Sample design for quality monitoring and measurement error evaluation of large-scale longitudinal surveys

Cristiano Ferraz
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Statistics and Probability Commons

Recommended Citation
Ferraz, Cristiano, "Sample design for quality monitoring and measurement error evaluation of large-scale longitudinal surveys" (2004). Retrospective Theses and Dissertations. 1160.
https://lib.dr.iastate.edu/rtd/1160
Sample design for quality monitoring and measurement error evaluation of large-scale longitudinal surveys

by

Cristiano Ferraz

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Sarah M. Nusser, Co-major Professor
Jean D. Opsomer, Co-major Professor
Huaiqing Wu
William Duckworth
Joseph Herriges

Iowa State University
Ames, Iowa
2004
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Graduate College
Iowa State University

This is to certify that the doctoral dissertation of
Cristiano Ferraz
has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member
Signature was redacted for privacy.

Committee Member
Signature was redacted for privacy.

Committee Member
Signature was redacted for privacy.

Co-major Professor
Signature was redacted for privacy.

Co-major Professor
Signature was redacted for privacy.

For the Major Program
DEDICATION

This dissertation is dedicated to my parents, Euclides Ferraz Filho and Maria Margarida de Assis Ferraz, and to the memory of my friend, Ivan Brondi de Carvalho Filho.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. vii
LIST OF FIGURES ................................................................. ix

CHAPTER 1. Nonsampling Errors and Survey Quality in Large-scale Surveys 1
1.1 Introduction ........................................................................... 1
1.2 Literature Review on Quality Monitoring ................................. 4
1.3 Literature Review on Measurement Error in Surveys ................. 7
1.4 The National Resources Inventory ........................................... 9
1.5 The Current NRI Sample Design .......................................... 10
1.6 Considerations about the NRI Measurement Process ............ 11
1.7 Quality Monitoring Procedures in Prior NRI Surveys ............ 12
1.8 Prior NRI Quality Evaluation Study .................................. 13
1.9 A Measurement Error Model for the NRI Measurement Process . 13

CHAPTER 2. Sample Design for Monitoring the Data Collection

Process of Large-scale Longitudinal Surveys ................................ 16
2.1 Designing a Sample for Quality Monitoring Purposes ............ 16
2.2 Building First-order Inclusion Probabilities .......................... 18
2.3 Alternative Designs ............................................................. 19
  2.3.1 Poisson Sampling (POI) ............................................. 19
  2.3.2 Systematic Sampling with Probability Proportional to Size (SYS/PPS) ...................................................... 20
  2.3.3 A One Unit-per-wave Sampling with Probability Proportional to Size (1PW/PPS) ............................................. 21
2.4 Assessment of Pattern-recognition Constraints ..................... 21
2.5 Sample Design Assessment for Quality Monitoring: Application to The National Resources Inventory ............................................. 22
   2.5.1 Survey Setting .................................................................. 22
   2.5.2 Determining the Size Variable ............................................. 23
   2.5.3 Simulation Design .............................................................. 25
   2.5.4 Performance Evaluation ..................................................... 27
   2.5.5 Evaluation of Pattern-recognition Constraints ...................... 30
2.6 Summary of Simulation Results ............................................... 32
2.7 A Quality Monitoring Sample Design for Large-scale Longitudinal Surveys ................................................................. 34
   2.7.1 Overview ....................................................................... 34
   2.7.2 One Unit-per-wave Two-stage PPS Sample (1PW2/PPS) ........ 35
   2.7.3 Simulation Design ............................................................. 35
   2.7.4 Performance Evaluation ..................................................... 37
   2.7.5 Evaluation of Pattern-recognition Constraints ...................... 37
2.8 Conclusions ........................................................................ 39

CHAPTER 3. A Sample Design for Measurement Error Evaluation in Large-scale Longitudinal Surveys ................................................. 41
3.1 Introduction ........................................................................ 41
3.2 Model Development ............................................................. 41
   3.2.1 Introduction ................................................................... 41
   3.2.2 A Measurement Error Model .......................................... 43
3.3 Estimation of the Variance Components .................................... 44
3.4 The Subsample Design .......................................................... 63
3.5 A Sample Allocation Simulation Study ....................................... 65
   3.5.1 A Sensitivity Analysis ....................................................... 70
3.6 Simulation Conclusions ........................................................ 73

CHAPTER BIBLIOGRAPHY ............................................................ 76
LIST OF TABLES

Table 1.1 NRI Current Panel Sample Design ............... 10
Table 2.1 Frequency Distribution of Segments Collected by Data Gatherers ... 23
Table 2.2 Rules for Forming Waves ....................... 26
Table 2.3 Simulation Results for Sample Sizes per Data Gatherer for Poisson Design 28
Table 2.4 Number of Data Gatherers with no Sample Size Drawn .......... 28
Table 2.5 Total Number of Segments for Data Gatherers with no Sample Drawn 28
Table 2.6 Simulation Results for Poisson Design ............. 29
Table 2.7 Simulation Results for Systematic with PPS Design .......... 29
Table 2.8 Simulation Results for One Unit-per-wave with PPS Design ...... 30
Table 2.9 Comparison of Sample Designs Based on satisfaction of Pattern-recognition Constraints for Intersample Mean Distance (Φ) ............... 31
Table 2.10 Comparison of Sample Designs Based on Satisfaction of Pattern-recognition Constraints for Intersample Standard Deviation Distance (Δ) .... 32
Table 2.11 Monte Carlo Percentiles of the Min. of Δ by Sample Design .... 33
Table 2.12 Number of Data Gatherers with Samples Generating Δ = 0 and, the Mean and Maximum of the Total Number of Segments Collected by These Data Gatherers ........................................ 34
Table 2.13 Waves for Each Data Gatherer (DG) by Level of Experience .... 36
Table 2.14 Simulation Results for 1PW2/PPS Design .................. 37
Table 2.15 Results Regarding Pattern-recognition Constraints for Intersample Mean (Φ) ........................................ 38
Table 2.16 Results Regarding Pattern-recognition Constraints for Intersample Standard Deviation (Δ) ................................................................. 38
Table 2.17 Monte Carlo Percentiles of Min. of Δ(S) for 1PW2/PPS Design .... 39
Table 2.18 Number of Data Gatherers with Samples Generating Δ = 0, and the Mean and Maximum of the Total Number of Segments Collected by These Data Gatherers for 1PW2/PPS Design .................................... 39

Table 3.1 Allocation Types Summary ............................................................ 65
Table 3.2 Group Description and Size ............................................................ 66
Table 3.3 Values of Proportion of Measurement Error, Variance of Soil Erosion Loss and Measurement Error Variance by Group Used in the Simulation 67
Table 3.4 Simulation Values of the Variance of \( \hat{\theta}_1 \) by Allocation Type and Sample Size ................................................................. 69
Table 3.5 Lower Values of \( v_g \) and Corresponding Values of \( \sigma_g^2 \) for Soil Erosion Loss by Group ................................................................. 70
Table 3.6 Simulation values of the Variance of \( \hat{\theta}_1 \) by Allocation Type and Sample Size for Lower Values ................................................................. 73
Table 3.7 Higher Values of \( v_g \) and \( \sigma_g^2 \) for Soil Erosion Loss ................. 74
Table 3.8 Simulation Values of the Variance of \( \hat{\theta}_1 \) by Allocation Type and Sample Size for Higher Values ................................................................. 75
LIST OF FIGURES

Figure 3.1  Graphical Analysis of Allocation Types .......................... 68
Figure 3.2  Sensitivity Analysis of Allocation Types for Lower Values .... 71
Figure 3.3  Sensitivity Analysis of Allocation Types for Higher Values .... 72
CHAPTER 1. Nonsampling Errors and Survey Quality in Large-scale Surveys

1.1 Introduction

The goal of producing high quality survey data and estimates has been an important factor in the design and implementation of sample surveys from the field’s earliest days. [Biemer and Lyberg, 2003, p.13] define survey quality as a tri-dimensional characteristic of survey data. The three dimensions are accuracy, timing, and accessibility. This is consistent with the interpretation of [Juran and Jr, 1980, p.1] quality definition in the industrial production setting. In order to have good survey quality, the survey data must be sufficiently accurate to achieve their designed purpose in a suitable time period and in an accessible way. In their general definition of survey quality, [Biemer and Lyberg, 2003] emphasize the importance of the accuracy since data with errors are of little help regardless of release timing and accessibility.

It is useful to identify the different kinds of error a survey is subject to. In a general framework, the total survey error is the sum of two major types of error: sampling error and nonsampling error. Sampling error is the error due to the fact that only a subset of the population (the sample) is observed. Survey sampling theory has developed effective tools to design a survey sample that controls for this source of error. Nonsampling error is composed of those errors related to operational issues of the survey. [Biemer and Lyberg, 2003, p.38] subdivide nonsampling errors into five categories: specification error, frame error, nonresponse error, processing error and measurement error. A brief description of each category follows.

• Specification Errors: Specification errors are essentially conceptual errors. This type of
error happens when there is disagreement between the intended construct to be evaluated and the one that is actually measured. For example, the construct “salary” is not necessarily the same as the amount earned by someone. Hence, if researchers intend to study the variable salary but use a survey question such as “How much do you make a year?”, they are inducing a specification error. In this example, there will be a tendency to overestimate the variable of interest, salary. Specification errors may be avoided with careful planning of survey questions and good communication between researchers and questionnaire designers.

- **Frame Errors:** Frame errors are typically errors associated with the existence of duplicate frame records and the level of coverage of the sampling frame used to select a probability sample. The existence of duplicated records is undesirable as it affects the actual selection probabilities of a sample design. On the other hand, frames containing more elements than the target population lead to overcoverage problems, while frames failing to cover all elements of the target population generate undercoverage problems. The latter is a more serious problem than the former. [Kish, 1962, pp.529-532] discusses the topic of undercoverage under a broader view of errors of nonobservation. Sometimes, the use of multiple frame designs ([Hartley, 1962]) may help to avoid undercoverage problems.

- **Nonresponse Error:** Nonresponse error is the error associated with the absence of response. For example, in a natural resource survey, a nonresponse may occur by an inability to physically access a certain geographical area leading to missing information regarding that unit. In human population surveys, refusal to answer survey questions or the entire survey are sources of nonresponse error. In order to efficiently handle this type of error, it is necessary to make assumptions about the underlying mechanism which generates the nonresponse. Techniques of statistical imputation or weighting can be used to help to address the problem.

- **Processing Errors:** Processing errors are errors related to post-data collection processes, including coding, keyentry, editing and statistical estimation procedures such as imputa-
tion and weighting. The more complex data processing is, the more vulnerable a survey is to processing errors. Mistakes in the coding process and keyentry are common examples of processing errors. Alternatively, error in computer code can induce a systematic error in the survey data or weights. Processing errors may be controlled by appropriate training of people in charge of operational tasks and by supervision of work. Procedures to monitor the quality of the data processing step are also an important tool to help diminish the problem.

- **Measurement Error**: Measurement error is the error associated with imperfections in the measurement process. A non-calibrated instrument, for instance, is a source of measurement error. False information from the respondent is also an example of this type of error. Compared to the other categories of nonsampling errors, measurement errors have received the majority of attention in survey methods research. In order to help diminish this problem, training of data collectors, calibrating instruments, and monitoring the data collection process are all essential. Special studies may be implemented to estimate the properties of measurement error for a survey.

In this dissertation, attention will be given to the design of samples to monitor the quality of the data being collected (data quality monitoring) and to evaluate properties of measurement errors for the survey (measurement error evaluation) in the context of large-scale longitudinal surveys. Longitudinal surveys provide historical information that allows more complex designs than the simple ad hoc approaches used in one-time surveys. The properties and feasibility of several classes of probability sample designs for data quality monitoring will be investigated. The problem of assessing the properties of measurement error will be addressed with emphasis on designing a subsample to estimate the measurement error contribution to the variance of a sample estimator. The investigation is motivated by potential applications to the United States Department of Agriculture's National Resources Inventory.

The dissertation is structured in three major chapters. The first chapter presents a brief introduction to the subject of survey quality. A literature review for quality monitoring and measurement errors in surveys is presented. Also, the National Resources Inventory is intro-
duced and its measurement process is discussed. Chapter 2 presents the problem of designing a sample for data quality monitoring purposes. Several classes of sample designs are investigated. Based on the performance of these designs, a sample design is presented for application, motivated by the NRI context. In Chapter 3, the problem of designing a sample for measurement error evaluation is discussed. Based on a measurement error model, a stratified design for subsampling sampling units is proposed in order to assess the contribution of measurement error to the variance of an estimate. Different types of sample allocation are compared in a simulation study. A sensitivity analysis is conducted to evaluate the robustness of the results.

1.2 Literature Review on Quality Monitoring

According to the [US Federal Committee on Statistical Methodology, 1990], editing is defined as a set of ‘procedures designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures.’ [Biemer and Lyberg, 2003] enumerate three goals of the editing phase: to provide information about the data quality, to provide useful information for future survey improvements and to ‘clean-up’ the data for further processing.

Although editing is a very important phase of data processing, [Grankist and Kovar, 1997], warn of the risks of overediting. A data set is said to be overedited when it is submitted to extensive editing without major quality improvement, causing an unnecessary increase on the survey cost.

In an attempt to avoid overediting, [Biemer and Lyberg, 2003] advocate the use of selective editing. Adopting a selective method of editing means that not necessarily all the suspicious data are subjected to checking. Instead, only a subset of the data, chosen based on different factors, such as the importance of the sampling unit, the relevance of the variable of interest and the cost of investigation, are investigated. Methods from statistical quality assurance theory can then be applied to monitor the quality of the data. In particular, the techniques
of acceptance sampling and process monitoring can be used to ensure the errors are under control. An example of the application of these tools in a clerical operation survey scenario is given by [Rosander, 1977, p.187]. [Vardeman and Jobe, 1999] apply these methods in an engineering setting. The traditional methods of statistical quality assurance, invented originally for the industrial settings, were later applied to the survey sample context as administrative applications of quality control ([Biemer and Lyberg, 2003, p.220]).

A typical setup for applying acceptance sampling in a survey requires the existence of 'batches of data'. For example, in natural resource surveys, these could be sets of field reports. From each batch, a random sample of field reports is taken and checked for data inconsistencies. If the estimated percentage of reports with inconsistencies is high, the whole batch is rejected in the sense that all its items are submitted to evaluation. Otherwise, no further checking is deemed necessary. Although some statistical agencies use this technique with success, [Deming, 1986] and [Biemer and Caspar, 1994] have criticisms with respect to its implementation since, as the methods of mass-inspection, acceptance sampling may have a high cost, be inefficient at identifying errors and, without any adjustment, present no feedback that allows for continuous improvement of quality.

An alternative approach is to engage in on-going monitoring of survey quality. Among the techniques for quality monitoring, the general class of Shewhart charts, along with the EWMA and CUSUM charts, are the main tools for the application of quality monitoring in real time.

Shewhart charts are based on the belief that the variation in a process is the sum of two components: inherent variation (that cannot be suppressed) and variation that can be suppressed. The main goal of the charts is to provide a tool to detect the variation that can be avoided, giving the chance to eliminate it by means of physical intervention.

In order to implement a Shewhart chart, samples of the process under observation are taken periodically. Based on these samples, key statistics are calculated. Their values are plotted in the order they were generated and the pattern of the chart is compared with some control limits. Whenever the values of the charts start falling outside the control limits, there is evidence that something has occurred to the process that needs investigation.
The EWMA charts are based on ‘exponentially weighted moving averages’. In this type of chart, the sequence of values used for a Shewhart chart is smoothed through the use of a recursive expression. Let the sequence of the key statistic values plotted in the Shewhart chart be denoted by $T_1, T_2, \ldots$. Let the EWMA sequence of statistics to be plotted be denoted by $E_0, E_1, E_2, \ldots$, where $E_0$ is some starting value. Then the EWMA recursive expression is given by

$$E_k = \lambda T_k + (1 - \lambda)E_{k-1} \tag{1.1}$$

where $\lambda$ is a constant that may assume a value over the interval $(0, 1]$. Note that for $\lambda = 1$, the EWMA is equivalent to the Shewhart chart. The smaller the value of $\lambda$, the smoother the pattern of the chart is. In this type of plot, smaller changes and weak trends on a process under monitoring may be easier to detect, possibly at an earlier stage, than on a Shewhart chart.

As with the EWMA, the CUSUM chart intends to smooth the Shewhart chart. CUSUM is the abbreviation for cumulative sum based charts. Retaining the notation for the key statistic values plotted in the Shewhart chart, and denoting the CUSUM sequence of values as $C_0, C_1, C_2, \ldots$, where $C_0$ is some starting value, the CUSUM recursive expression is given by

$$C_k = (T_k - \alpha) + C_{k-1} \tag{1.2}$$

where $\alpha$ is some constant.

Interpretations of the CUSUM charts requires more attention as the values of the slopes on this graphics are the indicators for the sizes of the statistics been observed. A positive trend on a CUSUM chart indicates that values larger than $\alpha$ are been observed while a negative trend indicates the opposite. Experienced CUSUM chart analysts argue that it is easier to see small changes in the process through this type of chart than through a Shewhart chart.

Sometimes, combinations of these monitoring charts are also employed in practice.

