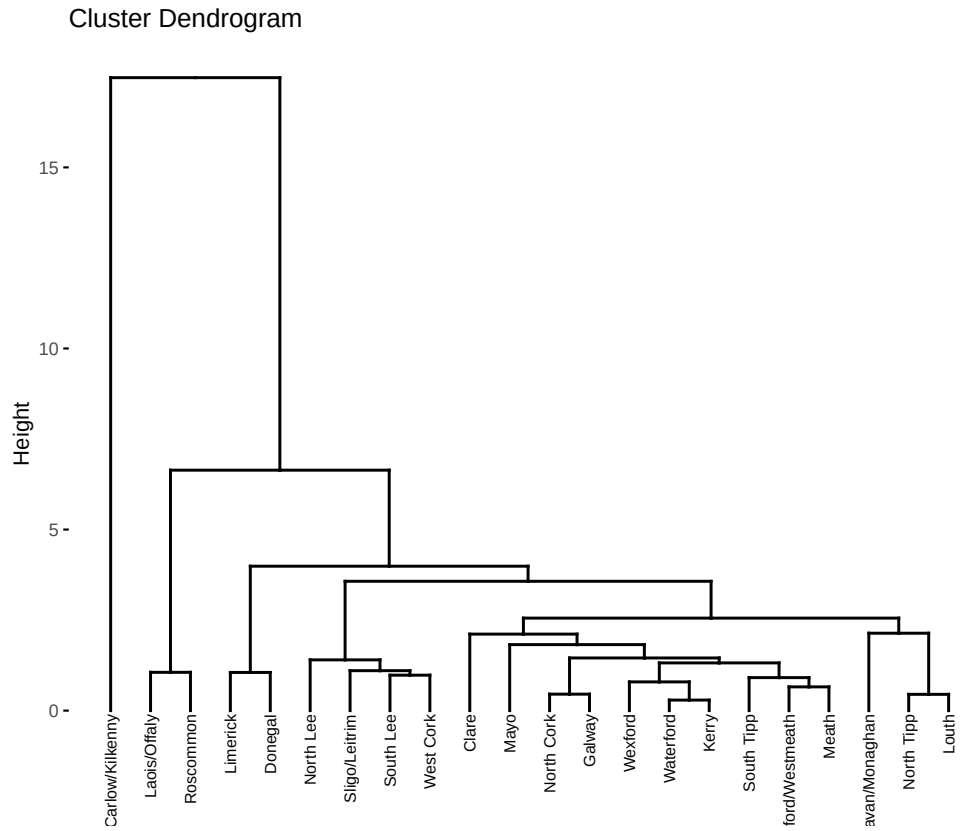


Figure E.12: Agglomerative Clustering Dendrogram (Average Distance Linkage)

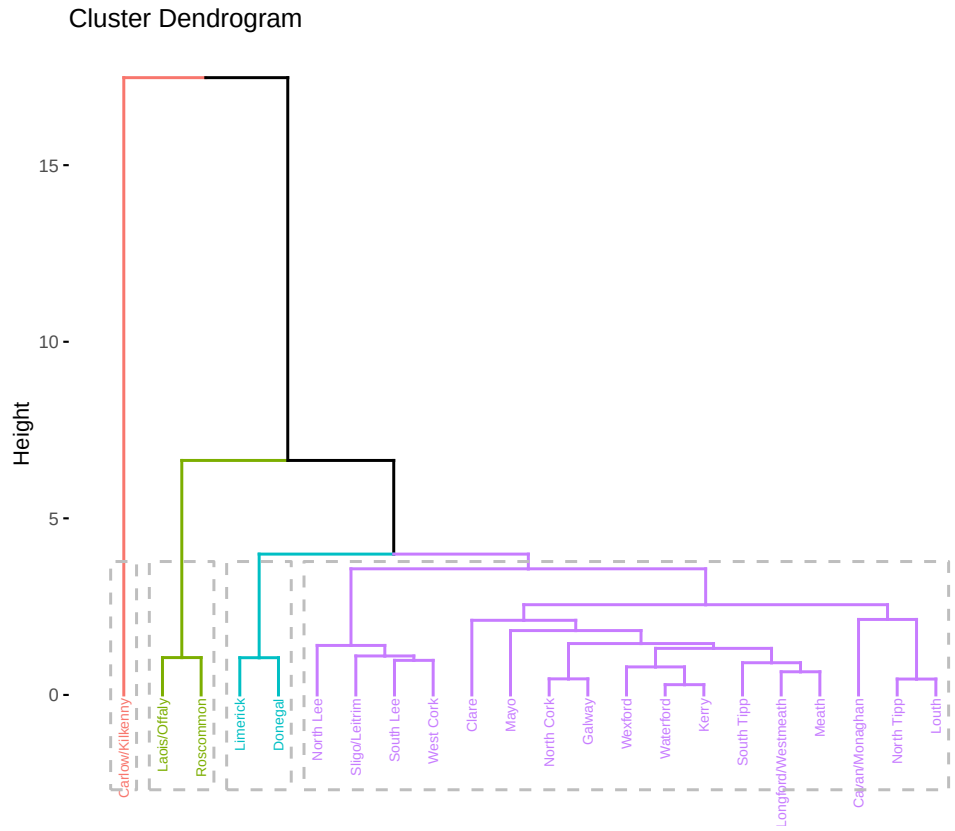


Figure E.13: Agglomerative Clustering Dendrogram (Average Distance Linkage) [k=4 Groups]

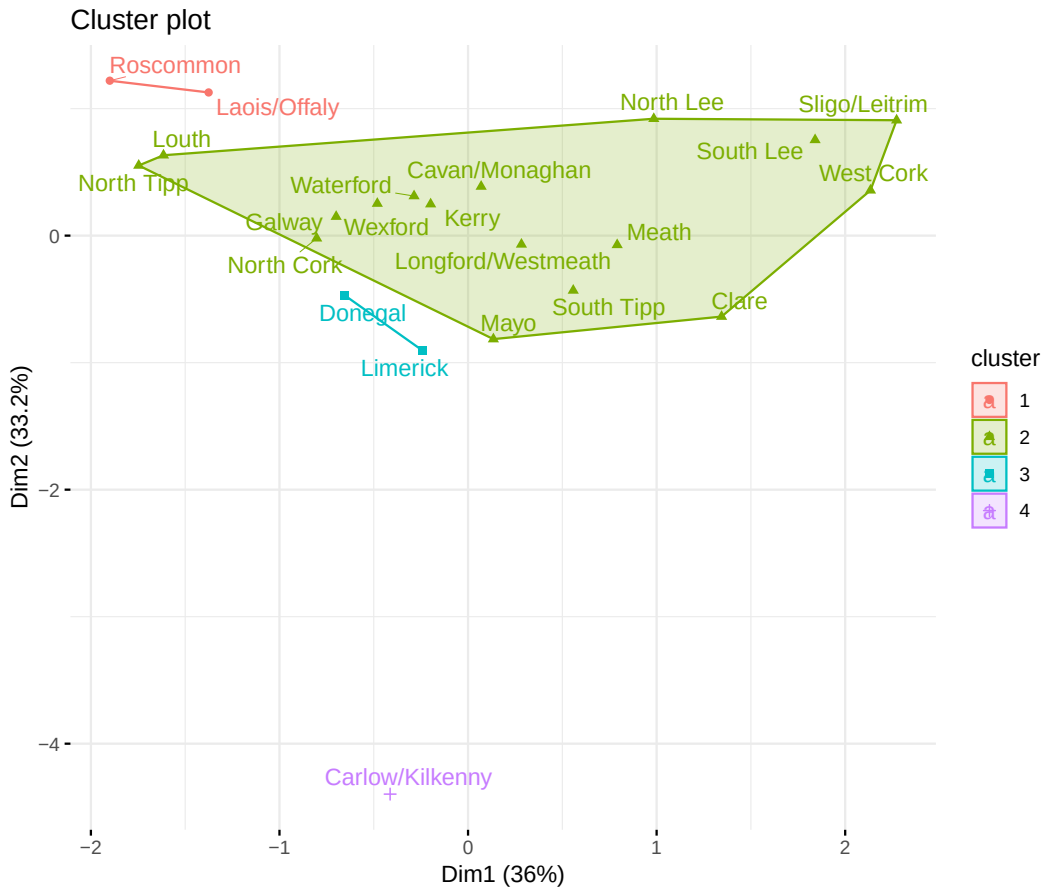


Figure E.14: Plot of Clusters by Principal Components One and Two for the Average Distance Linkage using Agglomerative Clustering

Appendix F

Linear Model Diagnostics

F.1 Linear Regression Diagnostics (dmft)

The first step will examine if there are any cases of high leverage in the regression model, and this is shown in Figure F.1.