In order to apply any of these techniques, a subset of sampling units must be chosen as the units comes from the field. It is usual practice to choose this subset based on specific ad hoc
rules serving as general guidelines that amount to non-probabilistic sample designs.

1.3 Literature Review on Measurement Error in Surveys

Historically, literature concerning errors of measurement may date back as far as 1902 with Karl Pearson's publication 'On the mathematical theory of errors of judgment'. In this work, [Pearson, 1902] investigated measurement error behavior when a human being was performing measurement of a fixed and well defined quantity.

An important early discussion about measurement error in survey sampling is found in [Mahalanobis, 1946]. In this work, Mahalanobis addressed the problem of designing a sample in which nonsampling error due to interviewer influence would be under some control. Mahalanobis used interpenetrating subsamples, which rely on randomly assigned interviews to each available interviewer. Later, [Hansen et al., 1951], and [Sukhatme and Seth, 1952] worked on measurement error models with variance components for nonsampling errors. In particular, motivated by surveys carried out by the Indian Council of Agricultural Research, [Sukhatme and Seth, 1952] formulated a general linear model believed to cover commonly found conditions in agricultural and socio-economic surveys. Let $y_i$ be a certain variable of interest. Let $Y_{ijk}$ be the $k^{th}$ measurement of the variable of interest, observed by the $j^{th}$ interviewer on the $i^{th}$ unit. Let $y_i$ be the true value of the variable of interest associated with the $i^{th}$ unit in the sample, selected from a finite or infinite population of mean $\mu$ and variance $\sigma^2$. Then, Sukhatme and Seth's model can be expressed as

$$M_1 : Y_{ijk} = y_i + \alpha_j + \delta_{ij} + e_{ijk},$$

(1.3)

where $\alpha_j$ is a bias effect associated with the $j^{th}$ interviewer, $\delta_{ij}$ is an interaction term between the $j^{th}$ interviewer and the $i^{th}$ unit, and $e_{ijk}$ is a random error term. Additional assumptions include $E_{M_1}(e_{ijk}|i,j) = 0$, $E_{M_1}(\delta_{ij}|j) = 0$, $E_{M_1}(e_{ijk}^2|i,j) = \sigma^2_e$ and $E_{M_1}(\delta_{ij}^2|j) = \sigma^2_\delta$, where $E_{M_1}(\cdot)$ denotes expectation with respect to model $M_1$. In their paper, [Sukhatme and Seth, 1952] note that, in order to estimate the different variance components, replication of the observations is necessary.
The conceptualization of a model to describe measurement error in a survey implies an extra source of variation, in addition to the fundamental randomization process applied in any probability sample. [Hansen et al., 1961] introduce a decomposition of the mean square error taking into account both sources of variation: from the randomization and from the model. Although they assume a survey as either a census or an equal selection probability sample, the idea is general enough to extend for any sample design \( p \) and measurement error model \( M \). Let \( \theta \) be a parameter of interest related to a finite population. Consider a sample estimator of \( \theta \), given by \( \hat{\theta} \). Let \( E_p(.) \) denotes the expectation over all possible samples under design \( p \) and over all possible measurements under model \( M \). Also, define \( E_p(.) \) and \( E_M(.) \) as the expected value with respect to the sample design \( p \) and model \( M \), respectively. Then, \( E_p M(.) = E_p[E_M(.)] \). Hansen, Hurwitz and Bershad's MSE-decomposition can be expressed as

\[
E_p M(\hat{\theta} - \theta)^2 = [E_p M(\hat{\theta}) - \theta]^2 + E_p M(\hat{\theta} - E_M(\hat{\theta}))^2 + E_p M[E_M(\hat{\theta}) - E_p M(\hat{\theta})]^2.
\]

The first term on the right side is the squared bias of the sample estimator. The second term can be interpreted as the response variance, while the last one can be interpreted as the sample variance. By definition, the response variance is given by

\[
E_p M[\hat{\theta} - E_M(\hat{\theta})]^2 = E_p \{E_M[\hat{\theta} - E_M(\hat{\theta})]^2\}
= E_p \text{Var}_M(\hat{\theta}).
\]

On the other hand, the sample variance is given by

\[
E_p M[E_M(\hat{\theta}) - E_p M(\hat{\theta})]^2 = E_p \{E_M(\hat{\theta}) - E_p[E_M(\hat{\theta})]\}^2
= \text{Var}_p E_M(\hat{\theta}),
\]

where \( \text{Var}_p(.) \) and \( \text{Var}_M(.) \) are the variances with respect to the sample design \( p \) and model \( M \), respectively. Let \( \text{Var}_p M(.) \) denote the variance of the sample estimator over all possible samples under design \( p \) and over all possible measurements under model \( M \). Then,
\[ \text{Var}_{PM}(\hat{\theta}) = \text{E}_p \text{Var}_M(\hat{\theta}) + \text{Var}_p \text{E}_M(\hat{\theta}). \]

Based on this last expression, the response variance corresponds to the measurement error contribution to the variance of the sample estimator. [Hansen et al., 1961] worked on analysis and estimation of the response variance.

Investigations concerning model use and the effects of measurement error in the analysis of survey data were done by [Cochran, 1968], [Bailar and Dalenius, 1969] and [Chandhok, 1982]. [Lessler, 1984] made a comprehensive review of terminology associated with measurement errors. The importance of the subject lead to the publication of books by [Biemer et al., 1991], “Measurement Errors in Surveys”, and, by [Lyberg et al., 1997], “Survey Measurement and Process Quality”. [Fuller, 1995] adresses aspects of data analysis when the observations are subjected to measurement error.

1.4 The National Resources Inventory

Over the years, the United States (U.S.) government has supported the implementation of a variety of large-scale natural resource surveys. Some examples of them are the U.S. Forest Service’s Forest Inventory and Analysis Program, the U.S. Geological Survey’s National Water Quality Assessment Program, and the U.S. Department of Agriculture’s (USDA) National Resources Inventory (NRI). In this dissertation, sample designs for quality monitoring and measurement error evaluation will be investigated within the context of large-scale longitudinal surveys, specifically the NRI.

The NRI was carried out every five years from 1982 to 1997, and has been conducted annually since 2000. Its main goal is to provide an assessment of soil, water and related natural resources conditions over time on non-federal lands of United States. The NRI is conducted by the USDA’s Natural Resources and Conservation Service (NRCS), in cooperation with the Iowa State University Center for Survey Statistics and Methodology. A historical overview of the NRI is found in [Nusser and Goebel, 1997].
1.5 The Current NRI Sample Design

Up through 1997, the NRI sample design was a stratified two stage area sample, carried out every five years. The NRI sample design incorporates features such as geographic dispersion and the ability to provide adequate sample sizes and units for estimates of longitudinal and level parameters. In a typical Public Land Survey County, strata are defined to be 2 mi x 6 mi sections. An area segment corresponds to a quarter section of 0.5 mi x 0.5 mi or 160 acres in size. Typically 2 segments are selected per stratum, but the rates may vary based on the heterogeneity of the landscape and domain considerations. In most segments, 3 points are selected using a restricted randomization design to disperse the points geographically. The Foundation NRI sample corresponds to the 1997 NRI survey and is composed of about 300,000 area segments, and approximately 800,000 points.

Research was conducted in order to implement the NRI on an annual basis from 2000 on ([Breidt and Fuller, 1999]). The current NRI sample design is a two-phase supplemented panel design where the first phase is the Foundation sample. Starting in year 2000, samples from the Foundation sample are taken annually following a supplemented panel design, as illustrated in Table 1.1. Approximately 40,000 Foundation area segments are revisited every year, referred to as the core panel. The core panel is supplemented with a rotation panel of roughly 30,000 segments, called the rotation panel. A new rotation panel is selected from the Foundation each year. Eventually, rotation segments will cycle back into the survey.

Table 1.1 NRI Current Panel Sample Design

<table>
<thead>
<tr>
<th>Panel name</th>
<th>Panel type</th>
<th>Foundation Sample (1997)</th>
<th>Years of Data Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2000</td>
</tr>
<tr>
<td>P00</td>
<td>Core</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P01</td>
<td>Rotation</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>P02</td>
<td>Rotation</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>P03</td>
<td>Rotation</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
1.6 Considerations about the NRI Measurement Process

The National Resources Inventory (NRI) relies on a measurement protocol involving many steps. In its current form, most of the collected data come from photo interpretation, although occasionally data come from on-site evaluation. In this dissertation, the observations are assumed to be generated using remote sensing techniques only. Data collectors examine low-altitude aerial photographs and other auxiliary resources such as topographic and soil maps, for the current and for a prior year. They then outline polygons that denote the boundaries of landscape features (e.g. water body, built-up land) in the area segment, and observe conditions (e.g. land cover, cropping practice, forest type) at each sample point.

In the survey, it is possible to identify three major factors that influence the measurement process: the object under observation, the instrument (or collection of instruments) used to make the observation, and the observer. In the NRI, the objects under observation are typically features in an area segment or conditions at a point. The instruments include the photo and auxiliary maps, while the observer is a data gatherer. In order to have a more comprehensive understanding of the nature of measurement error in the NRI, it is relevant to understand the role of each one of these three factors in the measurement process. As a point of reference for discussing these roles, two characteristics on the land are considered: the land cover/use for a point, and the area for a feature in the segment. In addition, we assume the perspective of evaluating these characteristics at a given point in time.

Consider first the role of the object under observation, an area segment. It is reasonable to assume that the landscape in the segment has an impact on the measurement process. As an example, suppose the land in question is cropland. It is known that if the observation is made by interpreting a photographic image taken at early stages of the crop growing season, it is not a simple task to identify the land as cultivated with corn or soybean. This suggests that, because of its nature (if an area segment is a cropland or noncropland), some observational units may be harder to measure than others. The research conducted in this dissertation considers landscape characteristics at the sample design stage.

The second factor is the photographic image and other auxiliary resources. A low-altitude
image is acquired from an airplane. Factors such as camera quality and the atmospheric conditions lead to different levels of the image quality. Even a good photograph can be difficult to interpret due to differences in the quality and appearance of the current and prior years images, positional mismatches with auxiliary maps, or the obscuring of features boundaries as a result of shadows. A potentially large source of error occurs for rotation segments because the quality of the prior year (1997) image is very poor compared to the quality of images obtained since 2000. Thus, panel type is considered as a feature in this dissertation's sample design investigations.

The last factor is the observer, who interprets the available imagery and auxiliary resources. In this interpretation process, the experience of the technician has an impact on the observations. For the purpose of photo interpretation, the technician has access to available information from past surveys regarding the location under observation and it is critical to relocate the points where they have been previously observed. Further, some features can be very challenging to assess. Because data collection is a difficult process, the more experienced the data collector, the more accurate the observations tend to be. In this dissertation, data gatherer experience is considered in designs to monitor the data collection process.

1.7 Quality Monitoring Procedures in Prior NRI Surveys

Data quality monitoring procedures in prior NRI surveys have relied on ad hoc methods. In the past, NRI data have been collected by USDA employees, most recently organized through 18 Inventory Coordination and Collection Sites. Each of these sites has a supervisor to oversee data collection. Supervisors are trained in remote sensing and natural resource science, but are generally not trained in survey methods. Supervisors are expected to implement a quality monitoring plan, but past experiences indicated that the effectiveness of monitoring plans and the degree to which data gatherers are reviewed varied widely across data collection sites. The research described in this dissertation on alternative quality monitoring designs for longitudinal surveys was motivated by the desire to implement a formal monitoring procedure for the NRI. The goals were to create a procedure that could be uniformly implemented across data
collection sites via the computer-assisted data collection system for NRI surveys, and that take into account prior information on the propensity for area segments to be subject to measurement errors.

1.8 Prior NRI Quality Evaluation Study

A quality evaluation study has not been conducted since the 1982 NRI. [Francisco, 1986] describe the quality evaluation study for the 1982 NRI in which the properties of measurement error of that survey are investigated. Altogether, 3426 area segments were selected for the study accordingly to a stratified two-stage sample design. Each state was stratified based on cropland acreage and amount of conservation operations technical assistance funds received between 1978 and 1981. Inside each stratum, counties were systematically selected, and for each selected county, a number of area segments (six, eight or ten, depending on what region the state was located) was chosen. Estimation of measurement error contribution to the variance of estimates were obtained based on the measurement error model given by

\[ Y_{hji} = y_{hji} + e_{hji}, \]  \hspace{1cm} (1.4)

where \( Y_{hji} \) is the value for the variable of interest \( Y \) observed at area segment \( j \) at stratum \( h \), for category of land use \( l \). \( y_{hji} \) is the true value of the variable of interest \( Y \) associated with area segment \( hjl \), and \( e_{hji} \) are random error terms assumed to be independent with \( \text{E}_0(e_{hji}) = 0 \) and \( \text{E}_0(e_{hji}^2) = \sigma^2_{el} \). The problem of designing a sample for measurement error evaluation, addressed in Chapter 3, is motivated by the need to incorporate quality evaluation studies in future NRI surveys.

1.9 A Measurement Error Model for the NRI Measurement Process

Given the major factors described in Section 1.6 and the way they may influence the measurement process, a measurement error model is postulated. Let \( \mathcal{Q}_1 \) be a finite set of levels of experience of a data gatherer. Let \( \mathcal{Q}_2 \) be a finite set of distinct types of area segments to be observed. It should be noted that given the area segment, historical information is available,
although its recency depends on whether the area segment is part of the core panel (last year's survey) or the rotation panel (1997 survey). Define $y_k$ to be the true value of the variable of interest $y$ for the $k^{th}$ area segment. Let $Y_{jk}$ denote the observed value of the variable of interest $y$ measured at area segment $k$ by the $j^{th}$ data gatherer. Assuming a linear additive model, $Y_{jk}$ may be expressed as a function of $y_k$, given by

$$ \xi_1: Y_{jk} = y_k + D_j + e_{jk}, \quad (1.5) $$

where the $D_j$ is a random effect of the $j^{th}$ data collector and $e_{jk}$ is a random term due to the measurement error associated with the observation of the $k^{th}$ area segment, measured by the $j^{th}$ data gatherer. Model $\xi_1$ assumes the random terms are uncorrelated among themselves and between them, with

$$ D_j \sim (\mu_D, \sigma^2_D), \text{ if data gatherer } j \text{ has level of experience } g_1 \in G_1, $$

and

$$ e_{jk} \sim (0, \sigma^2), \text{ if area segment } k \text{ belongs to category } g_2 \in G_2. $$

These assumptions reflect the belief that data gatherer experience, the type of area segment and the recency historical information have an effect on the variance of the measurement error.

The set $G_2$ is assumed to be composed of four types of area segments defined by the presence of cropland and panel membership: a type 1 area segment belongs to the core part of the sample and is classified as cropland; a type 2 area segment belongs to the core part of the sample and is classified as noncropland; a type 3 area segment belongs to the rotation part of the sample and is classified as cropland, while a type 4 area segment belongs to the rotation part of the sample and is classified as noncropland. Based on these different types of area segments, the variance of the $e_{jk}$ terms is given by

$$ \text{Var}_{\xi_1}(e_{jk}) = \sigma^2_1, \text{ if segment } k \text{ is type 1} $$
\begin{align*}
= \sigma_2^2, \text{ if segment } k \text{ is type 2} \\
= \sigma_3^2, \text{ if segment } k \text{ is type 3} \\
= \sigma_4^2, \text{ if segment } k \text{ is type 4.}
\end{align*}

The investigation of sample designs conducted in Chapter 2 rely on the postulated model $\xi_1$. The sample design developed in Chapter 3 considers a simplified version of model $\xi_1$, with the assumption that $\mu_D = 0$ and $\sigma_{g1}^2 = 0$, leading to a measurement error model equivalent to $\xi_0$ used in the 1982 NRI quality evaluation study.
CHAPTER 2. Sample Design for Monitoring the Data Collection Process of Large-scale Longitudinal Surveys

2.1 Designing a Sample for Quality Monitoring Purposes

In any survey of large scale, selecting specific units for systematic checking during the data collection process is an integral part of a data quality monitoring plan. Since at this stage there is no statistical inference involved, the use of ad hoc or systematic sample designs for the inspection of the quality of the data collected is not uncommon. There are instances, however, in which the use of a probabilistic sample design is more advantageous. This is particularly the case when dealing with longitudinal surveys. In such studies, the collection of data over time offers important information that can be used to improve forthcoming surveys. It also may provide the means to design an efficient probability sample for quality monitoring. Among the advantages of using a probabilistic approach are that it is an objective way of selecting units for quality inspection and it provides a predictable degree of control over sample rates. The goal of this paper is to investigate designs for monitoring the data collection process in large-scale longitudinal surveys. In particular, the properties and feasibility of several classes of statistical sample designs will be investigated.

In large-scale longitudinal surveys, the data collection process is carried out by data gatherers who are responsible for collecting information on each one of an assigned set of sampling units. For this reason, each data gatherer can be viewed as a data 'producer' where the order in which the sampling units are observed is the 'stream' of the 'produced' data. Based on this fact, the sample designs for quality monitoring considered in this paper, are applied to each data gatherer separately.

Under the quality monitoring context, there is interest in selecting sampling units with
higher potential to contain errors at a higher rate. Characteristics of the object under obser-
vation (sample unit), the instruments and auxiliary information used to collect the data, and
the characteristics of the observer (data gatherer) all affect the propensity for measurement
errors.

With the availability of historical information, it is possible to identify sampling units that
are more likely to experience errors in data collection. The identified characteristics of the
sampling units that may explain any difference in their propensity for errors are referred to in
this paper as error risk factors.

In order to increase the chances of selecting units with higher propensity to induce errors,
first order inclusion probabilities should be related to the degree of difficulty of observing the
sampling units. This can be achieved by building inclusion probabilities that are proportional
to a function of the identified error risk factors. Such a function is typically referred to as a
'size variable' in the survey literature.

In addition to considering risk factors, it is preferable that the rule for selecting the sampling
units for monitoring should not be evident to data gatherers so that they do not change their
behavior in response to the inspection process. A design satisfying this feature will be referred
to in this paper as a design that satisfies pattern-recognition constraints.

Other factors may be important in creating a design. For example, new data gatherers are
generally more prone to making mistakes. Data quality problems may also be more frequent
at the beginning and end of the data collection process since at the beginning stage, the data
gatherers are getting familiar with the measurement process, and, at the end, survey time
constraints may lead them to make more mistakes.

The idea of analyzing a sample with respect to pattern-recognition constraints, the use
of a size variable, and the inclusion of other factors in a design, are the building blocks for
assembling and evaluating a quality monitoring sample design. In practice, the final sample
design should take into account the specific survey context. Once a set of error risk factors
has been identified, a question may arise as to whether to use factors in stratification or
in building the size variable. Sometimes, the nature of the error risk factor does not allow
to use it as a stratification variable but it is possible to use it in the size variable building process. This is the case when it is not possible to identify the value of the error risk factor associated to the sampling unit before the sampling unit itself is observed. The number of warnings issued by an editing software associated with a given sampled unit is an example of such an error risk factor. On the other hand, if an error risk factor can be used for both stratification and the size variable, the choice should be made keeping in mind that the factors used for stratification are those for which one wants to guarantee the observation of sampling units with that characteristic. However, if the intention is to increase the chance of observing a sampling unit with a given characteristic, the error risk factor defining the characteristic should be used to build a size variable.

2.2 Building First-order Inclusion Probabilities

Given a size variable, a sampling scheme should be chosen such that the resulting sample design satisfies pattern-recognition constraints. In order to formally explore the building process of a size variable and its use under a sampling scheme, the following notation is introduced. Let \( \mathcal{U} = \{1, 2, \ldots, N\} \) be the set of indices of the population units. Assume \( q \) error risk factors are identified and let \( \mathcal{X} = \{1, 2, \ldots, q\} \) be the set of their indices. Further, define \( R_{ik} > 0 \) as the value of the risk factor \( i \) variable assigned to sampling unit \( k \) in a data gatherer’s data stream.

Once the values of \( R_{ik} \) are available for every \( i \in \mathcal{I} \), and each sampling unit \( k \), a function is needed that combines these values in an appropriate way. In this paper, a multiplicative function will be used, generating a size variable \( \lambda_k \), as follows:

\[
\lambda_k = \prod_{i \in \mathcal{I}} R_{ik}.
\]  

(2.1)

Using (2.1), the magnitude of the value \( R_{ik} \) has interpretation such that if \( R_{ik} \) is twice the value of \( R_{ik'} \), \( (k \neq k') \), then the sampling unit \( k \) has twice as much weight in the size variable composition as sampling unit \( k' \).

Based on \( \lambda_k \), a scaled size variable \( p_k \), where
\[ p_k = \frac{\lambda_k}{\lambda}, \text{ where } \lambda = \sum_{k \in \mathcal{U}} \lambda_k, \] (2.2)

is defined. It should be noted that the use of \( p_k \) as a size variable is restricted by the fact that the value of \( \lambda_k \) (i.e., \( R_{ik} \)) must be known for all sampling units of the population under study before the implementation of the quality monitoring sample design. Alternatively, it is possible to use an estimate \( \hat{p}_k \) based on historical information available in longitudinal surveys.

Since the inclusion probabilities should be proportional to \( \lambda_k \), the higher the value of \( \lambda_k \), the higher the chances of selecting sampling unit \( k \) for quality inspection. The next step is to choose a sample scheme that links the values of the inclusion probabilities to the size variable \( \lambda_k \), for every \( k \in \mathcal{U} \).

### 2.3 Alternative Designs

Three sample designs are suggested as potential candidates for application: Poisson sampling with probability proportional to size (POI), systematic sampling with probability proportional to size (SYS/PPS), and one sampling unit-per-wave sampling with probability proportional to size design (1PW/PPS). A brief description of the sample schemes related to each sample design is presented next.

#### 2.3.1 Poisson Sampling (POI)

The sample scheme that generates a Poisson sampling design is a generalization of the sample scheme that generates a Bernoulli sampling design. In a Bernoulli design, each element of the population has the same first order inclusion probability while in a Poisson design, the inclusion probabilities may vary.

Let \( \pi_k \) denote the first order inclusion probability for sampling unit \( k \). Under the Poisson scheme, \( \pi_k = n_0 \hat{p}_k \) for every \( k \in \mathcal{U} \), where \( n_0 \) is the expected sample size. Further, if for each \( k \), \( \epsilon_k \) denotes an independent random variable with uniform distribution on \((0,1)\), then sampling unit \( k \) is selected for inspection if \( \epsilon_k < \pi_k \).
One advantage of the Poisson scheme is its practicality for in-stream monitoring. Its implementation can be done each time a sampling unit is available from the data collection process of a given data gatherer. Also, by its nature, this scheme is expected to generate a sample that satisfies pattern-recognition constraints. One disadvantage, however is that under the Poisson scheme, the sample size $n_s$ is a random variable. Its expectation and variance are respectively given by

$$E_p(n_s) = \sum_{k \in \mathcal{U}} \pi_k = n_0 \sum_{k \in \mathcal{U}} p_k = n_0,$$  \hspace{1cm} (2.3)

$$\text{Var}_p(n_s) = \sum_{k \in \mathcal{U}} \pi_k (1 - \pi_k) = n_0 (1 - n_0 \sum_{k \in \mathcal{U}} \hat{p}_k^2).$$  \hspace{1cm} (2.4)

Since $n_s$ is a random variable, an undesired potential side effect of this scheme is the possibility that no sample unit is selected for inspection or that too many units are selected relative to available resources. Further, no control can be exerted over the sampling process for a data gatherer.

2.3.2 Systematic Sampling with Probability Proportional to Size (SYS/PPS)

The systematic sampling with PPS scheme was originally proposed by [Madow, 1949] and modified by [Goodman and Kish, 1950]. Under a systematic with PPS sample design, $\pi_k = np_k$ for every $k \in \mathcal{U}$, where $n$ is the desired sample size. Denote a cumulative total size variable as $T_l$ such that $T_0 = 0$ and, for $l = 1, 2, \ldots N$,

$$T_l = \sum_{k=1}^{l} n_0 \hat{p}_k.$$

(2.5)

Let $d = \sum_{k \in \mathcal{U}} \hat{p}_k$ and $\epsilon$ be a single realization from an uniform distribution on $(0,d)$. Then, on the $r^{th}$ draw, sampling unit $k$ is selected for quality inspection whenever $T_{k-1} \leq \epsilon + (r - 1)d < T_k$, for $r = 1, 2, \ldots n$.

Among the advantages of this scheme are the fact that the sample is spread along the data gatherer's data stream, its implementation is simple, and pattern-recognition characteristics are likely to be better than an equal probability systematic sample. However, a disadvantage
is the fact that a sampling unit can be selected more than once, especially if the sampling rate is high or the $p_k$ vary greatly.

2.3.3 A One Unit-per-wave Sampling with Probability Proportional to Size (1PW/PPS)

In order to select a quality monitoring sample of size $n$, the data stream of a data gatherer is partitioned into $n$ waves of sizes $N_1, N_2, ..., N_n$ such that $\sum_{i=1}^{n} N_i = N$. Then, for each wave, a sample of size one is selected with probability proportional to $\lambda_k$. Under this scheme, if $W_i = \{1, 2, ..., N_i\}$ denote the set of indices for wave $i$, then $\pi_k = (\sum_{k \in W_i} \lambda_k)^{-1} \lambda_k$, for every $k \in W_i$.

One advantage of this scheme is that the sample is spread across an individual’s data stream according to the wave definitions. Also, this scheme is expected to generate a sample that satisfies pattern-recognition constraints.

2.4 Assessment of Pattern-recognition Constraints

The assessment of any sample scheme with respect to pattern-recognition constraints is an important aspect of choosing a design for the quality monitoring of a data collection process. Assume that the desired sample size for quality inspection is large enough to allow the detection of a systematic pattern. In order to objectively evaluate these situations, the first order differences of the ranks of the collected units is considered. The rank of the collected units is defined as the order in which the information regarding sampling units are available from the data collection process.

Let $d_i$ denote the difference (distance) between the ranks of the $(i + 1)^{th}$ and the $i^{th}$ sampled units $(i = 1, 2, ...n - 1)$. Then for any sample $S$ of size $n$, define

$$\Phi(S) = (n-1)^{-1} \sum_{i=1}^{n-1} d_i.$$  \hspace{1cm} (2.6)

and
\[
\Delta(S) = \sqrt{(n - 2)^{-1} \sum_{i=1}^{n-1} [d_i - \Phi(S)]^2},
\]

as the intersample mean distance and the intersample standard deviation of the distances of sample S, respectively.

The intersample mean distance may be interpreted as a measure of the average distance between sampled units if they were to be described as equally spaced apart. The intersample standard deviation is a measure of the variability of the distances between sampled units. In this paper, it is interpreted as a measure of the degree to which the pattern-recognition constraint is satisfied. For a given sample, the closer its value is to zero, the more equally spaced are the sampled points. In particular, if a given sample \( S_0 \) is composed of every \( m^{th} \) unit, \( \Delta(S_0) \) assumes value 0 and \( \Phi(S_0) \) takes on the value \( m \). Larger values of the intersample standard deviation (\( \Delta \)) may be generated by irregular sampling patterns, or by intersample distance outliers.

2.5 Sample Design Assessment for Quality Monitoring: Application to The National Resources Inventory

2.5.1 Survey Setting

A sample design assessment study was conducted to investigate the performance of the designs suggested for data quality inspection. The results of this study contributed to the formulation of the 2003 NRI quality control sample design.

The corresponding sample schemes under study were applied to the available data set from the 2001 NRI sample. At the time the study was conducted, Alaska, Washington D.C. (District of Columbia), Guam, Northern Marianas (Pacific Islands) and Caribbean presented high percentages of area segments that had not been completed and therefore, information from these states are omitted from the study.

The 2001 NRI sample data set used in this study contains information on 72,090 area segments, collected by 340 data gatherers. Approximately 44% of the data gatherers collected information from at most 100 area segments and were shorter-term employees. Only 10% of
the data gatherers in the data set collected information on more than 500 area segments (See Table 2.1). On average, each data gatherer collected information on 212 area segments, with standard deviation of 268. The minimum number of area segments collected by a data gatherer is 2 and the maximum, 2730.

<table>
<thead>
<tr>
<th>Number of segments</th>
<th>Absolute frequency</th>
<th>Relative frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-50</td>
<td>86</td>
<td>25.29</td>
</tr>
<tr>
<td>51-100</td>
<td>64</td>
<td>18.82</td>
</tr>
<tr>
<td>101-200</td>
<td>69</td>
<td>20.29</td>
</tr>
<tr>
<td>201-300</td>
<td>48</td>
<td>14.12</td>
</tr>
<tr>
<td>301-500</td>
<td>39</td>
<td>11.47</td>
</tr>
<tr>
<td>501-</td>
<td>34</td>
<td>10.00</td>
</tr>
</tbody>
</table>

The following variables are included in the record for a given area segment:

- Area segment identification;
- Data gatherer who collected on the segment;
- Prior survey experience of the data gatherer;
- Rank of data collection for the specified data gatherer (using first check-out time);
- Rotation or core sample indicator;
- Indicator of serious error detected;
- Total number of error warnings detected.

2.5.2 Determining the Size Variable

The sample designs considered by the study have selection probabilities proportional to factors identified by NRI experts as being potentially related to the presence of errors in a sampling unit.
The first error risk factor \( R_1 \) is based on information from a data checking process executed by the NRI Data Review System. A serious error is an error that indicates a major inconsistency in the data for a segment. If a serious error was issued for a given area segment, this area segment is assumed to have a higher potential to contain errors in the measurement process. It is also possible for a sampling unit to have no serious error, but several minor errors or warning messages, which is viewed as having some error potential, but probably less than if a serious error message was triggered. Based on these considerations, the following values were assigned to this factor:

\[
R_{1k} = \begin{cases} 
  2 & \text{if at least one serious error} \\
  1.5 & \text{if no serious error, but at least 10 minor errors and/or warnings} \\
  1 & \text{if neither of above apply}
\end{cases}
\]

The rotation factor \( R_2 \) is related to the type of sample unit: rotation or core segment. This distinction is important as the quality of historical imagery used in the measurement process of a rotation segment may be inferior when compared to the quality historical imagery used in the measurement process of a core segment. Based on this fact, segments belonging to the core part of the sample are assumed to have less potential to present errors. Therefore, the values assigned to this factor were:

\[
R_{2k} = \begin{cases} 
  1.5 & \text{if segment belongs to rotation panel} \\
  1 & \text{if segment belongs to core panel}
\end{cases}
\]

The timing factor \( R_3 \) assumes that the segments observed at the beginning or the end of the data collection tend to be more error-prone than the others. The reasoning behind this assertion is that in the beginning, there exists a warming up period in which data gatherers are getting familiar with the measurement task. Hence, the values assigned to this risk factor were:

\[
R_{3k} = \begin{cases} 
  2 & \text{if segment is one of the first twenty collected} \\
  1 & \text{otherwise}
\end{cases}
\]
The increased chance of error that occurs at the end of the data collection process is considered in Section 2.7.

The data gatherer experience ($R_4$) may also be a factor, assuming that the more experience a data gatherer has, the less potential there is for errors. This factor will be considered in Section 2.7.

The value of $\lambda_k$ was calculated as suggested previously: $\lambda_k = R_{1k}R_{2k}R_{3k}$.

Define $\lambda^{(j)} = \sum_{k \in U_j} \lambda_k$, where index $j$ indicates the $j^{th}$ data gatherer and $U_j$ represents the set of segments collected by data gatherer $j$. For the simulation, since data collected for each data gatherer is available, the value of the scaled size variable for unit $k$ collected by data gatherer $j$ can be calculated:

$$p_{kj} = \frac{\lambda_k}{\lambda^{(j)}} \text{ for each } k \in U_j.$$  \hfill (2.8)

The first order inclusion probabilities were calculated based on an overall expected sample fraction of 0.04 (4%), or 2884 segments for the whole sample. It should be noted that in an equal probability sample design, it would be expected to inspect 4 for every 100 segments.

Seventeen percent (17%) of the 2001 NRI data set sample units have serious errors, while fifty six percent (56%) are of the rotation type. Also, 5% of the sample units have at least 10 minor errors and/or warnings. Based on the values assigned to the error risk factors used on the size variable composition for Poisson and systematic designs, it is expected that twenty six percent (26%) of the sampling units selected for inspection contain serious errors, while sixty three percent (63%) are of the rotation type.

2.5.3 Simulation Design

In order to investigate the performance of each sample design, a Monte Carlo simulation with 100 replicates was implemented. In each Monte Carlo run, a quality monitoring sample was taken from each data gatherer's stream using the 2001 NRI data set. Using traditional notation, let $U_j = \{1, 2, ..., N_j\}$ denote the set of primary sampling units collected by data gatherer $j$. A set of size variables $\{p_{kj} : k \in U_j\}$ is defined for each data gatherer $j$ as a
function of the factors discussed on Section 2.5.2. The stream of sample units for a data gatherer is ordered by the first check out time. A Poisson sample scheme was implemented as outlined in Section 2.3.1, using the size variable \( p_k \) described in Section 2.5.2, for each data gatherer. In order to achieve the 4% overall sample fraction target, an expected sample fraction of 4% per data gatherer was set. The same was done for systematic sampling with probability proportional to size scheme (Section 2.1.2). In order to implement a sample by the proposed one unit-per-wave sampling with probability proportional to size scheme (Section 2.3.3), the population consisting of all segments collected by the data gatherer under consideration was partitioned in disjoint waves, according to the order that was used to complete data collection (see Table 2.2). In each wave, a sample of size one is selected using a probability proportional to size scheme, as discussed previously. After the 60\(^{th}\) segment, the scheme selects one among each 50 consecutive segments for inspection.