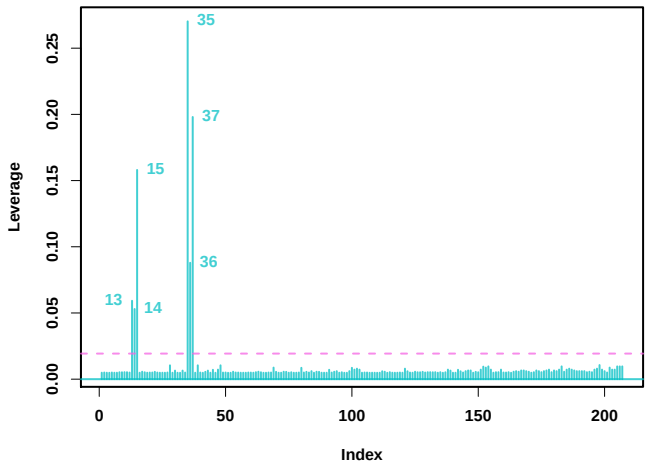


Figure F.1: Leverage values from model of $log(DE)$ on n.

Using the leverage cut off $(2 * p')/n$ there are 6 cases of high leverage present in the model. The respective case numbers are $14, 15, 16, 36, 37$ and 38. Cases of high leverage are usually cases which have an independent variable value that is far from the mean value of the independent variable. These cases are assessed as their omission from the model may provide a better model, if the justification

for removing them is valid. The next plot will identify any outlying cases in the dataset. These outliers will be shown in Figure F.2.

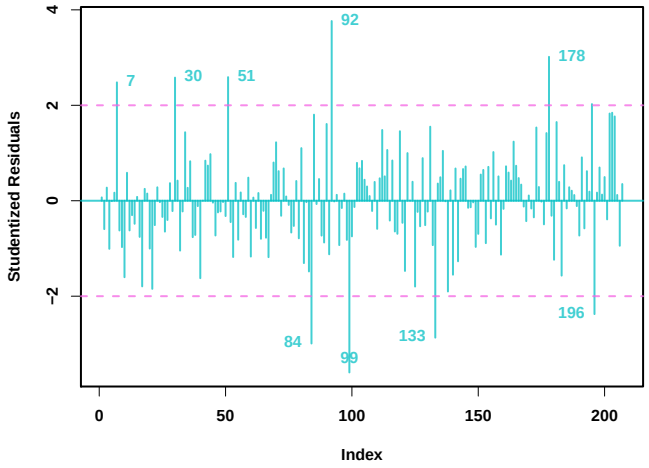


Figure F.2: Residual values from model of $log(DE)$ on n.

The plot of the residuals show there are nine cases that are classified as outliers (i.e. $r > 2$) where r is the standardised residuals from the model. The case numbers of the outliers are $8, 31, 52, 85, 93, 100, 134, 179$ and 197. There cases have a large difference between their observed values and expected values. Again, these cases need to be investigated to see if removing them may lead to a better regression model. The final plot (Figure F.3) is the plot of Cook's Distances and this is used to identify any cases of high influence in the regression model.

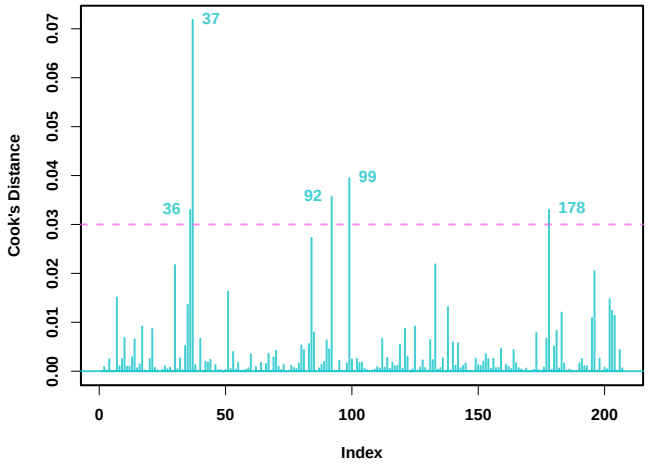


Figure F.3: Cook's distance values from model of $log(DE)$ on n.

Cases of high influence can have a large effect on the regression model. These cases are often outliers or cases with high leverage values. Again, Cook's Distances is used to assess if any of the cases may be omitted from the model due to some peculiarity in that particular case. Although Cook's Distance doesn't have a hard cut off value, the plot is usually inspected visually and in this plot five cases were shown to have large Cook's Distance values. These cases are $37, 38, 93, 100$ and 179.

Using the above diagnostic plots, cases $37, 38, 93, 100$ and 179 were investigated to see why they were identified in the above tests. Cases 37 and 38 were identified as their n values were both in the fourth quartile, while there was no obvious justification for cases 93, 100 and 179. The assumptions of the residuals, namely linearity, constant variance and normality were assessed visually, and the results are shown in Figures F.4, F.5, F.18 and F.7.

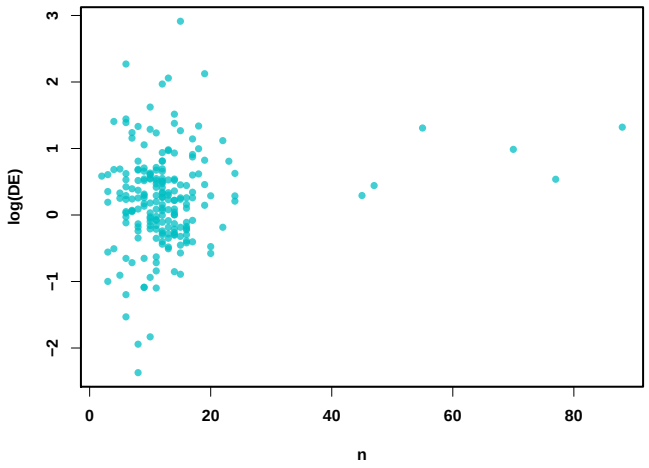


Figure F.4: Scatter plot of $log(DE)$ vs n for dmft.

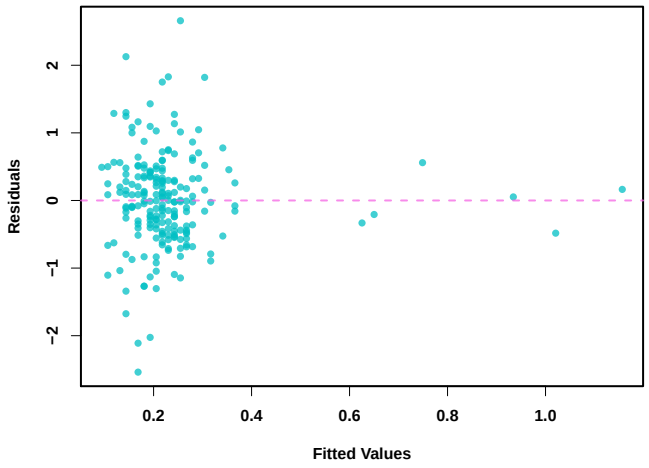


Figure F.5: Plot of residuals vs fitted values for model of $log(DE)$ on n for dmft.

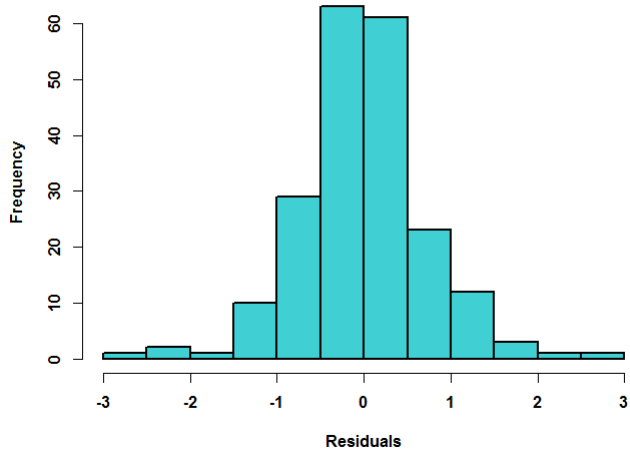


Figure F.6: Histogram of residuals for model of $log(DE)$ on n for dmft.

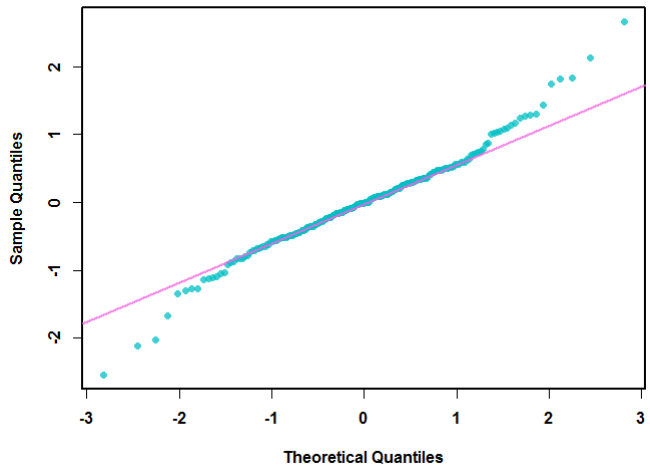


Figure F.7: Normal probability plot of residuals for model of $log(DE)$ on n for dmft.

Observing the above plots, the residuals have some slight deviations, however these are within a tolerable level and the residual assumptions are said to be satisfied.