<table>
<thead>
<tr>
<th>Wave ((i))</th>
<th>Rank order of (N_i) data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>from 1 to 5</td>
</tr>
<tr>
<td>2</td>
<td>from 6 to 10</td>
</tr>
<tr>
<td>3</td>
<td>from 11 to 20</td>
</tr>
<tr>
<td>4</td>
<td>from 21 to 40</td>
</tr>
<tr>
<td>5</td>
<td>from 41 to 60</td>
</tr>
<tr>
<td>(i = 6,\ldots)</td>
<td>from 61 on</td>
</tr>
</tbody>
</table>

In each simulation run for a given design, the actual sample size, sample fraction, proportion of sampled units with serious error warning issued, proportion of sampled units of the rotation type, the intersample mean of the sampled units and the intersample standard deviation of sampled units were calculated for each data gatherer. To summarize the outcome for a simulation run for each design, basic descriptive statistics (mean, standard deviation, minimum and maximum) were calculated across data gatherers for each of these performance measures. Finally, a single set of summary statistics (average, standard deviation, minimum and maximum) across simulation runs were generated for each design.

Let \(\theta_{pj}\) be the performance measure of a given design for data gatherer \(j\) in simulation run
Let $N_f$ be the number of data gatherers for whom a quality monitoring sample was taken. For each simulation run $p$, the following summary measures were calculated:

$$\bar{\theta}_p = \frac{1}{N_D} \sum_{j=1}^{N_D} \theta_{pj};$$

$$S_p = \left( \frac{1}{N_D - 1} \sum_{j=1}^{N_D} (\theta_{pj} - \bar{\theta}_p)^2 \right)^{1/2};$$

$$m_p = \min_j \theta_{pj};$$

$$M_p = \max_j \theta_{pj}.$$

For each of these four summary measures, the simulation performance across the 100 replicates was summarized using the mean (average), the standard deviation, the minimum and the maximum descriptive statistics.

2.5.4 Performance Evaluation

2.5.4.1 Poisson Design

The mean overall sample fraction for the entire sample (across all data gatherers), was 0.04 with standard deviation 0.0007. Regarding the total sample size, on average, this number was 3071.34, with standard deviation of 48.70.

In order to investigate the variability of the observed sample size per data gatherer under the Poisson design, descriptive statistics related to the sample sizes and sample fractions observed for each data gatherer were recorded. The results are shown in Table 2.3.

It can be seen that on average, the Poisson design yielded a mean sample fraction per data gatherer of 0.06, with standard deviation 0.003. As for the sample size, on average, 10 units per data gatherer were selected for inspection. Looking at Table 2.4, it can be seen that on average, using the Poisson design resulted in approximately 24 data gatherers (out of 340) with no sample taken (sample size is zero). In addition, Table 2.5 shows that, on average, the mean of the total collected segments for those data gatherers with no sample size was approximately 28 and on average, the maximum total of collected segments was approximately 80. Also, in
Table 2.3 Simulation Results for Sample Sizes per Data Gatherer for Poisson Design

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>Mean ($\bar{\theta}_p$)</td>
<td>Average 9.59 S.D. 0.20 Min 9.13 Max 10.28</td>
</tr>
<tr>
<td>Sample</td>
<td>S.D. ($S_p$)</td>
<td>Average 11.21 S.D. 0.40 Min 10.29 Max 12.40</td>
</tr>
<tr>
<td>Size per DG</td>
<td>Min ($m_p$)</td>
<td>Average 0.00 S.D. 0.00 Min 0.00 Max 0.00</td>
</tr>
<tr>
<td>DG</td>
<td>Max ($M_p$)</td>
<td>Average 107.930 S.D. 10.82 Min 87.00 Max 137.00</td>
</tr>
<tr>
<td>Sample</td>
<td>Mean ($\bar{\theta}_p$)</td>
<td>Average 0.060 S.D. 0.003 Min 0.051 Max 0.064</td>
</tr>
<tr>
<td>Fraction per DG</td>
<td>S.D. ($S_p$)</td>
<td>Average 0.070 S.D. 0.010 Min 0.030 Max 0.100</td>
</tr>
<tr>
<td></td>
<td>Min ($m_p$)</td>
<td>Average 0.009 S.D. 0.002 Min 0.004 Max 0.015</td>
</tr>
<tr>
<td></td>
<td>Max ($M_p$)</td>
<td>Average 0.800 S.D. 0.240 Min 0.250 Max 1.000</td>
</tr>
</tbody>
</table>

Table 2.4 Number of Data Gatherers with no Sample Size Drawn

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of DG with no sample</td>
<td>Average 23.96 S.D. 4.31 Min 14 Max 39</td>
</tr>
</tbody>
</table>

at least one simulation, a maximum of 203 segments were collected by the unsampled data gatherers.

The composition of the quality monitoring samples was also evaluated. Table 2.6 shows that, on average, 28% of sampled units per data gatherer contained serious errors and 61% of the sampled units belongs to the rotation panel. These numbers are consistent with the expected values described in Section 2.5.2, and indicate oversampling with respect to population figures in Section 2.5.2.

Although the sampling fraction for the entire sample was very consistently 4%, the simulation has shown a significant number of data gatherers with no sample size drawn in practice.

Table 2.5 Total Number of Segments for Data Gatherers with no Sample Drawn

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of segments</td>
<td>Mean ($\bar{\theta}_p$)</td>
<td>Average 28.14 S.D. 4.93 Min 13.67 Max 40.18</td>
</tr>
<tr>
<td>per DG</td>
<td>Max ($M_p$)</td>
<td>Average 79.86 S.D. 26.90 Min 41 Max 203</td>
</tr>
<tr>
<td></td>
<td>Min ($m_p$)</td>
<td>Average 3.49 S.D. 2.61 Min 2 Max 17</td>
</tr>
</tbody>
</table>
This fact led to the exclusion of this sample design as a potential candidate for use in practice.

Table 2.6  Simulation Results for Poisson Design

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Per simulation summary</th>
<th>Simulation Descriptive Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least one serious error</td>
<td>Mean ($\theta_p$)</td>
<td>Average 0.28 S.D. 0.011 Min 0.25 Max 0.31</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>Average 0.22 S.D. 0.005 Min 0.21 Max 0.24</td>
</tr>
<tr>
<td>Belongs to rotation panel</td>
<td>Mean ($\theta_p$)</td>
<td>Average 0.61 S.D. 0.01 Min 0.59 Max 0.63</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>Average 0.21 S.D. 0.006 Min 0.19 Max 0.23</td>
</tr>
</tbody>
</table>

2.5.4.2 Systematic Sample with PPS Design

A total of 3064 sampling units were selected, corresponding to an overall sample fraction of 4%. On average, 9 sampling units per data gatherer were selected for inspection with a standard deviation of 10.74. This corresponds to an approximate sample fraction of 5% per data gatherer (with standard deviation 5%). The small difference with the prescribed 4% sample fraction value may be due to rounding the sample size values for the next integer for each data gatherer.

Table 2.7  Simulation Results for Systematic with PPS Design

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least one serious error</td>
<td>Mean ($\theta_p$)</td>
<td>Average 0.27 S.D. 0.01 Min 0.26 Max 0.31</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>Average 0.28 S.D. 0.008 Min 0.25 Max 0.30</td>
</tr>
<tr>
<td>Belongs to rotation panel</td>
<td>Mean ($\theta_p$)</td>
<td>Average 0.61 S.D. 0.01 Min 0.59 Max 0.63</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>Average 0.28 S.D. 0.006 Min 0.267 Max 0.294</td>
</tr>
</tbody>
</table>

Table 2.7 shows that, under the systematic with PPS design, on average, 27% of the sampled units per data gatherer contained serious error and 61% of the sampled units belongs to the rotation panel. These numbers are consistent with the expected values described in section 2.5.2. and indicate oversampling with respect to population values described in Section 2.5.2.
2.5.4.3 One Unit-per-wave Design

Altogether, a total of 2809 area segments were selected for inspection, corresponding to an overall sample fraction of 3.9%. On average, 8.26 sampling units per data gatherer were selected for inspection, with standard deviation of 5.64. This corresponds to a sample fraction of 8% per data gatherer (with standard deviation 7%). It should be noted that the wave sizes for this design were determined expecting to achieve only the overall sample fraction target of 4%. No further expectation regarding sample fraction per data gatherer was made.

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>At least one serious error</td>
<td>Mean ($\theta_p$)</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>0.23</td>
</tr>
<tr>
<td>Belongs to rotation panel</td>
<td>Mean ($\theta_p$)</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Based on Table 2.8, under the one unit-per-wave with PPS design, on average, 27% of the selected segments contained serious errors and 62% belonged to the rotation panel. These numbers are consistent with the expected values described in Section 2.5.2 and indicate oversampling with respect to the population figures in Section 2.5.2.

2.5.5 Evaluation of Pattern-recognition Constraints

The simulation performance of the Poisson, systematic with PPS and one unit-per-wave PPS designs with respect to the satisfaction of pattern-recognition contraints were also analyzed. After each Monte Carlo run, basic descriptive statistics per data gatherer related to the intersample mean distance (equation 2.7) and intersample standard deviation distance (equation 2.8) variables were recorded and summarized using the approach outlined in Section 2.5.4. The results are summarized in Table 2.9 and Table 2.10.

On average, all designs have a mean intersample mean distance around 20 segments. This means that, if the selection of units had been equally spaced apart, on average, every 20<sup>th</sup> unit would be selected for inspection.
Table 2.9 Comparison of Sample Designs Based on satisfaction of Pattern-recognition Constraints for Intersample Mean Distance (\( \Phi \))

<table>
<thead>
<tr>
<th>Per simulation summary</th>
<th>Sample Design</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Mean (( \bar{\theta}_p ))</td>
<td>POI</td>
<td>20.41</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>22.25</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>20.55</td>
</tr>
<tr>
<td>S.D. (( S_p ))</td>
<td>POI</td>
<td>11.31</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>11.06</td>
</tr>
<tr>
<td>Min (( m_p ))</td>
<td>POI</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>10.33</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>1.82</td>
</tr>
<tr>
<td>Max (( M_p ))</td>
<td>POI</td>
<td>86.64</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>30.38</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>46.85</td>
</tr>
</tbody>
</table>

The systematic with PPS design tends to have, on average, a low intersample standard deviation (5.46) per data gatherer, compared to the Poisson (18.76) and the one-per-wave (15.39) designs. For all three designs, Table 2.10 shows that, the minimum observed value among Monte Carlo replicates of the minimum intersample standard deviation per data gatherer is zero. This means that, in at least one Monte Carlo replicate, there was at least one data gatherer with a quality monitoring sample with equally spaced samples. In order to further investigate this case, the Monte Carlo simulation percentiles of the minimum value of \( \Delta \) across data gatherers was recorded. The results are shown in Table 2.11. In addition, results concerning the number of samples with \( \Delta = 0 \) in each Monte Carlo replicate was recorded. The results are shown in Table 2.12.

Table 2.11 shows that for at least one data gatherer, a quality monitoring sample with zero intersample standard deviation distance (\( \Delta = 0 \)) was generated in 50% of the Monte Carlo runs for Poisson and one-per-wave designs while this was rarer for the systematic PPS design. Information in Table 2.12, on the other hand, shows that, for any design, on average, only one to five data gatherers had the corresponding quality monitoring sample with \( \Delta = 0 \), indicating that this happens in only a few cases. Under the one unit-per-wave design, these
Table 2.10  Comparison of Sample Designs Based on Satisfaction of Pattern-recognition Constraints for Intersample Standard Deviation Distance ($\Delta$)

<table>
<thead>
<tr>
<th>Per simulation summary</th>
<th>Sample Design</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Mean ($\overline{\theta}_p$)</td>
<td>POI</td>
<td>18.76</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>5.46</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>15.39</td>
</tr>
<tr>
<td>S.D. ($S_p$)</td>
<td>POI</td>
<td>10.86</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>8.46</td>
</tr>
<tr>
<td>Min ($m_p$)</td>
<td>POI</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>0.13</td>
</tr>
<tr>
<td>Max ($M_p$)</td>
<td>POI</td>
<td>74.04</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>16.02</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>32.76</td>
</tr>
</tbody>
</table>

Table data gatherers have collected a reasonable small number of area segments (on average, 16), with correspondingly small monitoring sample sizes, increasingly the potential for $\Delta = 0$. On the other hand, under the Poisson and the systematic with PPS designs, the few data gatherers with quality monitoring sample with $\Delta = 0$ collected, on average, as many as 50 and 70 area segments, respectively. The monitoring sample sizes for these data gatherers is small, increasing the chances for $\Delta = 0$.

2.6 Summary of Simulation Results

A study consisting of a Monte Carlo simulation for each class of sample design under investigation was conducted. The sample designs were evaluated based on the objectives of furnishing a predictable degree of control over sample rates, selection of error-prone sampling units at a higher rate, and satisfaction of pattern-recognition constraints.

The results have pointed out that the Poisson with PPS design selects high risk sampling units at a higher rate while satisfying pattern-recognition constraints. However, it was evident that the Poisson design does not provide the necessary level of control over sample rates for
a quality monitoring sample, as there was no sample taken for inspection for several data gatherers at the simulation study.

Concerning the systematic with PPS sample design, the results have shown that it selects error-prone sampling units at a higher rate and under a predictable degree of control over the sample rates. However, it was clear that, compared with the other designs, the systematic sample approach has poorer performance regarding pattern-recognition constraints properties.

Finally, the results have shown that the one unit-per-wave with PPS design provides the means for selecting more error-prone units at a higher rate while satisfying pattern-recognition constraints. It also provides a predictable degree of control of sample rates as a function of the number of waves used for sampling. However, on all designs, without formal assignments of data gatherers to segments, it is not possible to calculate or even estimate \( p_k \) very easily. The one unit-per-wave design presents the best situation. However, if waves are large, a sizable number of sample units could still be caught in a “waiting” state while accumulating the full set of completed sample units, which slows down the post-processing step when conducting a survey.

Since the one-per-wave PPS design has shown satisfactory performance during this study, efforts were made to improve its performance and practicality of implementation. The resulting sample design is introduced and discussed in the next section.
Table 2.12 Number of Data Gatherers with Samples Generating $\Delta = 0$ and, the Mean and Maximum of the Total Number of Segments Collected by These Data Gatherers

<table>
<thead>
<tr>
<th>Performance measures</th>
<th>Sample Design</th>
<th>Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Runs</td>
</tr>
<tr>
<td>Number of DG samples with $\Delta = 0$</td>
<td>POI</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>79</td>
</tr>
<tr>
<td>Mean number of segments</td>
<td>POI</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>79</td>
</tr>
<tr>
<td>Max number of segments</td>
<td>POI</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>SYS/PPS</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1PW/PPS</td>
<td>79</td>
</tr>
</tbody>
</table>

2.7 A Quality Monitoring Sample Design for Large-scale Longitudinal Surveys

2.7.1 Overview

In order to improve the performance and implementation of the one-per-wave PPS approach, a stratified two-stage structure is incorporated into the design. This is done by dividing waves into smaller clusters of sampling units, and at the beginning of the wave, selecting a cluster to be held for selecting a sampling unit for quality monitoring. This will support quality monitoring sample selection without retaining a large number of sampling units belonging to a given wave.

A second change was made in order to implement different sample rates for data gatherers depending on their levels of experience (error risk factor $R_4$). The first set of waves for inexperienced data gatherers are smaller and more numerous so that initial sampling rates are higher for new data gatherers. In addition, this design included an increased selection rate (smaller waves) at the end of the data collection process.
2.7.2 One Unit-per-wave Two-stage PPS Sample (1PW2/PPS)

Assume the population of segments completed by a data gatherer is partitioned into waves such that \( N_h \) is the population size of wave \( h \). Let the population inside each wave be partitioned into clusters of a fixed number of sampling units \( N_{hc} \). Assume \( N_h = D_h N_{hc} + r \), where \( D_h \) is the number of size-\( N_{hc} \)-formed clusters inside wave \( h \) and \( r \) (\( r < N_{hc} \)) is the number of sampling units in the last cluster.

The total number of clusters in the wave is \( M_h = D_h \) or \( M_h = D_h + 1 \) depending on \( r = 0 \) or \( 0 < r < N_{hc} \), respectively. A cluster from the \( M_h \) clusters is selected with equal probability. Let \( C_{hi} \) represent the index set of sampling units composing the \( i^{th} \) selected cluster in stratum \( h \). Then, for each \( k \in C_{hi} \), select a sampling unit for inspection with probability \( \pi_{k|C_{hi}} = (\sum_{k \in C_{hi}} \lambda_k)^{-1}\lambda_k \). The first order inclusion probability of sampling unit \( k \) in stratum \( h \), under this scheme, is given by \( \pi_k = \pi_{C_{hi}} \pi_{k|C_{hi}} \), where \( \pi_{C_{hi}} \) is the probability of selecting cluster \( i \) in wave \( h \) at the first stage.

2.7.3 Simulation Design

A one unit-per-wave two-stage sample scheme was applied to each data gatherer’s stream using the same data set used in the preliminary study. The design considered was that ultimately used for the 2003 NRI quality monitoring process. A target sampling rate of 4% for the entire NRI survey was considered. The waves for each data gatherer were defined using two error risk factors, the timing of the data collection effort and data gatherer experience. The size of \( N_h \) is smaller at the beginning and end of the data collection effort to increase sampling rates for these times. In addition, waves for the first 100 segments are smaller for inexperienced data gatherers than they are for experienced data gatherers.

Waves with 10 or fewer area segments are defined to have only one cluster. Waves with more than 10 area segments are divided into clusters of size \( N_{hc} = 10 \) segments. The wave and the cluster definitions for the 2001 simulation are indicated in Table 2.13.