F.2 Logistic Regression Diagnostics for dmft

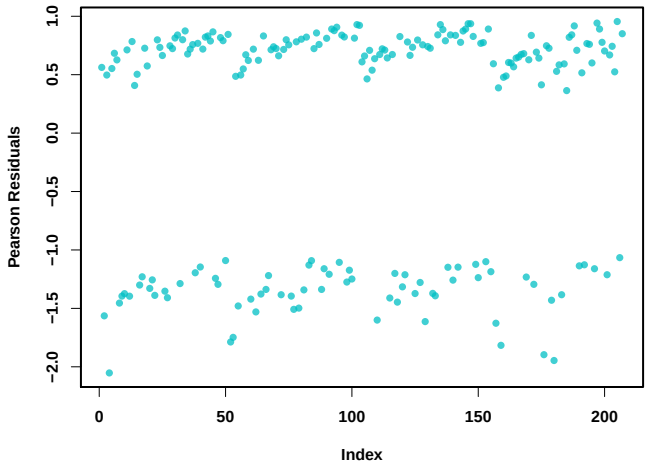


Figure F.8: Index plot of Pearson residuals for model $logit(p)$ on $\bar{m}$ for dmft

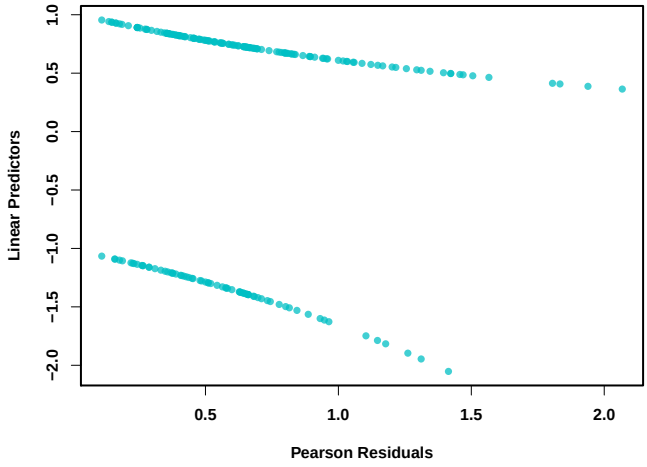


Figure F.9: Plot of Pearson Residuals vs Linear Predictor for model $logit(p)$ on $\bar{m}$ for dmft

The residual plots show no reason for concern in the logistic regression model.

F.3 Linear Regression Diagnostics (Caries Free)

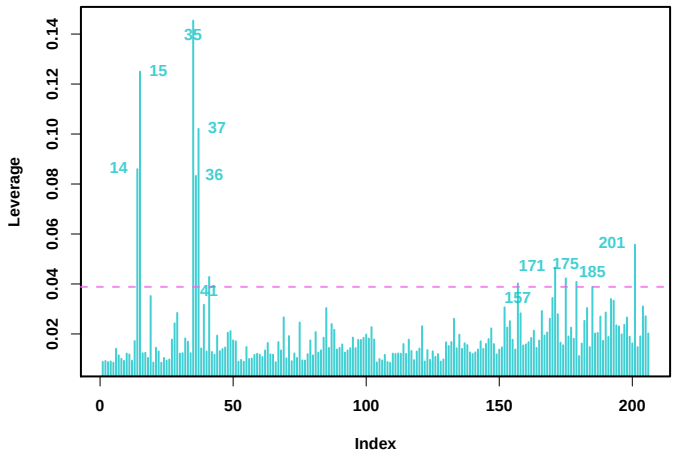


Figure F.10: Residual values from model of $log(DE)$ on Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{M}$ for caries free.

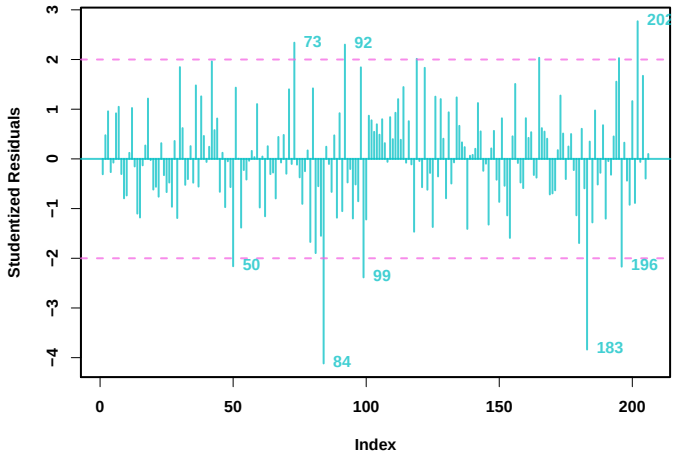


Figure F.11: Residual values from model of $log(DE)$ on Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{M}$ for caries free.

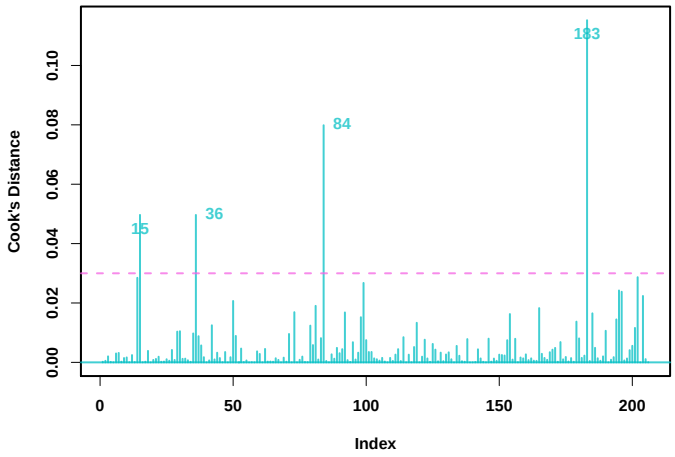


Figure F.12: Residual values from model of $log(DE)$ on Fluoridation Status, $\frac{n}{N}$ and $\frac{\bar{m}}{\bar{M}}$ for caries free.

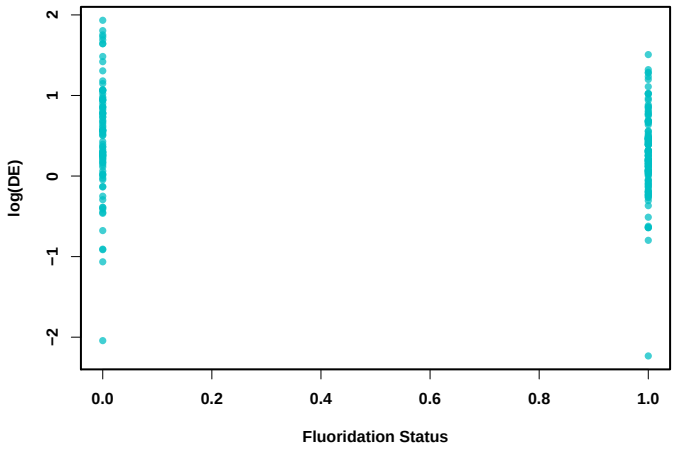


Figure F.13: Scatter plot of $log(DE)$ vs $FluoridationStatus$ for Caries Free.

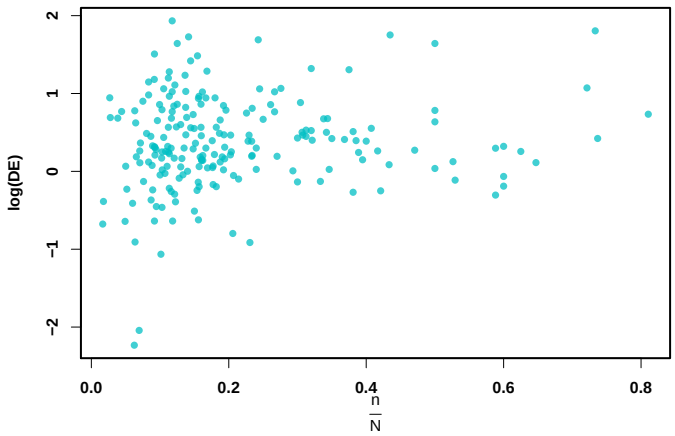


Figure F.14: Scatter plot of $log(DE)$ vs $\frac{n}{N}$ for Caries Free.

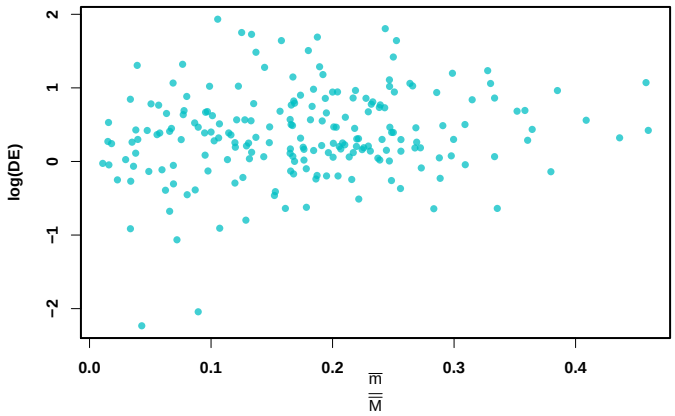


Figure F.15: Scatter plot of $log(DE)$ vs $\frac{\bar{m}}{\bar{M}}$ for Caries Free.

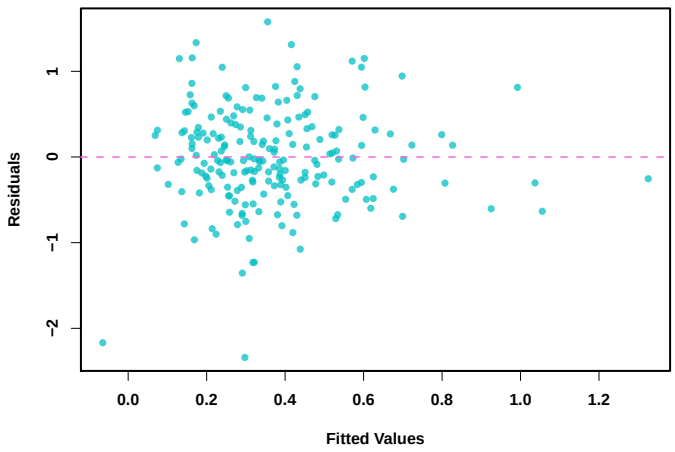


Figure F.16: Plot of residuals vs fitted values for model of $log(DE)$ on n for Caries Free.

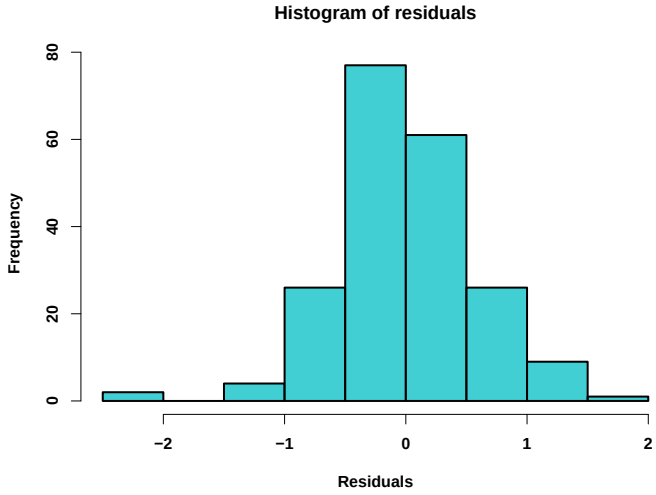


Figure F.17: Histogram of residuals for model of $log(DE)$ on n for Caries Free.

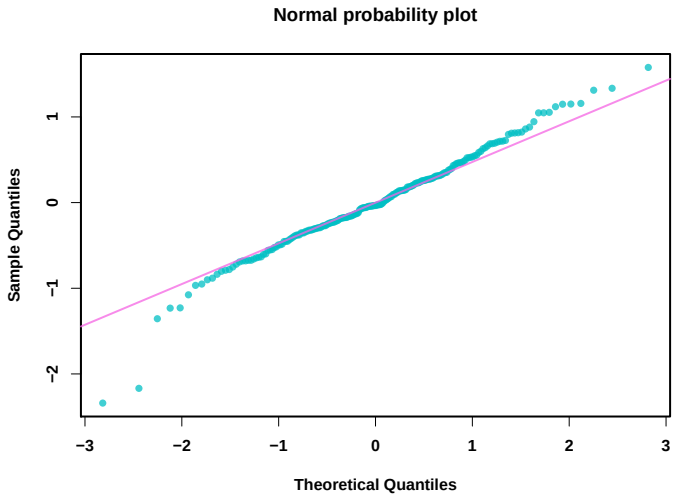


Figure F.18: Normal probability plot of residuals for model of $log(DE)$ on n for Caries Free.

Observing the above plots, the residuals have some slight deviations, however these are within a tolerable level and the residual assumptions are said to be satisfied.

F.4 Logistic Regression Diagnostics for Caries Free

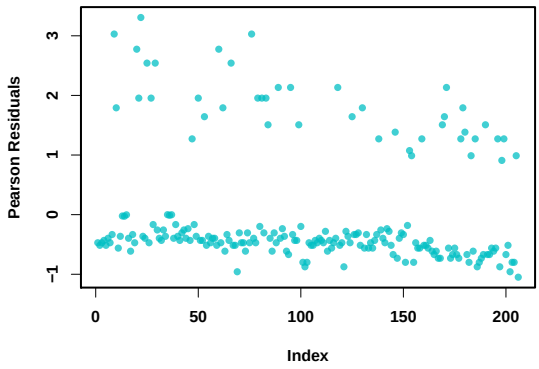


Figure F.19: Index plot of Pearson residuals for model $logit(p)$ on n for Caries Free

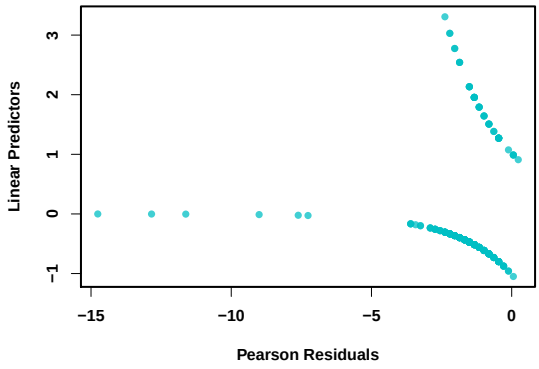


Figure F.20: Plot of Pearson Residuals vs Linear Predictor for model $logit(p)$ on n for Caries Free

The residual plots show no reason for concern in the logistic regression model.

Appendix G

Reducing Number of Clusters Samples (Caries Free)

Table G.1: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	5	10	16	23	31	42	55	55 (73)	55 (94)	55 (117)
CCA 2	3	7	11	16	21	28	28 (36)	28 (45)	28 (52)	28
CCA 3	3	7	11	15	19	24	30	30 (35)	30(41)	30(39)
CCA 4										
CCA 5	4	9	14	20	27	35	45	57	57 (71)	57 (89)
CCA 6	4	8	13	19	26	34	34 (45)	34 (59)	34 (78)	34
CCA 7	4	8	12	17	22	28	34	42	50	50 (59)
CCA 8	4	9	14	20	27	36	46	46 (58)	46 (71)	46 (74)
CCA 9	3	6	10	13	18	22	27	33	39	45
CCA 10	**3**	**7**	11	16	22	28	36	45	56	70
CCA 11	**5**	**10**	17	26	40	40 (59)	40 (91)	40 (144)	40	40
CCA 12	3	7	9	12	13	13 (13)	13 (10)	13 (5)	13	13
CCA 13	3	7	11	15	19	24	30	36	44	52
CCA 14	2	3	5	7	8*	10*	12*	14*	16*	18*
CCA 15	2	3	5	7	8*	10*	12*	14*	16*	18*
CCA 16	1	2	4	5*	6*	7*	9*	10*	11*	13*
CCA 17	4	8	13	18	25	32	40	50	62	62 (77)
CCA 18	6	12	20	28	28 (37)	28 (45)	28 (40)	28	28	28
CCA 19	3	7	11	16	20	25	31	37	43	50
CCA 20	5	10	16	23	30	38	47	47 (56)	47 (61)	47 (54)
CCA 21	**4**	8	13	19	25	33	41	52	65	81
CCA 22	**6**	12	20	30	41	56	56 (74)	56 (93)	56 (100)	56
CCA 23	**3**	6	10	14	19	24	30	37	45	54
CCA 24	3	6	10	14	18	23	28	35	43	54
CCA 25	4	9	15	21	29	39	50	64	82	82 (103)
CCA 26	**4**	**8**	12	18	24	31	39	49	61	61 (76)
CCA 27	5	11	17	26	35	47	63	63 (84)	63 (111)	63 (124)
CCA 28	**5**	**11**	17	25	34	45	45 (56)	45 (65)	45 (60)	45
CCA 29	2	4	7	9	12	15	18	21*	25*	29*
CCA 30	**4**	8	14	20	27	36	46	57	69	69 (80)
Col Avg	4	8	12	18	23	29	35	39	44	47
Col Max	6	12	20	30	41	56	63	64	82	82
Col 90th Percentile	5	11	17	26	35	43	51	57	63	69
Num CCAs	29	29	29	29	29	29	29	29	29	29

- **NUMBER** = DE< 1 vs original SRS SE even with NUMBER % increase in SE
- **CCA X** = CCA X has DE < 1 in original data
- NUM = $n < 5$ clusters left in CCA so estimate for n=5 (NUM) used to fill row
- (NUM) = actual estimate of SE with $n < 5$ clusters
- * = resampling used rather than taking all replications