One area segment is selected from each wave for each data gatherer. Whenever the wave is composed of more than one cluster, a cluster is randomly selected with equal probability. Then,
Table 2.13 Waves for Each Data Gatherer (DG) by Level of Experience

<table>
<thead>
<tr>
<th>Wave</th>
<th>Segment rank</th>
<th>Number of Clusters</th>
<th>Wave</th>
<th>Segment rank</th>
<th>Number of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-5</td>
<td>1</td>
<td>1</td>
<td>1-5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6-10</td>
<td>1</td>
<td>2</td>
<td>6-10</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>11-20</td>
<td>1</td>
<td>3</td>
<td>11-15</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>21-40</td>
<td>2</td>
<td>5</td>
<td>21-30</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>41-60</td>
<td>2</td>
<td>7</td>
<td>41-50</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>61-100</td>
<td>4</td>
<td>9</td>
<td>61-80</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>101-150</td>
<td>5</td>
<td>11</td>
<td>101-150</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>151-200</td>
<td>5</td>
<td>12</td>
<td>151-200</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>201-300</td>
<td>10</td>
<td>13</td>
<td>201-300</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>301-400</td>
<td>10</td>
<td>14</td>
<td>301-400</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>801-840</td>
<td>4</td>
<td>18</td>
<td>801-840</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>841-880</td>
<td>4</td>
<td>19</td>
<td>841-880</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After 90% of all segments have been completed, wave size is 40

one area segment is selected from the sampled (or certainty) cluster using unequal selection probabilities. The unequal selection probabilities are proportional to a size variable built based on the error and rotation risk factors, with some modification of the risk functions defined in Section 2.5.3.

The value of $\lambda_{jk}$ was calculated as $\lambda_{jk} = R_{1k}R_{2k}$, where

$$R_{1k} = \begin{cases} 
2 & \text{if a serious error was detected for segment } k \\
1 & \text{if otherwise} 
\end{cases}$$

$$R_{2k} = \begin{cases} 
2 & \text{if segment } k \text{ belongs to the rotation panel} \\
1 & \text{if otherwise} 
\end{cases}$$
The conditional selection probability for segment $k$ collected by data gatherer $j$ is given by

$$\pi_{k|C_n} = \left( \sum_{k \in C_n} \lambda_{jk} \right)^{-1} \lambda_{jk}.$$ 

### 2.7.4 Performance Evaluation

A Monte Carlo simulation, with 100 replicates, was implemented in order to evaluate the one unit-per-wave two-stage sample design. In each Monte Carlo run, a sample was taken from the 2001 NRI data set, and basic descriptive statistics were generated as discussed in Section 2.5.4.

A total of 2924 area segments were selected for inspection, corresponding to an overall sample fraction of 4%. On average, 8.6 sampling units per data gatherer were selected for inspection, with standard deviation of 3.95. This corresponds to a sample fraction of 9% per data gatherer (with standard deviation 7%). For this simulation, no expectation of sample fraction per data gatherer was made, since the wave sizes were determined expecting to achieve the overall 4% of sample fraction target.

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>At least one serious error</td>
<td>Mean ($\theta_p$)</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>0.23</td>
</tr>
<tr>
<td>Belongs to rotation panel</td>
<td>Mean ($\theta_p$)</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>S.D. ($S_p$)</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Table 2.14 shows that, under the one unit-per-wave two-stage design, on average, 25% of the sampled segments contained serious errors and 65% belonged to the rotation panel. These numbers are consistent with the expected values described in Section 2.5.2. and they indicate oversampling with respect to population values described in Section 2.5.2.

### 2.7.5 Evaluation of Pattern-recognition Constraints

The performance of the one unit-per-wave two-stage design with respect to the satisfaction of pattern-recognition contraints was also analysed. After each Monte Carlo run, basic
descriptive statistics per data gatherer related to the \( \Phi \) and \( \Delta \) variables were recorded. The results are summarized in Table 2.15 and Table 2.16.

Table 2.15 Results Regarding Pattern-recognition Constraints for Intersample Mean (\( \Phi \))

<table>
<thead>
<tr>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Mean (( \bar{\theta}_p ))</td>
<td>20.31</td>
</tr>
<tr>
<td>S.D. (( S_p ))</td>
<td>14.98</td>
</tr>
<tr>
<td>Min (( m_p ))</td>
<td>1.84</td>
</tr>
<tr>
<td>Max (( M_p ))</td>
<td>82.20</td>
</tr>
</tbody>
</table>

Table 2.16 Results Regarding Pattern-recognition Constraints for Intersample Standard Deviation (\( \Delta \))

<table>
<thead>
<tr>
<th>Per simulation summary</th>
<th>Descriptive Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Mean (( \bar{\theta}_p ))</td>
<td>17.50</td>
</tr>
<tr>
<td>S.D. (( S_p ))</td>
<td>14.68</td>
</tr>
<tr>
<td>Min. (( m_p ))</td>
<td>0.13</td>
</tr>
<tr>
<td>Max. (( M_p ))</td>
<td>58.41</td>
</tr>
</tbody>
</table>

The one unit-per-wave two-stage design presented, on average, an intersample mean distance of 20, showing similar performance to the other designs.

Based on Table 2.16, the one unit-per-wave two-stage design presented, on average, intersample standard deviation of 17.50 per data gatherer, performance comparable with the Poisson design (18.76), and showing a slight improvement compared with the one-per-wave PPS design (15.39). The minimum observed value among Monte Carlo replicates of the minimum intersample standard deviation per data gatherer is zero. This means that, in at least one Monte Carlo replicate, there was at least one data gatherer's quality monitoring sample with intersample standard deviation equal to zero. Based on Table 2.17 and Table 2.18, the explanation of this event is similar to the one given for the other designs, i.e., on average, this event happens for only 2 (out of 340) data gatherers. Further, these data gatherers have collected information from, on average, only 15 area segments, with monitoring sample size small, increasing the potential for \( \Delta = 0 \).
Table 2.17 Monte Carlo Percentiles of Min. of $\Delta (S)$ for 1PW2/PPS Design

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Observed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% (Max)</td>
<td>0.71</td>
</tr>
<tr>
<td>99%</td>
<td>0.71</td>
</tr>
<tr>
<td>95%</td>
<td>0.71</td>
</tr>
<tr>
<td>90%</td>
<td>0.64</td>
</tr>
<tr>
<td>75% (Q3)</td>
<td>0</td>
</tr>
<tr>
<td>50% (Median)</td>
<td>0</td>
</tr>
<tr>
<td>25% (Q1)</td>
<td>0</td>
</tr>
<tr>
<td>10%</td>
<td>0</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
</tr>
<tr>
<td>1%</td>
<td>0</td>
</tr>
<tr>
<td>0%</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.18 Number of Data Gatherers with Samples Generating $\Delta = 0$, and the Mean and Maximum of the Total Number of Segments Collected by These Data Gatherers for 1PW2/PPS Design

<table>
<thead>
<tr>
<th>Performance measures</th>
<th>Statistics for all Simulation Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Runs</td>
</tr>
<tr>
<td>Number of DG samples with $\Delta = 0$</td>
<td>79</td>
</tr>
<tr>
<td>Mean number of segments</td>
<td>79</td>
</tr>
<tr>
<td>Max number of segments</td>
<td>79</td>
</tr>
</tbody>
</table>

2.8 Conclusions

The investigated classes of sampling designs were evaluated with respect to satisfaction of pattern-recognition constraints. The main criterion was a low value for the proposed intersample standard deviation measure. Using this criterion, the systematic with PPS sample was deemed too predictable in the sense that, under this design, it is possible to have an idea about the order in time in which sample units are selected for inspection, when compared to other designs. It should be noted, however, that an alternative type of predictability exists,
which the proposed intersample standard deviation measure is not designed to detect. This alternative type of predictability may be induced by setting extremely high values for risk factors when building first-order inclusion probabilities. In this case, because the inclusion probabilities will be very high, there will be a strong tendency to select units with the given risk factor characteristics, which may be identified as a rule of selection.

The use of previous NRI survey data in the simulation provided the necessary information to calculate the scaled size variable \( p_k \) for every unit \( k \) collected by each data gatherer. Unless similar information is available in practice, the \( p_k \) values cannot be calculated and, in order to implement designs such as the Poisson and systematic with PPS sampling, its estimation is necessary. However, for some types of error risk factors, estimation of \( p_k \) is not feasible. The risk factor “data gatherer experience”, considered in the one unit-per-wave two-stage design, is an example of such a factor. Without prior information about which sampling units are assigned to each data gatherer, it is not possible to estimate the scaled size variable. Alternatively, the effect of such risk factors may be taken into account at the design stage using stratification, as was done in the one-unit-per-wave two-stage design.

In large-scale longitudinal samples, ad hoc methods are often used for selecting a quality monitoring sample of the data collection process. However, based on the results of this paper, the availability of past information from such surveys may be used to design a more efficient probability sample design for that purpose. Such a probability sample can incorporate the ability to sample error-prone units at a higher rate while satisfying pattern-recognition constraints.

The proposed one unit-per-wave two-stage sample design provides a means of controlling the sample selection rate depending on the number of strata used. At the same time, practical implementation of this design relies upon defining clusters within each wave so that larger numbers of sampling units do not get backlogged in preparation for selection.
CHAPTER 3. A Sample Design for Measurement Error Evaluation in Large-scale Longitudinal Surveys

3.1 Introduction

In any survey, the data collected are subjected to different kinds of errors, as described in Chapter 1. This paper focuses on designing a sample for evaluating measurement error properties. In order to evaluate the contribution of measurement errors to the total survey error, studies of survey error evaluation are conducted. Such studies are classified as postsurvey evaluations since they are usually conducted after the survey data has been collected. The results of such studies serve as indicators of the data quality. The goal of this paper is to investigate optimal sample designs for measurement error evaluation in large-scale longitudinal surveys. A simple measurement error model with error variance components that vary across groups of sampling units is considered. Based on this model, an optimal sample design for estimating the contribution of measurement errors to the variance of estimates is developed. The investigation is illustrated by potential applications to the USDA’s National Resources Inventory (NRI). This design will be evaluated with an existing sample used to collect data for the survey.

3.2 Model Development

3.2.1 Introduction

Under the usual sampling scenario, let $\mathcal{U}$ denote the set of population units, of size $N$, from which a sample $\mathcal{S}$, of size $n_1$, is selected. The population is partitioned into $G$ groups according to variables that define separate groups with different measurement error variance parameters.
For the NRI application, for instance, $G = 4$: areas segments with cropland and on the core part of the sample, areas segments with cropland and on the rotation part of the sample, areas segments with no cropland and on the core part of the sample, and areas segments with no cropland and on the rotation part of the sample. Let $G = \{1, \ldots, G\}$ be the set of groups in which the population is partitioned and $U_g$ be the set of population units belonging to group $g \in G$. Then, $\mathcal{U} = \bigcup_{g \in G} U_g$, with $U_g \cap U_{g'} = \emptyset \ \forall \ g \neq g'$. Also, whenever needed, define $S_g = S \cap U_g$. Let $I_{1k}$ be the original sample $S$ membership indicator, i.e., $I_{1k} = 1$ if $k \in S$ and $I_{1k} = 0$ otherwise. Assume $S$ is taken under a pre-defined sample design $p_1$ given by

\[
p_1: \quad E_{p_1}(I_{1k}) = \pi_{1k},
E_{p_1}(I_{1k}I_{1l}) = \pi_{1kl},
E_{p_1}(I_{1k}I_{1l}I_{1m}) = \pi_{1klm},
E_{p_1}(I_{1k}I_{1l}I_{1m}I_{1n}) = \pi_{1klmn},
\]

where $E_{p_1}(.)$ represents the expected value under sample design $p_1$. In particular, $\pi_{1k}$ is the first-order sample inclusion probability of population unit $k$, $\pi_{1kl}$ is the second-order sample inclusion probability for population units $k$ and $l$, $\pi_{1klm}$ is the third-order inclusion probability for units $k$, $l$ and $m$, and $\pi_{1klmn}$ is the fourth-order inclusion probability of units $k$, $l$, $m$ and $n$. Once the sample is taken, the measurement process takes place.

For a given variable of interest $y$, let $y_k$ be the true (fixed) value of $y$ associated with population unit $k \in \mathcal{U}$. Due to inherent imperfections in the measurement process, these true values $\{y_k\}_{k \in S}$ are observed with error. Reflecting this assumption, denote by $Y_{jk}$ the $j^{th}$ observed value for unit $k$. $Y_{jk}$ measures $y_k$ with error. In this paper, the index $j$ assumes the values 1 or 2 depending on the observation being made during the original survey ($k \in S$) or during the quality evaluation survey ($k \in \mathcal{R}$, where $\mathcal{R} \subset (S)$), respectively.

It is assumed there is a parameter $\mu = N^{-1} \sum_{k \in \mathcal{U}} y_k$ of interest, estimated based on $S$ by the Horvitz-Thompson estimator ([Horvitz and Thompson, 1952]) $\hat{\mu} = N^{-1} \sum_{k \in S} Y_{1k}\pi_{1k}^{-1}$. Since the estimator uses the observations subjected to measurement error, the goal is to design the
subsample \( R \subset S \) that, along with a suitable measurement error model, allows for the estimation of the extent to which the measurement error contributes to the variance of the \( \hat{\mu} \) estimator.

### 3.2.2 A Measurement Error Model

Measurement error model A, denoted \( \xi_A \), assumes that the observed value is equal to the true value plus a measurement error. In this model, the random terms \( \omega_{1k} : k \in S_g \ \forall \ g \in G \) are uncorrelated and have a symmetric distribution about zero. Also, their variances are constant inside each group \( g \). Let the symbol \( E_{\xi_A}(\cdot) \) represent the expected value under model A assumptions. Then, model A can be specified by the following description:

\[
\xi_A : Y_{1k} = y_k + \omega_{1k},
\]

with,

\[
E_{\xi_A}(\omega_{1k}\omega_{1l}) = E_{\xi_A}(\omega_{1k})E_{\xi_A}(\omega_{1l}) = 0 \ \forall \ k \neq l \in S; \\
E_{\xi_A}(\omega_{1k}) = 0 \ \forall \ k \in S; \\
E_{\xi_A}(\omega_{1k}^2) = \sigma_g^2 \ \forall \ k \in S_g; \\
E_{\xi_A}(\omega_{1k}^4) = 0 \ \forall \ k \in S; \\
E_{\xi_A}(\omega_{1k}^4) = \eta_g \ \forall \ k \in S_g.
\]

Consider the estimation of \( \mu \) under the randomization imposed by design \( p_1 \) and the measurement error structure assumed under model A. Then, the estimator \( \hat{\mu} \) is unbiased for \( \mu \):

\[
E_{p_1\xi_A}(\hat{\mu}) = E_{p_1}[N^{-1} \sum_{k \in S} \frac{E_{\xi_A}(Y_{1k})}{\pi_{1k}}] = E_{p_1}[N^{-1} \sum_{k \in S} \frac{y_k}{\pi_{1k}}] = E_{p_1}[N^{-1} \sum_{k \in U} \frac{y_k}{\pi_{1k}}I_{1k}] = N^{-1} \sum_{k \in U} \frac{y_k}{\pi_{1k}} E_{p_1}(I_{1k})
\]
Also, its variance is decomposed in two parts, \( V_1 \) and \( V_2 \):

\[
\text{Var}_{p_1, \xi_A}(\hat{\mu}) = \text{E}_{p_1} \text{Var}_{\xi_A}(\hat{\mu}) + \text{Var}_{p_1} \text{E}_{\xi_A}(\hat{\mu}) = V_1 + V_2.
\]

The components of the estimator variance are given by

\[
V_1 = \text{E}_{p_1} \text{Var}_{\xi_A}(N^{-1} \sum_{k \in G} Y_{1k}) = N^{-2} \text{E}_{p_1} \left[ \sum_{k \in G} \frac{\text{Var}_{\xi_A}(Y_{1k})}{\pi_{1k}} + \sum_{k \in G, l \neq k \in G} \frac{\text{Cov}_{\xi_A}(Y_{1k}, Y_{1l})}{\pi_{1k} \pi_{1l}} \right]
\]

\[
= N^{-2} \text{E}_{p_1} \left( \sum_{g \in G} \sum_{k \in S_g} \frac{\sigma_g^2}{\pi_{1k}} \right) = N^{-2} \text{E}_{p_1} \left( \sum_{g \in G} \sum_{k \in U_g} \frac{\sigma_g^2}{\pi_{1k}} I_{1k} \right)
\]

\[
= N^{-2} \sum_{g \in G} \sum_{k \in U_g} \frac{\sigma_g^2}{\pi_{1k}},
\]

since \( \text{Cov}_{\xi_A}(Y_{1k}, Y_{1l}) = 0 \) for \( k \neq l \), and

\[
V_2 = \text{Var}_{p_1} (N^{-1} \sum_{k \in G} \frac{E_{\xi_A}(Y_{1k})}{\pi_{1k}}) = \text{Var}_{p_1} (N^{-1} \sum_{k \in G} \frac{Y_{1k}}{\pi_{1k}} I_{1k})
\]

\[
= N^{-2} \sum_{k \in G} \sum_{l \in U_k} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_{1k} \pi_{1l}},
\]

where \( \Delta_{1kl} = \text{Cov}_{p_1}(I_{1k}, I_{1l}) = \pi_{1kl} - \pi_{1k} \pi_{1l} \). Component \( V_2 \) is the usual Horvitz-Thompson estimator variance under no measurement error. \( V_1 \), on the other hand, is a function of the variance of the random term \( \omega_{1k} \) due to the presence of measurement error. Hence, component \( V_1 \) is regarded as the measurement error contribution to the estimator variance.