Table G.2: Percentage Bias of Mean Caries Free ($\bar{p_r}$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-2)
CCA 2	0	0	-1	-1	-1	-2	-2 (-2)	-2 (-3)	-2 (-4)	-2
CCA 3	0	0	0	1	1	1	2	2 (3)	2 (4)	2 (7)
CCA 4										
CCA 5	0	0	0	0	0	0	-1	-1	-1 (-1)	-1 (-2)
CCA 6	0	0	0	-1	-1	-1	-1 (-2)	-1 (-2)	-1 (-4)	-1
CCA 7	0	0	0	0	0	0	0	0	0	0 (-1)
CCA 8	0	0	0	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-3)
CCA 9	0	0	0	0	1	1	1	1	2	2
CCA 10	**0**	**0**	0	0	0	0	0	0	0	0
CCA 11	**0**	**0**	-1	-1	-1	-1 (-2)	-1 (-3)	-1 (-4)	-1	-1
CCA 12	0	0	-1	-1	-2	-2 (-2)	-2 (-4)	-2 (-4)	-2	-2
CCA 13	0	0	0	-1	-1	-1	-2	-2	-3	-3
CCA 14	0	0	0	0	0*	0*	0*	0*	0*	0*
CCA 15	0	0	0	0	0*	0*	0*	0*	0*	0*
CCA 16	0	0	0	0*	0*	0*	0*	0*	0*	0*
CCA 17	0	0	0	0	0	0	1	1	1	1 (2)
CCA 18	0	0	0	-1	-1 (-1)	-1 (-1)	-1 (-2)	-1	-1	-1
CCA 19	0	0	0	0	0	-1	-1	-1	-1	-2
CCA 20	0	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-1)
CCA 21	**0**	0	0	0	0	0	1	1	1	1
CCA 22	**0**	0	0	0	1	1	1 (2)	1 (2)	1 (4)	1
CCA 23	**0**	0	0	0	0	0	0	0	1	1
CCA 24	0	0	-1	-1	-1	-2	-2	-3	-4	-5
CCA 25	0	0	0	0	0	0	0	-1	-1	-1 (-1)
CCA 26	**0**	**0**	0	0	0	0	0	0	-1	-1 (-1)
CCA 27	0	0	0	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-4)
CCA 28	**0**	**0**	0	0	0	0	0 (0)	0 (-1)	0 (-1)	0
CCA 29	0	0	0	0	0	0	0	0*	0*	0*
CCA 30	**0**	0	0	0	0	0	0	1	1	1 (2)
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	0	0	0	1	1	1	2	2	2	2
Col 90th Percentile	0	0	0	0	0	0	1	1	1	1
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.3: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	4	8	13	18	24	31	40	50	50 (63)	50 (77)
CCA 2	**5**	**12**	**19**	**28**	28 (39)	28 (49)	28 (47)	28	28	28
CCA 3	2	4	7	9	11	14	14 (20)	14 (31)	14 (43)	14
CCA 4	4	8	13	18	24	30	37	46	55	67
CCA 5	4	9	14	19	26	33	41	41 (51)	41 (64)	41 (81)
CCA 6	4	9	14	20	27	35	44	56	71	71 (89)
CCA 7	3	6	10	14	18	22	26	31	35	35 (39)
CCA 8	6	12	20	29	41	56	56 (76)	56 (103)	56 (130)	56
CCA 9	**4**	**8**	**13**	18	24	32	40	51	64	80
CCA 10	4	9	14	20	27	35	45	45 (58)	45 (75)	45 (105)
CCA 11	**6**	**13**	22	34	48	48 (66)	48 (88)	48 (105)	48	48
CCA 12	5	11	17	24	24 (32)	24 (39)	24 (36)	24	24	24
CCA 13	4	8	12	18	23	30	38	47	57	71
CCA 14	4	8	12	17	22	28	35	43	43 (50)	43 (57)
CCA 15	**4**	**8**	13	18	24	31	38	45	53	53 (58)
CCA 16	1	2	3	4	6	9	9 (12)	9 (15)	9 (15)	9
CCA 17	6	12	20	30	41	55	55 (71)	55 (88)	55 (93)	55
CCA 18	(-4)	(-16)								
CCA 19	3	6	10	14	19	24	30	38	47	57
CCA 20	4	9	14	20	26	34	43	43 (51)	43 (55)	43 (38)
CCA 21	4	9	15	21	28	36	45	45 (54)	45 (64)	45 (72)
CCA 22	2	4	6	9	9 (13)	9 (17)	9 (17)	9	9	9
CCA 23	3	7	10	14	19	24	30	36	43	50
CCA 24	5	10	16	23	31	40	52	52 (65)	52 (79)	52 (87)
CCA 25	**2**	**5**	**8**	**11**	14	17	20	24	28	32
CCA 26	4	8	12	18	24	31	39	49	62	62 (79)
CCA 27	4	9	15	22	29	38	48	48 (58)	48 (67)	48 (61)
CCA 28	**5**	**10**	**17**	**25**	**34**	46	46 (58)	46 (68)	46 (61)	46
CCA 29	2	5	7	10	13	16	19	23	27*	31*
CCA 30	**5**	**10**	**16**	**22**	**30**	38	38 (45)	38 (51)	38 (49)	38
Col Avg	4	8	13	19	25	31	36	39	43	45
Col Max	6	13	22	34	48	56	56	56	71	80
Col 90th Percentile	5	12	19	28	36	46	49	52	58	67
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.4: Percentage Bias of Mean Caries Free ($\bar{p_r}$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	0	0	0	0	0	0	0 (0)	0(0)
CCA 2	**0**	**0**	**0**	**0**	0 (0)	0 (0)	0 (1)	0	0	0
CCA 3	0	0	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-2)	-1
CCA 4	0	0	0	0	0	0	0	0	0	0
CCA 5	0	0	0	0	0	0	0	0 (0)	0 (0)	0(0)
CCA 6	0	0	0	0	0	0	0	0	0	0 (0)
CCA 7	0	0	0	0	0	0	0	0	0	0 (0)
CCA 8	0	0	0	0	-1	-1	-1 (-2)	-1 (-3)	-1 (-6)	-1
CCA 9	**0**	**0**	**0**	0	0	0	0	0	0	-1
CCA 10	0	0	0	0	0	0	-1	-1 (-1)	-1 (-1)	-1 (-2)
CCA 11	**0**	**0**	0	0	0	0 (-1)	0 (-1)	0 (-1)	0	0
CCA 12	0	0	0	0	0 (0)	0 (0)	0 (1)	0	0	0
CCA 13	0	0	0	0	0	0	0	0	-1	-1
CCA 14	0	0	0	0	0	0	0	0	0 (-1)	0 (-1)
CCA 15	**0**	**0**	0	0	0	0	0	0	0	0 (0)
CCA 16	0	1	1	1	2	2	2 (3)	2 (3)	2 (4)	2
CCA 17	0	0	0	0	0	0	0 (0)	0 (0)	0 (0)	0
CCA 18	(1)	(2)								
CCA 19	0	0	0	0	0	0	0	0	0	0
CCA 20	0	0	0	0	-1	-1	-1	-1 (-2)	-1 (-2)	-1 (-2)
CCA 21	0	0	0	0	0	1	1	1 (1)	1 (2)	1 (3)
CCA 22	1	1	2	3	3 (5)	3 (7)	3 (11)	3	3	3
CCA 23	0	0	0	0	0	0	0	0	0	0
CCA 24	0	0	0	0	0	0	0	0 (0)	0 (0)	0 (0)
CCA 25	**0**	**0**	**0**	**0**	0	0	0	0	0	0
CCA 26	0	0	0	0	0	0	0	0	-1	-1 (-1)
CCA 27	0	0	0	0	0	0	0	0 (0)	0 (0)	0 (0)
CCA 28	**0**	**0**	**0**	**0**	**0**	0	0 (0)	0 (0)	0 (0)	0
CCA 29	0	0	0	0	0	0	0	0	0*	0*
CCA 30	**0**	**0**	**0**	**0**	**0**	0	0 (0)	0 (0)	0 (1)	0
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	1	1	2	3	3	3	3	3	3	3
Col 90th Percentile	0	0	0	0	0	0	0	0	0	0
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.5: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	5	10	17	24	32	42	55	55 (70)	55 (91)	55 (118)
CCA 2	4	8	13	19	25	31	31 (38)	31 (43)	31 (39)	31
CCA 3	5	10	15	22	29	37	37 (46)	37 (56)	37 (71)	37
CCA 4	5	11	17	24	32	42	53	66	66 (82)	66 (99)
CCA 5	4	9	15	21	29	38	48	48 (62)	48 (80)	48 (112)
CCA 6	4	9	15	21	28	36	45	56	69	69 (84)
CCA 7	4	9	14	20	27	35	45	57	57 (71)	57 (88)
CCA 8	4	8	13	19	25	33	42	42 (53)	42 (66)	42 (76)
CCA 9	3	6	9	13	17	21	26	31	37	44
CCA 10	7	16	28	43	43 (64)	43 (93)	43 (126)	43	43	43
CCA 11	1	2	3	4	5	5	6	5	5 (4)	5 (0)
CCA 12	5	12	19	28	40	40 (53)	40 (70)	40 (84)	40	40
CCA 13	5	10	16	23	31	40	52	52 (65)	52 (79)	52 (81)
CCA 14	4	8	13	18	25	32	41	52	65	65 (80)
CCA 15	**3**	**6**	**9**	**14**	**19**	**25**	**32**	32 (42)	32 (55)	32 (73)
CCA 16	3	5	8	11	14	17	17 (20)	17 (20)	17 (12)	17
CCA 17	5	11	18	26	36	48	64	64 (84)	64 (110)	64 (140)
CCA 18	(9)	(16)	(12)							
CCA 19	2	4	6	8	10	13	16	18	22	25
CCA 20	4	8	13	19	26	34	43	55	69	88
CCA 21	5	10	16	23	31	41	53	53 (68)	53 (86)	53 (104)
CCA 22	**4**	**10**	**17**	**29**	29 (46)	29 (76)	29 (113)	29	29	29
CCA 23	2	4	7	9	12	15	18	22	26	30
CCA 24	4	8	12	17	23	29	36	44	54	66
CCA 25	2	5	7	10	13	16	19	22	26	31
CCA 26	5	10	16	22	30	38	38 (48)	38 (58)	38 (61)	38
CCA 27	4	8	12	18	23	23 (30)	23 (38)	23 (48)	23	23
CCA 28	5	11	17	25	34	34 (45)	34 (54)	34 (52)	34	34
CCA 29	3	7	12	16	22	28	35	43	52	63
CCA 30	6	14	23	34	47	47 (61)	47 (72)	47 (67)	47	47
Col Avg	4	9	14	20	26	31	37	40	43	45
Col Max	7	16	28	43	47	48	64	66	69	88
Col 90th Percentile	5	11	18	29	37	43	53	56	65	66
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.6: Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	0	0	-1	-1	-1	-1 (-2)	-1 (-2)	-1 (-4)
CCA 2	0	0	0	0	0	1	1 (1)	1 (2)	1 (3)	1
CCA 3	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-1)	0
CCA 4	0	0	0	0	0	-1	-1	-1	-1 (-2)	-1 (-3)
CCA 5	0	0	0	0	0	0	0	0 (0)	0 (0)	0 (0)
CCA 6	0	0	0	0	-1	-1	-1	-1	-2	-2 (-3)
CCA 7	0	0	0	0	-1	-1	-1	-1	-1 (-2)	-1 (-3)
CCA 8	0	0	0	1	1	1	2	2 (2)	2 (3)	2 (5)
CCA 9	0	0	0	0	1	1	1	1	1	2
CCA 10	0	-1	-1	-2	-2 (-3)	-2 (-4)	-2 (-7)	-2	-2	-2
CCA 11	0	-1	-1	-1	-2	-2	-3	-4	-4 (-5)	-4 (-6)
CCA 12	0	0	0	0	0	0 (0)	0 (-1)	0 (-1)	0	0
CCA 13	0	0	0	0	1	1	1	1 (1)	1 (1)	1 (0)
CCA 14	0	0	0	0	0	0	0	0	0	0 (1)
CCA 15	**0**	**0**	**0**	**0**	**0**	**-1**	**-1**	-1 (-1)	-1 (-2)	-1 (-3)
CCA 16	0	-1	-2	-2	-3	-5	-5 (-6)	-5 (-8)	-5 (-12)	-5
CCA 17	0	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-2)
CCA 18	(0)	(1)	(4)							
CCA 19	0	0	-1	-1	-1	-1	-2	-2	-3	-3
CCA 20	0	0	0	0	0	-1	-1	-1	-1	-1
CCA 21	0	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-3)
CCA 22	**0**	**-1**	**-1**	**-1**	-1 (-2)	-1 (-3)	-1 (-5)	-1	-1	-1
CCA 23	0	0	-1	-1	-1	-2	-2	-2	-3	-4
CCA 24	0	0	0	0	0	0	0	0	-1	-1
CCA 25	0	0	0	0	0	1	1	1	1	1
CCA 26	0	0	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1
CCA 27	0	-1	-1	-2	-3	-3 (-4)	-3 (-6)	-3 (-10)	-3	-3
CCA 28	0	0	-1	-1	-1	-1 (-2)	-1 (-4)	-1 (-7)	-1	-1
CCA 29	0	0	0	0	-1	-1	-1	-1	-2	-2
CCA 30	0	-1	-1	-2	-3	-3 (-4)	-3 (-7)	-3 (-11)	-3	-3
Col Avg	0	0	0	0	-1	-1	-1	-1	-1	-1
Col Max	0	0	0	1	1	1	2	2	2	2
Col 90th Percentile	0	0	0	0	0	1	1	1	1	1
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.7: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	5	10	15	21	28	35	43	43 (52)	4359 v	43 (52)
CCA 2	8	18	29	42	59	59 (82)	59 (113)	59 (158)	59	59
CCA 3	5	11	18	27	38	38 (52)	38 (72)	38 (95)	38	38
CCA 4	**5**	5 (10)	5 (14)	5 (13)	5	5	5	5	5	5
CCA 5	4	9	14	19	25	30	30 (36)	30 (39)	30 (31)	30
CCA 6	5	10	16	23	32	42	42 (54)	42 (69)	42 (82)	42
CCA 7	5	11	18	26	35	35 (44)	35 (55)	35 (57)	35	35
CCA 8	4	7	12	17	23	30	39	49	49 (61)	49 (73)
CCA 9	3	7	12	16	16 (22)	16 (27)	16 (26)	16	16	16
CCA 10	5	9	12	12 (14)	12 (14)	12 (12)	12	12	12	12
CCA 11	5	11	18	26	26 (35)	26 (44)	26 (47)	26	26	26
CCA 12	4	8	8 (15)	8 (23)	8 (37)	8	8	8	8	8
CCA 13	5	11	11 (19)	11 (28)	11 (34)	11	11	11	11	11
CCA 14	**6**	**14**	24	24 (37)	24 (56)	24 (77)	24	24	24	24
CCA 15	**4**	10	17	26	26 (38)	26 (54)	26 (69)	26	26	26
CCA 16	5	10	16	22	29	35	42	42 (48)	42 (49)	42 (39)
CCA 17	4	9	15	23	33	33 (48)	33 (68)	33 (92)	33	33
CCA 18	4	9	9 (14)	9 (21)	9 (31)	9	9	9	9	9
CCA 19	3	7	11	11 (15)	11 (20)	11 (26)	11	11	11	11
CCA 20	6	12	20	30	42	42 (58)	42 (80)	42 (106)	42	42
CCA 21	4	9	15	15 (21)	15 (28)	15 (36)	15	15	15	15
CCA 22	5	11	11 (17)	11 (20)	11 (14)	11	11	11	11	11
CCA 23	10	10 (20)	10 (29)	10 (32)	10	10	10	10	10	10
CCA 24	**6**	**13**	21	31	43	43 (58)	43 (77)	43 (96)	43	43
CCA 25	**5**	**11**	**11** (19)	**11** (31)	**11** (38)	11	11	11	11	11
CCA 26	5	11	19	19 (29)	19 (42)	19 (57)	19	19	19	19
CCA 27	10	10 (22)	10 (39)	10 (58)	10	10	10	10	10	10
CCA 28	(-2)									
CCA 29	3	8	13	21	21 (31)	21 (45)	21 65()	21	21	21
CCA 30	**6**	6 (13)	6 (23)	6 (38)	6	6	6	6	6	6
Col Avg	5	10	14	18	22	23	24	24	24	24
Col Max	10	18	29	42	59	59	59	59	59	59
Col 90th Percentile	7	12	20	27	39	42	42	43	43	43
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.8: Percentage Bias of Mean Caries Free ($\bar{p}_r$) by Reduction in the Number of Schools Selected (by CCA) for Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 1	0	0	-1	-1	-1	-2	-2	-2 (-3)	-2 (-4)	-2 (-6)
CCA 2	0	0	0	1	1	1 (1)	1 (2)	1 (4)	1	1
CCA 3	0	0	0	0	0	0 (0)	0 (0)	0 (0)	0	0
CCA 4	**0**	0 (-1)	0 (-1)	0 (-3)	0	0	0	0	0	0
CCA 5	0	0	0	1	1	1	1 (2)	1 (2)	1 (3)	1
CCA 6	0	0	0	0	0	0	0 (-1)	0 (-1)	0 (-2)	0
CCA 7	0	0	1	1	2	2 (3)	2 (5)	2 (9)	2	2
CCA 8	0	0	0	-1	-1	-1	-2	-2	-2 (-3)	-2 (-5)
CCA 9	0	0	0	0	0 (0)	0 (0)	0 (0)	0	0	0
CCA 10	0	-1	-1	-1 (-2)	-1 (-4)	-1 (-8)	-1	-1	-1	-1
CCA 11	0	1	1	1	1 (2)	1 (4)	1 (6)	1	1	1
CCA 12	0	0	0 (-1)	0 (-1)	0 (-3)	0	0	0	0	0
CCA 13	0	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1	-1	-1
CCA 14	**0**	**0**	0	0 (1)	0 (1)	0 (1)	0	0	0	0
CCA 15	**0**	0	0	1	1 (1)	1 (2)	1 (3)	1	1	1
CCA 16	0	0	0	1	1	1	1	1 (2)	1 (2)	1 (2)
CCA 17	0	0	0	0	0	0 (0)	0 (1)	0 (2)	0	0
CCA 18	-1	-2	-2 (-3)	-2 (-4)	-2 (-4)	-2	-2	-2	-2	-2
CCA 19	0	0	0	0 (0)	0 (1)	0 (2)	0	0	0	0
CCA 20	0	0	0	0	0	0 (0)	0 (0)	0 (0)	0	0
CCA 21	0	-1	-1	-1 (-2)	-1 (-2)	-1 (-4)	-1	-1	-1	-1
CCA 22	0	0	0 (-1)	0 (-1)	0 (-2)	0	0	0	0	0
CCA 23	1	1 (3)	1 (7)	1 (16)	1	1	1	1	1	1
CCA 24	**0**	**0**	0	1	1	1 (1)	1 (2)	1 (4)	1	1
CCA 25	**0**	**0**	**0 (0)**	**0 (0)**	**0 (-1)**	0	0	0	0	0
CCA 26	0	-1	-1	-1 (-2)	-1 (-3)	-1 (-4)	-1	-1	-1	-1
CCA 27	0	0 (0)	0 (0)	0 (0)	0	0	0	0	0	0
CCA 28	(0)									
CCA 29	0	0	-1	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1
CCA 30	**0**	0 (0)	0 (-1)	0 (-1)	0	0	0	0	0	0
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	1	1	1	1	2	2	2	2	2	2
Col 90th Percentile	0	0	0	1	1	1	1	1	1	1
Num CCAs	29	29	29	29	29	29	29	29	29	29