### 3.3 Estimation of the Variance Components

Unbiased estimation of \( V_1 \) and \( V_2 \) requires the use of a subsample \( R, R \subset S \). Assume \( R \) has a fixed size \( n_2 \) \( (n_2 < n_1) \) and has observations for each group \( g \in G \). Let \( R_g \) denote the set
of subsampling observations at group $g$. The problem is to design the subsample $\mathcal{R} = \bigcup_{g \in \mathcal{G}} \mathcal{R}_g$ in order to have good performance in the sense of minimizing the variance of an estimator $\hat{\theta}$ of a parameter $\theta = f(V_1, V_2)$, for a differentiable function $f$. In particular, two parameters of interest are $\theta_1 = f_1(V_1, V_2) = V_1$ and $\theta_2 = f_2(V_1, V_2) = V_1/(V_1 + V_2)$.

Let $I_{2k}$ be the subsample $\mathcal{R}$ membership indicator, i.e., $I_{2k} = 1$ if population unit $k$ is selected for subsample $\mathcal{R}$ and zero, otherwise. Let $p_2$ denote the sample design of the subsample $\mathcal{R}$, conditional on $\mathcal{S}$. The design for the subsample $p_2$ is defined by

$$
\begin{align*}
    p_2 : & \quad \mathbb{E}_{p_2}(I_{2k}|I_{1k}) = \pi_{2k}, \\
        & \quad \mathbb{E}_{p_2}(I_{2k}I_{2l}|I_{1k}I_{1l}) = \pi_{2kl}, \\
        & \quad \mathbb{E}_{p_2}(I_{2k}I_{2l}I_{2m}|I_{1k}I_{1l}I_{1m}) = \pi_{2klm}, \\
        & \quad \mathbb{E}_{p_2}(I_{2k}I_{2l}I_{2m}I_{2n}|I_{1k}I_{1l}I_{1m}I_{1n}) = \pi_{2kln}.
\end{align*}
$$

In particular, $\pi_{2k}$ is the first-order subsample conditional inclusion probability of unit $k$, given $k \in \mathcal{S}$, $\pi_{2kl}$ is the second-order subsample conditional inclusion probability of units $k$ and $l$, given $k, l \in \mathcal{S}$, $\pi_{2kln}$ is the third-order subsample conditional inclusion probability for units $k$, $l$ and $m$, and $\pi_{2klnn}$ is the fourth-order conditional subsample inclusion probability of units $k$, $l$, $m$ and $n$. Once the subsample $\mathcal{R}$ is taken, the observable values $\{Y_{2k}\}_{k \in \mathcal{R}}$ follow the same measurement error model as the sample values $\{Y_{1k}\}_{k \in \mathcal{S}}$, i.e.:

$$
\xi_A : Y_{2k} = y_k + \omega_{2k},
$$

with,

$$
\begin{align*}
    \mathbb{E}_{\xi_A}(\omega_{2k}\omega_{2l}) &= \mathbb{E}_{\xi_A}(\omega_{2k})\mathbb{E}_{\xi_A}(\omega_{2l}) = 0 \quad \forall \ k \neq l \in \mathcal{R}; \\
    \mathbb{E}_{\xi_A}(\omega_{2k}) &= 0 \quad \forall \ k \in \mathcal{R}; \\
    \mathbb{E}_{\xi_A}(\omega_{2k}^2) &= \sigma_g^2 \quad \forall \ k \in \mathcal{R}_g; \\
    \mathbb{E}_{\xi_A}(\omega_{2k}^3) &= 0 \quad \forall \ k \in \mathcal{R}; \\
    \mathbb{E}_{\xi_A}(\omega_{2k}^4) &= \eta_g \quad \forall \ k \in \mathcal{R}_g.
\end{align*}
$$
In addition, model A assumes $E_{\xi_A}(\omega_{1k}\omega_{2k}) = E_{\xi_A}(\omega_{1k})E_{\xi_A}(\omega_{2k}) = 0 \ \forall \ k \in R$. Under this setup, the sample design of the subsample $R$ is embedded on a particular case of a two-phase sample approach, in which the invariance and the independency properties are retained. Let $p$ denote this particular two-phase sample design. Then $p$ is defined by

\[
p : \quad E_p(I_{2k}) = \pi_{1k}\pi_{2k}, \quad E_p(I_{2k}I_{2l}) = \pi_{1kl}\pi_{2kl}, \quad E_p(I_{2k}I_{2l}I_{2m}) = \pi_{1kln}\pi_{2klm}, \quad E_p(I_{2k}I_{2l}I_{2m}I_{2n}) = \pi_{1klnm}\pi_{2klmn}.
\]

Also,

\[
\text{Var}_p(I_{2k}) = \text{Var}_p[I_{1k}(I_{2k}|I_{1k})] \\
= \text{Var}_{p_1}[E_p[I_{1k}(I_{2k}|I_{1k})]] + E_p_1\text{Var}_p[I_{1k}(I_{2k}|I_{1k})] \\
= \text{Var}_{p_1}[I_{1k}\pi_{2k}] + E_p_1[I_{1k}\pi_{2k}(1 - \pi_{2k})] \\
= \pi_{1k}(1 - \pi_{1k})\pi_{2k}^2 + \pi_{1k}\pi_{2k}(1 - \pi_{2k}) \\
= \pi_{1k}\pi_{2k}^2 - \pi_{1k}\pi_{2k}^2 + \pi_{1k}\pi_{2k} - \pi_{1k}\pi_{2k}^2 \\
= \pi_{1k}\pi_{2k}(1 - \pi_{1k}\pi_{2k});
\]

and

\[
\Delta_{2kl} = \text{Cov}_p(I_{2k}, I_{2l}) \\
= \text{Cov}_{p_1}\{E_{p_2}[I_{1k}(I_{2k}|I_{1k})], E_{p_2}[I_{1l}(I_{2l}|I_{1l})]\} + E_{p_1}\{\text{Cov}_{p_2}[I_{1k}(I_{2k}|I_{1k}), I_{1l}(I_{2l}|I_{1l})]\} \\
= \text{Cov}_{p_1}\{I_{1k}\pi_{2k}, I_{1l}\pi_{2l}\} + E_{p_1}\{I_{1k}I_{1l}(\pi_{2kl} - \pi_{2k}\pi_{2l})\} \\
= \pi_{2k}\pi_{2l}(\pi_{1kl} - \pi_{1k}\pi_{1l}) + \pi_{1kl}(\pi_{2kl} - \pi_{2k}\pi_{2l}) \\
= \pi_{1kl}\pi_{2kl} - \pi_{1k}\pi_{2k}\pi_{1l}\pi_{2l}.
\]

The following result presents an unbiased estimator for $\theta_1 = \nu_1$ and its variance.
Result 3.1 Under the randomization imposed by sample design \( p \) and measurement error
model \( A \),

\[
\hat{\theta}_1 = (2N^2)^{-1} \sum_{k \in R} \frac{(Y_{1k} - Y_{2k})^2}{\pi_{1k}^2 \pi_{2k}^2}
\]

is an unbiased estimator of \( \theta_1 \) and its variance has the form

\[
\text{Var}_p(\hat{\theta}_1) = (2N^4)^{-1} \sum_{g \in G} \sum_{k \in U_g} \frac{\eta_g + \sigma_g^4}{\pi_{1k}^3 \pi_{2k}^2} + N^{-4} \sum_{g \in G} \sum_{k \in U_g} \sum_{i \in U_g} \sum_{j \in U_g} \frac{\Delta_{2k} \sigma_g^2 \sigma_g^2}{\pi_{1k}^2 \pi_{2k}^2 \pi_{1l}^2 \pi_{2l}^2}.
\]

Proof:

\[
\text{E}_p \text{Var}_A(\hat{\theta}_1) = \text{E}_p \left[ (2N^2)^{-1} \sum_{g \in G} \sum_{k \in U_g} \text{E}_A \left( \frac{(Y_{1k} - Y_{2k})^2}{\pi_{1k}^2 \pi_{2k}^2} \right) \right]
\]

\[
= \text{E}_p \left[ N^{-2} \sum_{g \in G} \sum_{k \in U_g} \frac{\sigma_g^2}{\pi_{1k}^2 \pi_{2k}^2} I_{2k} \right]
\]

\[
= N^{-2} \sum_{g \in G} \sum_{k \in U_g} \frac{\sigma_g^2}{\pi_{1k}^2} = \theta_1.
\]

\[
\text{Var}_p(\hat{\theta}_1) = \text{E}_p \text{Var}_A(\hat{\theta}_1) + \text{Var}_p \text{E}_A(\hat{\theta}_1).
\]

\[
\text{E}_p \text{Var}_A(\hat{\theta}_1) = \text{E}_p \left[ (4N^4)^{-1} \sum_{g \in G} \sum_{k \in U_g} \frac{\text{Var}_A \left( \frac{(Y_{1k} - Y_{2k})^2}{\pi_{1k}^2 \pi_{2k}^2} \right)}{\pi_{4k}^2 \pi_{2k}^2} \right]
\]

\[
= \text{E}_p \left[ (2N^4)^{-1} \sum_{g \in G} \sum_{k \in U_g} \frac{(\eta_g + \sigma_g^4)}{\pi_{1k}^4 \pi_{2k}^2} \right]
\]

\[
= \text{E}_p \left[ (2N^4)^{-1} \sum_{g \in G} \sum_{k \in U_g} \frac{(\eta_g + \sigma_g^4)}{\pi_{1k}^4 \pi_{2k}^2} I_{2k} \right]
\]

\[
= (2N^4)^{-1} \sum_{g \in G} \sum_{k \in U_g} \frac{(\eta_g + \sigma_g^4)}{\pi_{1k}^4 \pi_{2k}^2},
\]

since

\[
\text{Var}_A(Y_{1k} - Y_{2k})^2 = \text{Var}_A(\omega_{1k} - \omega_{2k})^2
\]

\[
= \text{Var}_A(\omega_{1k}^2) + 4 \text{Var}_A(\omega_{1k} \omega_{2k}) + \text{Var}_A(\omega_{2k}^2)
\]

\[
= (\eta_g - \sigma_g^4) + 4 \sigma_g^4 + (\eta_g - \sigma_g^4) = 2(\eta_g + \sigma_g^4).
\]
\[ \text{Var}_p E_{\xi_A}(\hat{\theta}_1) = \text{Var}_p(2N^{-2} \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \frac{E_{\xi_A}(Y_{1k} - Y_{2k})^2}{\pi_{1k}^2 \pi_{2k}}) \]

\[ = N^{-4} \text{Var}_p \left( \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \frac{\sigma_g^2}{\pi_{1k}^2 \pi_{2k}} I_{2k} \right) \]

\[ = N^{-4} \sum_{g \in \mathcal{G}} \sum_{g' \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \sum_{l \in \mathcal{R}_{g'}} \frac{\Delta_{2k} \sigma_g^2 \sigma_{l_k}^2}{\pi_{1k}^2 \pi_{2k} \pi_{1l}^2 \pi_{2l}}. \]

Under the same two-phase sample design \( p \), consider estimating \( \theta_2 = V_1/(V_1 + V_2) \). For this goal, the following series of results are useful. First, an unbiased estimator for \( V_2 \) is introduced:

**Result 3.2** Under the randomization imposed by sample design \( p \) and measurement error model \( A \),

\[ \hat{V}_2 = N^{-2} \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_{1kl} \pi_{1k} \pi_{1l}} - N^{-2} \sum_{k \in \mathcal{R}} \frac{\Delta_{1kk} (Y_{1k} - Y_{2k})^2}{2} \]

is an unbiased estimator of \( V_2 \).

**Proof:**

Under measurement error model \( A \), the standard estimator of the Horvitz-Thompson sample variance, \( \hat{V}_2 = N^{-2} \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} \frac{\Delta_{1kl} y_{kl} y_{kl}}{\pi_{1kl} \pi_{1k} \pi_{1l}} \), is biased:

\[ E_{p_{\xi_A}}(\hat{V}_2) = N^{-2} E_{p_{\xi_A}} \left[ \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} \frac{\Delta_{1kk} (y_k + \omega_{1k})^2}{\pi_{1k}^2} + \sum_{k \in \mathcal{R}} \sum_{l \notin \mathcal{S}} \frac{\Delta_{1kl} (y_k + \omega_{1k})(y_l + \omega_{1l})}{\pi_{1kl} \pi_{1k} \pi_{1l}} \right] \]

\[ = N^{-2} E_{p} \left[ \sum_{k \in \mathcal{S}} \frac{\Delta_{1kk} E_{\xi_A}(y_k + \omega_{1k})^2}{\pi_{1k}^2} + \sum_{k \in \mathcal{R}} \sum_{l \notin \mathcal{S}} \frac{\Delta_{1kl} E_{\xi_A}(y_k + \omega_{1k})(y_l + \omega_{1l})}{\pi_{1kl} \pi_{1k} \pi_{1l}} \right] \]

\[ = N^{-2} E_{p} \left[ \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \frac{\Delta_{1kk} y_k y_l}{\pi_{1k}^2 \pi_{1k}} + \sum_{k \in \mathcal{R}} \sum_{l \notin \mathcal{S}} \frac{\Delta_{1kl} y_k y_l}{\pi_{1kl} \pi_{1k} \pi_{1l}} \right] \]

\[ = N^{-2} E_{p} \left[ \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \sum_{l \in \mathcal{R}_{g'}} \frac{\Delta_{1kl} y_k y_l I_{1k} I_{1l}}{\pi_{1k} \pi_{1k} \pi_{1l}} + \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \frac{\Delta_{1kk} \sigma_g^2}{\pi_{1k}^2} \right] \]

\[ = V_2 + N^{-2} \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{R}_g} \frac{\Delta_{1kk} \sigma_g^2}{\pi_{1k}^2}. \]

On the other hand,
\[ E_{p\xi_A}\left[N^{-2} \sum_{g \in G} \sum_{k \in \mathcal{R}_g} \frac{\Delta_{1kk}}{\pi_1^3 k \pi_{2k}} \left( Y_{1k} - Y_{2k} \right)^2 \right] = E_p\left[N^{-2} \sum_{g \in G} \sum_{k \in \mathcal{R}_g} \frac{\Delta_{1kk}}{\pi_1^3 k \pi_{2k}} \frac{E_{\xi_A}(Y_{1k} - Y_{2k})^2}{2} \right] = N^{-2} \sum_{g \in G} \sum_{k \in \mathcal{R}_g} \frac{\Delta_{1kk}}{\pi_1^2} \sigma_g^2. \]

Therefore, \( \hat{V}_2 \) is indeed unbiased.

\[ \text{Lemma 3.1} \text{ Let } \hat{V}_{HT} = \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} \frac{\Delta_{1kl}}{\pi_{1k} \pi_{1l}} \text{ denote the standard Horvitz-Thompson estimator of the variance of } \hat{\mu} \text{ under no measurement error. Its variance, with respect to the design } p, \text{ is given by} \]

\[ \text{Var}_p(\hat{V}_{HT}) = \text{Var}_p(N^{-2} \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} a_{1kl}) \]

\[ = N^{-4}\left[2 \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} a_{1kl}^2 \pi_{1kl}(1 - \pi_{1kl}) + 4 \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \sum_{m \neq k \in \mathcal{U}} a_{1kl} a_{1km} (\pi_{1klm} - \pi_{1kl} \pi_{1km}) + \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \sum_{m \neq k \in \mathcal{U}} \sum_{n \neq k \in \mathcal{U}} a_{1kl} a_{1mn} (\pi_{1klmn} - \pi_{1kl} \pi_{1mn}) \right]. \]

where \( a_{1kl} = \frac{\Delta_{1kl}}{\pi_{1kl} \pi_{1l}}. \)

\[ \text{Proof:} \]

\[ \text{Var}_p(\hat{V}_{HT}) = \text{Var}_p(N^{-2} \sum_{k \in \mathcal{S}} \sum_{l \in \mathcal{S}} a_{1kl}) \]

\[ = N^{-4}\text{Var}_p(\sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} a_{1kl} I_{1k} I_{1l}) \]

\[ = N^{-4}\sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \sum_{m \in \mathcal{U}} \sum_{n \in \mathcal{U}} a_{1kl} a_{1mn} \text{Cov}_p(I_{1k} I_{1l}, I_{1m} I_{1n}) \]

\[ = N^{-4}\left[2 \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} a_{1kl}^2 \text{Var}_p(I_{1k} I_{1l}) + 4 \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \sum_{m \neq k \in \mathcal{U}} a_{1kl} a_{1km} \text{Cov}_p(I_{1k} I_{1l}, I_{1k} I_{1m}) + \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \sum_{m \neq k \in \mathcal{U}} \sum_{n \neq k \in \mathcal{U}} a_{1kl} a_{1mn} \text{Cov}_p(I_{1k} I_{1l}, I_{1m} I_{1n}) \right]. \]
Since

\[ \text{Var}_p(I_{1k}I_{1l}) = \pi_{1kl}(1 - \pi_{1kl}), \]

\[ \text{Cov}_p(I_{1k}I_{1l}, I_{1k}I_{1m}) = \pi_{1klm} - \pi_{1kl}\pi_{1km}, \]

and

\[ \text{Cov}_p(I_{1k}I_{1l}, I_{1m}I_{1n}) = \pi_{1kln} - \pi_{1kl}\pi_{1mn}, \]

the desired result holds.