Table G.9: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	3	5	8	10	13	16	19	21	24	27
CCA 10	3	6	10	14	18	22	27	33	39	39 (45)
CCA 11	4	9	15	21	27	35	44	54	54 (65)	54 (77)
CCA 12	3	6	9	13	17	21	26	32	38	45
CCA 13	3	7	11	15	19	24	30	36	42	50
CCA 14	1	1	2	3*	3*	4*	5*	5*	6*	7*
CCA 15	0	1	1	1	1*	2*	2*	2*	3*	3*
CCA 16	1	3	4	5*	7*	8*	10*	11*	12*	14*
CCA 17	0	0	0	-1	-1	-1	-1	-1	0	0 (1)
CCA 18	2	5	8	11	14	17	21	24*	28*	32*
CCA 19	3	7	11	15	20	25	31	38	46	55
CCA 20	1	3	4	7	10	13	18	23	23 (30)	23 (37)
CCA 21	1	2	3	4	5	6	7	8	10	11
CCA 22	2	5	7	10	13	17	20	25	29	35
CCA 23	4	7	11	16	21	26	32	38	44	44 (50)
CCA 24	3	5	8	12	15	19	24	29	34	41
CCA 25	3	6	10	13	18	22	28	34	34 (42)	34 (53)
CCA 26	**6**	6 (13)	6 (21)	6 (22)	6	6	6	6	6	6
CCA 27	3	5	8	11	14	18	22	26*	30*	35*
CCA 28	3	7	11	16	22	28	35	44	54	67
CCA 29	**5**	**12**	**19**	**28**	**38**	**51**	51 (65)	51 (81)	51 (87)	51
CCA 30	3	6	9	12	16	19	23	27	27 (30)	27 (29)
Col Avg	3	5	8	11	14	18	22	26	29	32
Col Max	6	12	19	28	38	51	51	54	54	67
Col 90th Percentile	4	7	11	16	22	28	35	43	50	54
Num CCAs	22	22	22	22	22	22	22	22	22	22

Table G.10: Percentage Bias of Mean Caries Free ($\bar{p_r}$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 5

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	0	0	1	1	1	1	2	2	3	3
CCA 10	0	0	-1	-1	-1	-1	-2	-2	-2	-2 (-3)
CCA 11	0	0	0	-1	-1	-1	-2	-2	-2 (-3)	-2 (-4)
CCA 12	0	0	0	0	0	0	0	0	-1	-1
CCA 13	0	0	-1	-1	-1	-2	-2	-3	-3	-4
CCA 14	0	0	0	0*	0*	0*	0*	0*	0*	0*
CCA 15	0	0	0	0	0*	0*	0*	0*	0*	0*
CCA 16	0	0	0	0*	0*	0*	0*	0*	0*	0*
CCA 17	-1	-1	-2	-3	-4	-5	-7	-8	-10	-10 (-13)
CCA 18	0	0	0	0	0	0	0	0*	0*	0*
CCA 19	0	0	1	1	1	2	2	3	4	5
CCA 20	0	0	-1	-1	-1	-1	-1	-1	-1 (-1)	-1 (-1)
CCA 21	1	1	2	2	3	4	5	6	7	9
CCA 22	0	0	0	0	0	0	0	0	0	0
CCA 23	0	0	0	1	1	1	1	2	2	2 (3)
CCA 24	0	0	0	0	0	0	0	0	0	1
CCA 25	0	0	0	1	1	1	1	2	2 (2)	2 (3)
CCA 26	**0**	0 (0)	0 (1)	0 (1)	0	0	0	0	0	0
CCA 27	0	0	0	0	0	0	0	0*	0*	0*
CCA 28	0	0	0	0	0	0	0	0	-1	-1
CCA 29	**1**	**2**	**4**	**6**	**10**	**14**	14 (21)	14 (32)	14 (51)	14
CCA 30	0	0	0	1	1	2	2	3	3 (4)	3 (7)
Col Avg	0	0	0	0	0	1	1	1	1	1
Col Max	1	2	4	6	10	14	14	14	14	14
Col 90th Percentile	0	0	1	1	1	2	2	3	4	5
Num CCAs	22	22	22	22	22	22	22	22	22	22

Table G.11: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	3	6	9	13	16	20	25	30	35	41
CCA 10	**5**	**11**	**18**	**25**	34	44	44 (57)	44 (74)	44 (98)	44
CCA 11	**8**	**18**	**30**	**30 (45)**	**30 (70)**	**30 (123)**	30	30	30	30
CCA 12	3	6	10	14	18	22	27	32	38	38 (43)
CCA 13	4	10	16	22	22 (30)	22 (36)	22 (32)	22	22	22
CCA 14	3	7	10	14	19	24	30	36	43	51
CCA 15	2	4	6	9	12	15	18	18 (21)	18 (24)	18 (19)
CCA 16	**4**	**8**	**13**	18	24	29	34	34 (37)	34 (36)	34 (23)
CCA 17	3	7	11	15	19	25	30	37	44	44 (51)
CCA 18	2	5	8	11	14	18	21	26	30	35
CCA 19	1	1	1	1	1	0	-1	-2	-4	-8
CCA 20	6	12	19	26	26 (34)	26 (40)	26 (24)	26	26	26
CCA 21	7	14	21	21 (30)	21 (39)	21 (41)	21	21	21	21
CCA 22	**3**	**6**	**10**	**13**	**17**	**20**	23	23 (26)	23 (29)	23 (33)
CCA 23	3	6	10	13	17	21	26	31	36	41
CCA 24	4	8	12	18	24	31	39	49	49 (61)	49 (74)
CCA 25	-1	-2	-3	-4	-6	-7	-9	-10	-10 (-12)	-10 (-12)
CCA 26	**7**	**16**	**27**	**27 (40)**	**27 (57)**	27 (77)	27	27	27	27
CCA 27	3	6	9	13	16	21	25	30	35*	41*
CCA 28	6	6 (14)	6 (26)	6 (47)	6	6	6	6	6	6
CCA 29	(1)	(2)	(-13)							
CCA 30	4	4 (5)	4 (2)	4 (-8)	4	4	4	4	4	4
Col Avg	4	8	12	15	17	20	22	24	26	27
Col Max	8	18	30	30	34	44	44	49	49	51
Col 90th Percentile	7	14	21	26	27	30	34	37	44	44
Num CCAs	21	21	21	21	21	21	21	21	21	21

Table G.12: Percentage Bias of Mean Caries Free ($\bar{p_r}$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 8

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	0	0	0	0	0	0	0	0	0	-1
CCA 10	**0**	**0**	**0**	**0**	0	-1	-1 (-1)	-1 (-1)	-1 (-2)	-1
CCA 11	**0**	**0**	**0**	**0**	**0 (-1)**	**0 (-1)**	0 (-3)	0	0	0
CCA 12	0	0	0	0	0	0	0	0	0	0 (0)
CCA 13	0	0	0	0	0 (0)	0 (0)	0 (-1)	0	0	0
CCA 14	0	0	0	0	0	0	0	0	0	0
CCA 15	0	0	0	0	0	0	0	0 (1)	0 (1)	0 (1)
CCA 16	**0**	**0**	**0**	0	0	0	0	0 (0)	0 (0)	0 (1)
CCA 17	0	0	0	-1	-1	-1	-1	-2	-2	-2 (-3)
CCA 18	0	0	0	0	0	0	0	0	0	0
CCA 19	0	1	2	2	3	4	5	6	8	10
CCA 20	0	0	0	0	0 (-1)	0 (-1)	0 (-2)	0	0	0
CCA 21	0	0	-1	-1 (-1)	-1 (-2)	-1 (-3)	-1	-1	-1	-1
CCA 22	**0**	**0**	**0**	**0**	**-1**	**-1**	-1	-1 (-1)	-1 (-2)	-1 (-2)
CCA 23	0	0	0	0	0	0	0	0	0	0
CCA 24	0	0	0	0	0	0	0	-1	-1 (-1)	-1 (-1)
CCA 25	0	1	1	1	2	2	3	3	3 (4)	3 (5)
CCA 26	**0**	**0**	**-1**	**-1 (-1)**	**-1 (-2)**	-1 (-3)	-1	-1	-1	-1
CCA 27	0	0	0	0	0	0	0	0	0*	0*
CCA 28	0	0 (-1)	0 (-1)	0 (-3)	0	0	0	0	0	0
CCA 29	(0)	(-1)	(-2)							
CCA 30	1	1 (2)	1 (4)	1 (7)	1	1	1	1	1	1
Col Avg	0	0	0	0	0	0	0	0	0	0
Col Max	1	1	2	2	3	4	5	6	8	10
Col 90th Percentile	0	1	1	1	1	1	1	1	1	1
Num CCAs	21	21	21	21	21	21	21	21	21	21