The notation established for this Lemma is retained, whenever convenient. The next result introduces the variance of \( \hat{V}_2 \).

**Result 3.3** Under the randomization imposed by sample design \( p \) and measurement error model \( A \), the variance of \( \hat{V}_2 \) is given by

\[
\text{Var}_{p,\xi_A}(\hat{V}_2) = E_p \text{Var}_{\xi_A}(\hat{V}_2) + \text{Var}_p E_{\xi_A}(\hat{V}_2) \\
= N^{-1} \left[ 2 \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{1kk}^2 (4y_k^2 \sigma_g^2 + \eta_g - \sigma_\xi^2)}{\pi_{1kl}} \right] \\
+ 2 \sum_{g \in G} \sum_{g' \in G} \sum_{k \in U_{g'}} \frac{\Delta_{1kl}^2 (y_k^2 \sigma_g^2 + \eta_g - \sigma_\xi^2)}{\pi_{1kl} \pi_{1kl}^2 \pi_{1l}} \\
+ 8 \sum_{g \in G} \sum_{k \in U_g} \sum_{m \neq k \in U_g} \frac{\Delta_{1kk} \Delta_{1km} \Delta_{1l} \Delta_{1lm} \eta_m \sigma_g^2}{\pi_{1kl}^2 \pi_{1km} \pi_{1lm} \pi_{1klm} \pi_{1lm}} \\
+ 4 \sum_{g \in G} \sum_{g' \in G} \sum_{k \in U_{g'}} \sum_{m \neq k \in U_{g'}} \frac{\Delta_{1kk} \Delta_{1kl} \Delta_{1km} \Delta_{1l} \Delta_{1lm} \eta_m \sigma_g^2}{2 \pi_{1kl}^2 \pi_{2k}^2} \\
+ \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{1kk}^2 (1 - 2\pi_{2k}) + \sigma_\xi^2 (1 + 2\pi_{2k})}{2 \pi_{1kl}^2 \pi_{2k}^2} \\
+ 2 \sum_{k \in U} \sum_{l \in U} \left( \frac{\Delta_{1kl} y_k y_l \eta_l \eta_k}{\pi_{1kl} \pi_{1k} \pi_{1l}} \right)^2 \pi_{1kl}(1 - \pi_{1kl}) \\
+ 4 \sum_{k \in U} \sum_{l \in U} \sum_{m \neq k \in U} \frac{\Delta_{1kl} \Delta_{1km} \Delta_{1l} \Delta_{1lm} \eta_m}{\pi_{1kl} \pi_{1km} \pi_{1lm} \pi_{1klm}} \left( \frac{y_k}{\pi_{1kl}} \right)^2 \frac{y_l \eta_l \eta_m \eta_k}{\pi_{1l} \pi_{1m} \pi_{1klm}} (\pi_{1klm} - \pi_{1klm}) \\
+ \sum_{k \in U} \sum_{l \in U} \sum_{m \neq k \in U} \sum_{n \neq l \in U} \frac{\Delta_{1kl} \Delta_{1km} \Delta_{1ln} \Delta_{1ln} \eta_m \eta_l \eta_k}{\pi_{1kl} \pi_{1km} \pi_{1ln} \pi_{1l} \pi_{1m} \pi_{1ln} \pi_{1klm}} (\pi_{1klmn} - \pi_{1klmn}) \right].
\]
Proof:

\[ E_p \text{Var}_{\xi_A}(\tilde{V}_2) = N^{-4} \left[ E_p \text{Var}_{\xi_A} \left( \sum_{k, l \in S} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_{1kl} \pi_{1k} \pi_{1l}} - \sum_{k \in R} \frac{\Delta_{1kk} (Y_{1k} - Y_{2k})^2}{2} \right) \right] \]

Using the same notation introduced earlier, let \( \tilde{V}_2 = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_{1kl} \pi_{1k} \pi_{1l}} \). In addition, let \( b_k = \frac{\Delta_{1kk} (Y_{1k} - Y_{2k})^2}{\pi_{1k} \pi_{2k}} \) and \( b = \sum_{k \in R} b_k \). Then,

\[ E_p \text{Var}_{\xi_A}(\tilde{V}_2) = N^{-4} \left[ E_p \text{Var}_{\xi_A}(\tilde{V}_2 - b) \right] \]

\[ = N^{-4} \left[ E_p \text{Var}_{\xi_A}(\tilde{V}_2) + E_p \text{Var}_{\xi_A}(b) - 2E_p \text{Cov}_{\xi_A}(\tilde{V}_2, b) \right]. \]

Now, replacing the variance with respect to the design \( p \) by the variance with respect to the model \( \xi_A \) on the proof of Lemma 3.1, gives

\[ E_p \text{Var}_{\xi_A}(\tilde{V}_2) = E_p \left[ 2 \sum_{k \in S} \sum_{l \in S} \left( \frac{\Delta_{1kl}}{\pi_{1kl} \pi_{1k} \pi_{1l}} \right)^2 \text{Var}_{\xi_A}(Y_{1k} Y_{1l}) \right] \]

\[ + 4 \sum_{k \in S} \sum_{l \in S} \sum_{m \neq k, l \in S} \frac{\Delta_{1kl} \Delta_{1km}}{\pi_{1kl} \pi_{1km} \pi_{1l} \pi_{1m}} \text{Cov}_{\xi_A}(Y_{1k} Y_{1l}, Y_{1k} Y_{1m}) \]

\[ + \sum_{k \in S} \sum_{l \in S} \sum_{m \neq k, l \in S} \sum_{n \neq k, l \in S} \frac{\Delta_{1kl} \Delta_{1mn}}{\pi_{1kl} \pi_{1mn} \pi_{1l} \pi_{1m} \pi_{1n}} \text{Cov}_{\xi_A}(Y_{1k} Y_{1l}, Y_{1m} Y_{1n}). \]

Since

\[ \text{Var}_{\xi_A}(Y_{1k} Y_{1l}) = \text{Var}_{\xi_A} [(y_k + \omega_k)(y_l + \omega_l)] \]

\[ = \text{Var}_{\xi_A} [y_k y_l + y_k \omega_l + y_l \omega_k + \omega_k \omega_l] \]

\[ = y_k^2 \sigma_g^2 + y_l^2 \sigma_g^2 + \sigma_g^2 \sigma_g^2, \text{ if } k \neq l, \text{ and} \]

\[ = 4y_k^2 \sigma_g^2 + \eta_g - \sigma_g^4, \text{ if } k = l, \]

\[ \text{Cov}_{\xi_A}(Y_{1k} Y_{1l}, Y_{1k} Y_{1m}) = \text{Cov}_{\xi_A} [(y_k + \omega_k)(y_l + \omega_l), (y_k + \omega_k)(y_m + \omega_m)] \]

\[ = y_l y_m \text{Var}_{\xi_A}(\omega_{1k}) = y_l y_m \sigma_g^2 \text{ if } k \neq l \text{ (and } m \neq k, l), \text{ and} \]

\[ = 2y_m \sigma_g^2, \text{ if } k = l \text{ (and } m \neq k, l), \]
and

$$\text{Cov}_{\xi_A}(Y_{1k}Y_{1l}, Y_{1m}Y_{1n}) = 0, \text{ for } m \neq k, l \text{ and } n \neq k, l,$$

then,

$$\text{EpVar}_{\xi_A}(\tilde{V}_2) = \text{Ep} \left[ 2 \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{kk}^2}{\pi_{1k}^2} (4y^2_g \sigma_g^2 + \eta_g - \sigma_g^4) I_{1k} \right]$$

$$+ 2 \sum_{g \in G} \sum_{g' \in G} \sum_{k \in U_g} \sum_{l \neq k \in U_{g'}} \left( \frac{\Delta_{kl}}{\pi_{1k} \pi_{1l} \pi_{1l}} \right)^2 (y^2_{g'} \sigma_{g'}^2 + \eta_{g'} \sigma_{g'}^2 + \sigma_{g'}^2 \sigma_{g}^2) I_{1k} I_{1l}$$

$$+ 8 \sum_{g \in G} \sum_{k \in U_g} \sum_{m \neq k \in U_g} \frac{\Delta_{kk}}{\pi_{1k}^4} \frac{\Delta_{km}}{\pi_{1km} \pi_{1m} \pi_{1m}} y_m \sigma_g^2 I_{1k} I_{1m}$$

$$+ 4 \sum_{g \in G} \sum_{g' \in G} \sum_{g'' \in G} \sum_{k \in U_g} \sum_{l \neq k \in U_{g'}} \sum_{m \neq k \in U_{g''}} \frac{\Delta_{kl}}{\pi_{1kl} \pi_{1km} \pi_{1k} \pi_{1l} \pi_{1m} \pi_{1m}} y_m y_{g''} \sigma_g^2 I_{1k} I_{1l} I_{1m}$$

$$= \left[ 2 \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{kk}^2}{\pi_{1k}^2} (4y^2_g \sigma_g^2 + \eta_g - \sigma_g^4) \right]$$

$$+ 2 \sum_{g \in G} \sum_{g' \in G} \sum_{k \in U_g} \sum_{l \neq k \in U_{g'}} \frac{\Delta_{kl}^2}{\pi_{1kl}^2 \pi_{1l}^2} (y^2_{g'} \sigma_{g'}^2 + \eta_{g'} \sigma_{g'}^2 + \sigma_{g'}^2 \sigma_{g}^2)$$

$$+ 8 \sum_{g \in G} \sum_{k \in U_g} \sum_{m \neq k \in U_g} \frac{\Delta_{kk}}{\pi_{1k}^4} \frac{\Delta_{km}}{\pi_{1km} \pi_{1m}^2} y_m \sigma_g^2$$

$$+ 4 \sum_{g \in G} \sum_{g' \in G} \sum_{g'' \in G} \sum_{k \in U_g} \sum_{l \neq k \in U_{g'}} \sum_{m \neq k \in U_{g''}} \frac{\Delta_{kl} \Delta_{km}}{\pi_{1kl}^2 \pi_{1km} \pi_{1l} \pi_{1m}^2} y_m y_{g''} \sigma_g^2 \right].$$

Next,

$$\text{EpVar}_{\xi_A}(b) = \text{Ep} \left[ \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{kk}^2}{\pi_{1k}^2 \pi_{1k}^2} \frac{1}{4} \var{\xi_A}{(Y_{1k} - Y_{2k})^2} \right]$$

$$+ \sum_{k \in U, l \neq k \in U} \frac{\Delta_{kk} \Delta_{ll}}{\pi_{1k} \pi_{1k} \pi_{1l} \pi_{1l}} \frac{1}{4} \text{Cov}_{\xi_A}\left[(Y_{1k} - Y_{2k})^2, (Y_{1l} - Y_{2l})^2\right]$$

$$= \text{Ep} \left[ \sum_{k \in U} \frac{\Delta_{kk}^2}{\pi_{1k}^3 \pi_{1k}^2} \frac{1}{4} \left( \var{\xi_A}{(\omega_{1k}^2)} + \var{\xi_A}{(2\omega_{1k} \omega_{2k})} + \var{\xi_A}{(\omega_{2k}^2)} \right) \right]$$

$$= \text{Ep} \left[ \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{kk}^2}{\pi_{1k}^3 \pi_{1k}^2} \frac{(\eta_g + \sigma_g^4)}{2} I_{2k} \right]$$

$$= \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{kk}^2}{\pi_{1k} \pi_{1k}^2} \frac{\eta_g + \sigma_g^4}{2}.$$
as $\text{Cov}_\xi [(Y_{1k} - Y_{2k})^2, (Y_{1l} - Y_{2l})^2] = 0$ for $k \neq l$.

Finally,

$$2E_p \text{Cov}_\xi (\bar{V}_2, b) = 2E_p \left[ \sum_{k \in S} \sum_{l \in S} \sum_{k' \in R} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_{1kl} \pi_{1l}} \sum_{k'' \in R} \frac{\Delta_{1k''}^l (Y_{1k''} - Y_{2k''})^2}{2} \right]$$

$$= 2E_p \left[ \sum_{k \in R} \sum_{l \in R} \sum_{k' \in R} \frac{\Delta_{1kl} \Delta_{1k'} Y_{1k} Y_{1l} \text{Cov}_\xi (Y_{1k} Y_{1l}, (Y_{1k'} - Y_{2k'})^2)}{2\pi_{1kl} \pi_{1l}} \right].$$

The covariance term above can be expressed as follows:

$$\text{Cov}_\xi (Y_{1k} Y_{1l}, (Y_{1k} - Y_{2k})^2) = \text{Cov}_\xi [(y_k + \omega_{1k})(y_l + \omega_{1l}), (\omega_{1k'} - \omega_{2k'})^2]$$

$$= \text{Cov}_\xi (y_k y_l + y_k \omega_{1l} + y_l \omega_{1k} + \omega_{1k} \omega_{1l}, \omega_{1k'}^2 + \omega_{2k'}^2 + 2\omega_{1k'} \omega_{2k'})$$

$$= y_k [\text{Cov}_\xi (\omega_{1l}, \omega_{1k'}) + \text{Cov}_\xi (\omega_{1l}, \omega_{2k'}) + 2\text{Cov}_\xi (\omega_{1l}, \omega_{1k'} \omega_{2k'})]$$

$$+ y_l [\text{Cov}_\xi (\omega_{1k}, \omega_{1k'}) + \text{Cov}_\xi (\omega_{1k}, \omega_{2k'}) + 2\text{Cov}_\xi (\omega_{1k}, \omega_{1k'} \omega_{2k'})]$$

$$+ \text{Cov}_\xi (\omega_{1k} \omega_{1l}, \omega_{1k'}^2) + \text{Cov}_\xi (\omega_{1k} \omega_{1l}, \omega_{2k'}^2) + 2\text{Cov}_\xi (\omega_{1k} \omega_{1l}, \omega_{1k'} \omega_{2k'}).$$

For $k \neq l$ and $k = k'$, or, $k \neq l$ and $k \neq k'$, the covariance term above is zero. However, for $k = l$ and $k = k'$, this covariance is $(\eta_g - \sigma_g^4)$. Therefore,

$$2E_p \text{Cov}_\xi (\bar{V}_2, b) = 2E_p \left[ \sum_{g \in G} \sum_{k \in R_g} \frac{\Delta_{1kk}^l (\eta_g - \sigma_g^4)}{\pi_{1kk}^6 \pi_{2k}^2} \right]$$

$$= E_p \left[ \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{1kk}^l (\eta_g - \sigma_g^4)}{\pi_{1k}^6 \pi_{2k}^2} I_{2k} \right]$$

$$= \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{1kk}^l (\eta_g - \sigma_g^4)}{\pi_{1k}^6 \pi_{2k}^2}.$$

Hence,

$$E_p \text{Var}_\xi (b) - 2E_p \text{Cov}_\xi (\bar{V}_2, b) = \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{1kk}^l}{2\pi_{1k}^2 \pi_{2k}^2} [\eta_g (1 - 2\pi_{2k}) + \sigma_g^4 (1 + 2\pi_{2k})].$$

The last term follows from Lemma 3.3:
\[
\text{Var}_p E_{\xi A}(\hat{V}_2) = \text{Var}_p \left( N^{-2} \sum_{k \in U} \sum_{l \in U} \frac{\Delta_{1kl} \frac{y_k}{\pi_{1kl}} \frac{y_l}{\pi_{1kl}} I_{1k} I_{1l}}{\pi_{1kl} \pi_{11}} \right)
\]
\[
= N^{-4} \sum_{k \in U} \sum_{l \in U} \left( \frac{\Delta_{1kl} \frac{y_k}{\pi_{1kl}} \frac{y_l}{\pi_{1kl}}}{{\pi_{1kl} \pi_{11}^2}} \right)^2 \pi_{1kl} (1 - \pi_{1kl})
\]
\[
+ N^{-4} \sum_{k \in U} \sum_{l \in U} \sum_{m \neq k, l \in U} \frac{\Delta_{1kl} \Delta_{1km} \frac{y_k}{\pi_{1kl}} \frac{y_l}{\pi_{1lm}} \left( \frac{y_m}{\pi_{1lm}} \right)^2 (\pi_{1km} - \pi_{1kl} \pi_{1km})}{\pi_{1kl} \pi_{1km} \pi_{11} \pi_{1lm} \pi_{1ln}} \pi_{1km} - \pi_{1kl} \pi_{1mm})
\]

\[\boxed{\text{Result 3.4}}\]
Under the randomization imposed by sample design \( p \) and measurement error model \( A \), the covariance between \( \hat{V}_1 \) and \( \hat{V}_2 \) is given by

\[
\text{Cov}_{p\xi A}(\hat{V}_1, \hat{V}_2) = N^{-4} \left( \sum_{g \in G} \sum_{k \in U_g} \sum_{l \in U_g} \frac{\Delta_{1kl} \Delta_{1kk} \eta_g + \sigma_g^2}{\pi_{11} \pi_{12k}} \left( \eta_g + \sigma_g^2 \right) \right) (3.3)
\]
\[
- \sum_{g \in G} \sum_{k \in U_g} \frac{\Delta_{1kk} \eta_g + \sigma_g^2}{\pi_{11} \pi_{12k}} (3.4)
\]
\[
+ \sum_{g \in G} \sum_{k \in U_g} \sum_{l \in U_g} \frac{\Delta_{1kl} \Delta_{1km} \frac{y_k}{\pi_{1kl}} \frac{y_l}{\pi_{1lm}} \sigma_g^2 \pi_{1kl} \pi_{12k} (1 - \pi_{1kl} \pi_{12k})}{\pi_{1kl} \pi_{1km} \pi_{11} \pi_{1lm} \pi_{1ln}} \right] (3.5)
\]