Table G.13: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	4	9	14	19	25	32	39	39 (46)	39 (51)	39 (47)
CCA 10	2	5	8	11	15	15 (19)	15 (23)	15 (21)	15	15
CCA 11	4	8	13	19	25	31	39	47	47 (55)	47 (62)
CCA 12	3	7	11	16	21	27	33	40	49	59
CCA 13	**11**	11 (25)	11 (45)	11 (73)	11	11	11	11	11	11
CCA 14	3	6	9	12	16	20	24	29	34	39
CCA 15	4	8	12	17	23	29	36	43	51	51 (58)
CCA 16	4	9	14	21	28	36	45	45 (56)	45 (68)	45 (74)
CCA 17	2	4	6	8	10	12	15	18	20	23
CCA 18	3	6	9	12	16	20	25	30	35	42
CCA 19	4	9	15	22	30	39	39 (50)	39 (57)	39 (40)	39
CCA 20	6	14	22	22 (30)	22 (36)	22 (29)	22	22	22	22
CCA 21	**9**	**21**	21 (36)	21 (48)	21 (34)	21	21	21	21	21
CCA 22	8	17	17 (25)	17 (29)	17 (16)	17	17	17	17	17
CCA 23	4	8	12	17	23	29	35	43	50	50 (57)
CCA 24	3	7	10	14	19	24	30	36	44	52
CCA 25	3	6	9	13	18	23	29	36	45	55
CCA 26	10	10 (24)	10 (40)	10 (44)	10	10	10	10	10	10
CCA 27	3	5	8	11	15	18	22	26*	31*	36*
CCA 28	**(17)**	(44)								
CCA 29	**(-10)**									
CCA 30	8	8 (17)	8 (26)	8 (19)	8	8	8	8	8	8
Col Avg	5	9	12	15	18	22	26	29	32	34
Col Max	11	21	22	22	30	39	45	47	51	59
Col 90th Percentile	9	14	17	21	25	32	39	43	49	53
Num CCAs	20	20	20	20	20	20	20	20	20	20

Table G.14: Percentage Bias of Mean Caries Free ($\bar{p_r}$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 12

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	0	0	0	0	0	0	0	0 (1)	0 (1)	0 (2)
CCA 10	-1	-1	-2	-3	-5	-5 (-6)	-5 (-8)	-5 (-9)	-5	-5
CCA 11	0	0	0	0	0	0	-1	-1	-1 (-1)	-1 (-1)
CCA 12	0	0	0	0	0	0	0	0	0	0
CCA 13	**0**	0 (0)	0 (1)	0 (3)	0	0	0	0	0	0
CCA 14	0	0	0	0	0	0	0	0	0	0
CCA 15	0	0	0	0	0	0	0	0	0	0 (0)
CCA 16	0	0	1	1	1	2	3	3 (4)	3 (6)	3 (10)
CCA 17	0	1	1	1	2	2	3	3	4	5
CCA 18	0	0	0	-1	-1	-1	-1	-1	-2	-2
CCA 19	0	1	1	2	2	3	3 (4)	3 (5)	3 (5)	3
CCA 20	1	2	4	4 (6)	4 (10)	4 (18)	4	4	4	4
CCA 21	**2**	**6**	6 (11)	6 (21)	6 (44)	6	6	6	6	6
CCA 22	1	1	1 (2)	1 (4)	1 (9)	1	1	1	1	1
CCA 23	0	0	0	0	1	1	1	1	2	2 (2)
CCA 24	0	0	0	0	0	0	0	-1	-1	-1
CCA 25	0	0	0	0	0	0	0	0	-1	-1
CCA 26	0	0 (0)	0 (-1)	0 (-1)	0	0	0	0	0	0
CCA 27	0	0	0	0	0	0	0	0*	0*	0*
CCA 28	(1)	(1)								
CCA 29	(-1)									
CCA 30	1	1 (3)	1 (5)	1 (9)	1	1	1	1	1	1
Col Avg	0	1	1	1	1	1	1	1	1	1
Col Max	2	6	6	6	6	6	6	6	6	6
Col 90th Percentile	1	1	2	2	2	3	3	3	4	4
Num CCAs	20	20	20	20	20	20	20	20	20	20

Table G.15: Percentage Change in Mean SE of Estimate of Mean Caries Free by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	(3)	(5)	(4)							
CCA 10	11	11 (23)	11 (32)	11 (34)	11	11	11	11	11	11
CCA 11	9	18	18 (26)	18 (31)	18 (27)	18	18	18	18	18
CCA 12	7	14	23	23 (32)	23 (38)	23 (30)	23	23	23	23
CCA 13	**10**	20	33	33 (48)	33 (67)	33 (91)	33	33	33	33
CCA 14	5	12	20	20 (32)	20 (50)	20 (73)	20	20	20	20
CCA 15	12	25	39	39 (54)	39 (65)	39 (58)	39	39	39	39
CCA 16	5	11	17	24	31	31 (39)	31 (45)	31 (47)	31	31
CCA 17	5	10	13	16	16 (16)	16 (12)	16 (2)	16	16	16
CCA 18	6	11	17	23	29	29 (34)	29 (36)	29 (30)	29	29
CCA 19	**13**	**13 (30)**	**13 (49)**	13 (63)	13	13	13	13	13	13
CCA 20	(-1)	(-7)	(-21)							
CCA 21										
CCA 22	**5**	5 (12)	5 (20)	5 (28)	5	5	5	5	5	5
CCA 23	8	17	28	28 (40)	28 (48)	28 (42)	28	28	28	28
CCA 24	5	11	17	24	33	43	43 (54)	43 (66)	43 (75)	43
CCA 25	(2)	(-14)								
CCA 26	2	2 (2)	2 (0)	2 (-8)	2	2	2	2	2	2
CCA 27	1	1 (3)	1 (5)	1 (6)	1	1	1	1	1	1
CCA 28										
CCA 29	**(-5)**									
CCA 30	(-10)									
Col Avg	7	12	17	19	20	21	21	21	21	21
Col Max	13	25	39	39	39	43	43	43	43	43
Col 90th Percentile	11	20	31	31	33	37	37	37	37	37
Num CCAs	15	15	15	15	15	15	15	15	15	15

Table G.16: Percentage Bias of Mean Caries Free ($\bar{p_r}$) by Reduction in the Number of Schools Selected (by CCA) for Non-Fluoridated Children Aged 15

Clusters Dropped	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CCA 9	(-1)	(-4)	(-8)							
CCA 10	2	2 (5)	2 (11)	2 (26)	2	2	2	2	2	2
CCA 11	0	-1	-1 (-1)	-1 (-2)	-1 (-4)	-1	-1	-1	-1	-1
CCA 12	0	-1	-1	-1 (-2)	-1 (-3)	-1 (-7)	-1	-1	-1	-1
CCA 13	**1**	1	2	2 (4)	2 (7)	2 (12)	2	2	2	2
CCA 14	0	0	-1	-1 (-1)	-1 (-2)	-1 (-4)	-1	-1	-1	-1
CCA 15	0	1	1	1 (2)	1 (3)	1 (5)	1	1	1	1
CCA 16	0	0	0	1	1	1 (2)	1 (2)	1 (4)	1	1
CCA 17	1	2	3	4	4 (5)	4 (6)	4 (7)	4	4	4
CCA 18	-1	-1	-2	-3	-4	-4 (-5)	-4 (-7)	-4 (-10)	-4	-4
CCA 19	**3**	**3 (8)**	**3 (16)**	3 (33)	3	3	3	3	3	3
CCA 20	(-3)	(-7)	(-14)							
CCA 21										
CCA 22	**-1**	-1 (-2)	-1 (-5)	-1 (-10)	-1	-1	-1	-1	-1	-1
CCA 23	0	0	1	1 (1)	1 (2)	1 (4)	1	1	1	1
CCA 24	0	0	0	-1	-1	-2	-2 (-2)	-2 (-3)	-2 (-5)	-2
CCA 25	(-2)	(-5)								
CCA 26	3	3 (6)	3 (11)	3 (19)	3	3	3	3	3	3
CCA 27	0	0 (-1)	0 (-2)	0 (-3)	0	0	0	0	0	0
CCA 28										
CCA 29	**(0)**									
CCA 30	(-6)									
Col Avg	0	1	1	1	1	1	1	1	1	1
Col Max	3	3	3	4	4	4	4	4	4	4
Col 90th Percentile	2	2	3	3	3	3	3	3	3	3
Num CCAs	15	15	15	15	15	15	15	15	15	15