Proof:

\[
\text{Cov}_{p\xi A}(\hat{V}_1, \hat{V}_2) = \text{E}_p \text{Cov}_{\xi A}(\hat{V}_1, \hat{V}_2) + \text{Cov}_p [E_{\xi A}(\hat{V}_1), E_{\xi A}(\hat{V}_2)]
\]
\[
= \text{E}_p \text{Cov}_{\xi A} \left[ (2N^2)^{-1} \sum_{k \in R} \frac{(Y_{1k} - Y_{2k})^2}{\pi_{1k} \pi_{2k}}, N^{-2} \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_{1kl} \pi_{11}} \right]
\]
\[
- \text{E}_p \text{Cov}_{\xi A} \left[ (2N^2)^{-1} \sum_{k \in R} \frac{(Y_{1k} - Y_{2k})^2}{\pi_{1k} \pi_{2k}}, N^{-2} \sum_{k \in R} \frac{\Delta_{1kk} (Y_{1k} - Y_{2k})^2}{\pi_{1k} \pi_{2k}} \right]
\]
\[
+ \text{Cov}_p \left[ N^{-2} \sum_{g \in G} \sum_{k \in U_g} \frac{\sigma_g^2}{\pi_{1k} \pi_{2k}} , N^{-2} \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{1kl} \frac{y_k}{\pi_{1kl}} \frac{y_l}{\pi_{1kl}}}{\pi_{1kl} \pi_{11}} \right]
\]

Earlier, it was shown that
Next,

$$
E_p \text{Cov}_{\xi_A} \left[ (2N^2)^{-1} \sum_{k \in R} \frac{(Y_{1k} - Y_{2k})^2}{\pi_1^2 \pi_2^2}, N^{-2} \sum_{k \in R} \sum_{i \in S} \frac{\Delta_{1kl} Y_{1k} Y_{1l}}{\pi_1 \pi_{1k} \pi_{1l}} \right] =
$$

$$
N^{-4} \sum_{g \in G} \sum_{k \in U_g} \sum_{i \in U_g} \frac{\Delta_{1kl} \Delta_{1kk} \left( \eta_{g} + \sigma_{g}^2 \right)}{\pi_1^4 \pi_1 \pi_{1k} \pi_{2k}}.
$$

The last step is

$$
\text{Cov}_p \left[ N^{-2} \sum_{g \in G} \sum_{k \in R_g} \frac{\sigma_{g}^2}{\pi_1 \pi_2}, N^{-2} \sum_{k \in R} \sum_{i \in S} \frac{\Delta_{1kl} y_k y_l}{\pi_1 \pi_{1k} \pi_{1l}} \right] =
$$

$$
= \text{Cov}_p \left[ N^{-2} \sum_{g \in G} \sum_{k \in R_g} \frac{\sigma_{g}^2}{\pi_1 \pi_2}, N^{-2} \sum_{k \in R} \sum_{i \in S} \frac{\Delta_{1kl} y_k y_l}{\pi_1 \pi_{1k} \pi_{1l}} \right]
$$

$$
= N^{-4} \sum_{g \in G} \sum_{k \in U_g} \sum_{l \in U_g} \frac{\Delta_{1kl} y_k y_l \pi_1 \pi_{1k} \pi_{1l} \pi_{2k} \pi_{2l}}{\pi_1^3 \pi_{1k} \pi_{1l} \pi_{2k} \pi_{2l} \left( 1 - \pi_1 \pi_{2k} \right)} \text{Cov}_p(I_{2k'}, I_{2k} I_{2l})
$$

$$
= N^{-4} \sum_{g \in G} \sum_{k \in U_g} \sum_{l \in U_g} \frac{\Delta_{1kl} y_k y_l \pi_1 \pi_{1k} \pi_{1l} \pi_{2k} \pi_{2l} \left( 1 - \pi_1 \pi_{2k} \right)}{\pi_1^3 \pi_{1k} \pi_{1l} \pi_{2k} \pi_{2l}}
$$

In order to study the large sample properties of the estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \), assume there is a sequence of populations \( \{ U_{\nu} \} \) such that \( U_{\nu} \subset U_{\nu+1} \) for every \( \nu = 1, 2, \ldots \), each \( U_{\nu} \) composed of \( N_{\nu} \) elements of an infinite sequence of elements \( \{ u_k \} \). Hence, \( U_{\nu} = \{ u_k : k = 1, 2, \ldots N_{\nu} \} \) and \( N_{\nu} < N_{\nu+1} \) for every \( \nu \).
Let $\theta_1\nu$ and $V_{2\nu}$ be the parameters of interest $\theta_1$ and $V_2$ calculated over the $\nu^{th}$ population. For each $\mathcal{U}_\nu$, a subsample $\mathcal{R}_\nu$, of size $n_2\nu$ is taken accordingly to the previously defined design $p_\nu$. Let $I_{2k\nu}$ denote the subsample $\mathcal{R}_\nu$ membership indicator. Then,

$$
p_\nu : \quad E_{p_\nu}(I_{2k\nu}) = \pi_{1k\nu}\pi_{2k\nu},$$

$$E_{p_\nu}(I_{2k\nu}I_{2l\nu}) = \pi_{1kl\nu}\pi_{2kl\nu},$$

$$E_{p_\nu}(I_{2k\nu}I_{2l\nu}I_{2m\nu}) = \pi_{1klm\nu}\pi_{2klm\nu},$$

$$E_{p_\nu}(I_{2k\nu}I_{2l\nu}I_{2m\nu}I_{2n\nu}) = \pi_{1klmn\nu}\pi_{2klmn\nu}.$$

It should be noted that that $\nu \to \infty$ implies both $N_\nu \to \infty$ and $n_2\nu \to \infty$.

The following assumptions will be needed to prove the next results:

- **A1**: The inclusion probabilities are bounded from zero, i.e., there is a constant $M_1$ such that
  
  $$0 < M_1 < \pi_{jk\nu},$$
  
  $$0 < M_1 < \pi_{jkl\nu},$$
  
  $$0 < M_1 < \pi_{jklm\nu},$$
  
  and,
  
  $$0 < M_1 < \pi_{jklmn\nu},$$
  
  for $j = 1, 2$ and for every $\nu$;

- **A2**: $\eta_g + \sigma_g^4$ are bounded, i.e. there is a constant $M_2 < \infty$ such that
  
  $$\eta_g + \sigma_g^4 < M_2 \forall g;$$

- **A3**: $\max_{k \neq l} |\Delta_{jkl\nu}| = O(N_\nu^{-1})$, i.e. there is a constant $M_3 < \infty$ such that $\max_{k \neq l} |\Delta_{jkl\nu}| < M_3 N_\nu^{-1}$ for every $\nu$ and $j = 1, 2$;
• A4: $\sum_{k \in U_\nu} \frac{1}{N_\nu} y_k^6$ is bounded, i.e., there is a constant $M_4 < \infty$ such that $\sum_{k \in U_\nu} \frac{1}{N_\nu} y_k^6 < M_4$
for every $\nu$.

**Result 3.5** Under assumptions $A_1$-$A_3$, $\hat{\theta}_{1\nu} = \theta_{1\nu} + o_p(N_\nu^{-1})$.

**Proof:** It is desired to show that for any $\epsilon > 0$,

$$\lim_{\nu \to \infty} \Pr(N_\nu|\hat{\theta}_{1\nu} - \theta_{1\nu}| > \epsilon) = 0.$$ 

Using the Chebyshev’s inequality,

$$\Pr(N_\nu|\hat{\theta}_{1\nu} - \theta_{1\nu}| > \epsilon) \leq \frac{N_\nu^2 \text{Var}_{P_\nu \xi_A}(\hat{\theta}_{1\nu} - \theta_{1\nu})}{\epsilon^2}$$

$$= \frac{N_\nu^2}{\epsilon^2} \left[ \frac{1}{2N_\nu^2} \sum_{g \in G} \sum_{k \in U_\nu} \eta_g + \sigma_g^4 \right] + \frac{1}{N_\nu^4} \sum_{g \in G} \sum_{g' \in G} \sum_{k \in U_\nu} \sum_{l \in U_{l'}} \frac{\Delta_{2kk'} \sigma_g^2 \sigma_{g'}^2}{\pi_{1kk'} \pi_{2kk'} \pi_{1ll'} \pi_{2ll'}}$$

$$= B_{1\nu} + B_{2\nu} + B_{3\nu}.$$

It follows that

$$B_{1\nu} \leq \frac{1}{2\epsilon^2 N_\nu^2} \sum_{g \in G} \sum_{k \in U_\nu} \frac{\eta_g + \sigma_g^4}{M_1^4}$$

$$\leq \frac{1}{2\epsilon^2 N_\nu^2} \frac{N_\nu M_2}{M_1^4}$$

$$= \frac{1}{2\epsilon^2 N_\nu} \frac{M_2}{M_1^4} \to 0 \text{ as } \nu \to \infty,$$

by assumptions A1 and A2;

$$B_{2\nu} \leq \frac{1}{\epsilon^2 N_\nu^2} \sum_{g \in G} \sum_{k \in U_\nu} \frac{\Delta_{2kk'} \sigma_g^4}{M_1^6}$$

$$\leq \frac{1}{2\epsilon^2 N_\nu^2} \frac{N_\nu(1 - M_1) M_2}{M_1^6}$$

$$= \frac{1}{2\epsilon^2 N_\nu} \frac{(1 - M_1) M_2}{M_1^6} \to 0 \text{ as } \nu \to \infty,$$
by assumptions A1 and A2, and

\[ |B_{3\nu}| \leq \frac{1}{\epsilon^2 N^2_{\nu}} \sum_{g \in G} \sum_{g' \in G} \sum_{k \in \mathcal{U}_\nu \setminus \{k\}} \sum_{l \in \mathcal{U}'_{\nu}} \frac{|\Delta_{2klu}| \sigma^2_{g} \sigma^2_{g'}}{M^6_{1}} \]

\[ \leq \frac{1}{\epsilon^2 N^2_{\nu}} \frac{N^2_{\nu} M^3_{2} M^2_{1}}{2 N^6_{\nu} M^6_{1}} \rightarrow 0 \text{ as } \nu \rightarrow \infty, \]

by assumptions A2 and A3.

Therefore, the desired result that, for any \( \epsilon > 0 \),

\[ \lim_{\nu \rightarrow \infty} \Pr(\tilde{\theta}_{1\nu} - \theta_{1\nu} > \epsilon) = 0, \]

is verified.

**Result 3.6** Under assumptions A1-4, \( \tilde{V}_{2\nu} = V_{2\nu} + o_p(N^{-1}_{\nu}) \).

**Proof:** It is desired to show that for any \( \epsilon > 0 \),

\[ \lim_{\nu \rightarrow \infty} \Pr(\tilde{V}_{2\nu} - V_{2\nu} > \epsilon) = 0. \]

Using the Chebyshev’s inequality,

\[
\Pr(N_{\nu}(|\tilde{V}_{2\nu} - V_{2\nu}|) > \epsilon) \leq \frac{N^2_{\nu} \text{Var}_{p,\nu}(\tilde{V}_{2\nu})}{\epsilon^2}
\]

\[ = \frac{N^2_{\nu}}{\epsilon^2} \sum_{g \in G} \sum_{k \in \mathcal{U}_\nu \setminus \{k\}} \sum_{l \in \mathcal{U}'_{\nu}} \frac{\Delta^2_{kklu}(4y_k^2 \sigma^2_{g} + \eta_{g} - \sigma^4_{g})}{\pi^4_{1kuv}} \]

\[ + 2 \sum_{g \in G} \sum_{g' \in G} \sum_{k \in \mathcal{U}_\nu \setminus \{k\}} \sum_{l \in \mathcal{U}'_{\nu}} \frac{\Delta^2_{kklu}}{\pi^4_{1kuv}} (y_k^2 \sigma^2_{g'} + y_{g'}^2 \sigma^2_{g} + \sigma^2_{g} \sigma^2_{g'}) \]

\[ + 8 \sum_{g \in G} \sum_{k \in \mathcal{U}_\nu \setminus \{k\}} \sum_{l \in \mathcal{U}'_{\nu}} \frac{\Delta^4_{1kkuv} \Delta^4_{1kmv}}{\pi^4_{1kuv}} \frac{y_{m} \sigma^2_{g}}{\pi^2_{1kmv}} \]

\[ + 4 \sum_{g \in G} \sum_{g' \in G} \sum_{g'' \in G} \sum_{k \in \mathcal{U}_\nu \setminus \{k\}} \sum_{l \in \mathcal{U}'_{\nu}} \sum_{m \in \mathcal{U}_\nu} \frac{\Delta^4_{1kkuv} \Delta^4_{1kmv} \Delta^4_{1klu}}{\pi^4_{1klu}} \frac{y_{l} y_{m} \sigma^2_{g}}{\pi^2_{1klu} \pi^2_{1kmv} \pi^2_{1lmv}} \]

...
\[
+ \sum_{g \in G} \sum_{k \in U_{1g}} \frac{\Delta_{1kuv}^2}{2 \pi_{1kuv}^2 \pi_{2kuv}} \left[ \eta_g (1 - 2\pi_{2k}) + \sigma_g^4 (1 + 2\pi_{2k}) \right] \\
+ 2 \sum_{k \in U_v} \sum_{l \in U_v} \left( \frac{\Delta_{1kuv} y_k}{\pi_{1kuv} \pi_{1luv}} \right)^2 \pi_{1kluv} (1 - \pi_{1kluv}) \\
+ 4 \sum_{k \in U_v} \sum_{m \neq k} \sum_{l \in U_v} \left( \frac{\Delta_{1kuv} \Delta_{1lmuv} y_k y_l}{\pi_{1kuv} \pi_{1lmuv} \pi_{1lmuv}} \right) \pi_{1kluv} \pi_{1lmuv} \pi_{1lmuv} (1 - \pi_{1kluv}) \\
+ \sum_{k \in U_v} \sum_{l \in U_v} \sum_{m \neq k} \sum_{n \neq k} \sum_{l \in U_v} \left( \frac{\Delta_{1kuv} \Delta_{1lmuv} \Delta_{1mnuv} y_k y_l y_m y_n}{\pi_{1kuv} \pi_{1lmuv} \pi_{1lmuv} \pi_{1lmuv}} \pi_{1kluv} \pi_{1lmuv} \pi_{1lmuv} \pi_{1lmuv} \right) \\
= C_{1\nu} + C_{2\nu} + C_{3\nu} + C_{4\nu} + C_{5\nu} + C_{6\nu} + C_{7\nu} + C_{8\nu},
\]

It follows that,

\[
C_{1\nu} \leq \frac{2}{N_{\nu}^2 e^2} \sum_{g \in G} \sum_{k \in U_{1g}} \frac{\Delta_{1kk}^2}{M_1^5} (4y_k^2 \sigma_g^2 + \eta_g - \sigma_g^4) \\
\leq \frac{8}{N_{\nu}^2 e^2} \sum_{k \in U_v} \frac{\Delta_{1kk}^2 M_2 y_k^2}{M_1^3} \\
\leq \frac{8}{N_{\nu} e^2} \frac{(1 - M_1)^2 M_2}{M_1^5} \sum_{k \in U_v} \frac{y_k^2}{N_{\nu}} \\
\frac{8}{N_{\nu} e^2} \frac{(1 - M_1)^2 M_2}{M_1^5} M_4 \to 0 \text{ as } \nu \to \infty,
\]

under assumptions A1, A2 and A4;

\[
C_{2\nu} \leq \frac{1}{N_{\nu}^2 e^2} \sum_{g \in G} \sum_{k \in U_{1g}} \sum_{l \neq k} \frac{\Delta_{1klu}^2}{M_1^3} (y_k^2 \sigma_g^2 + y_l^2 \sigma_g^2 + \sigma_g^2 \sigma_g^2) \\
\leq \frac{2}{N_{\nu}^2 e^2} \sum_{k \in U_v} \sum_{l \neq k} \frac{\Delta_{1klu}^2}{M_1^3} M_2 (y_k^2 + y_l^2 + 1) \\
\leq \frac{2 M_2}{N_{\nu} e^2 M_1^5} \sum_{k \in U_v} \sum_{l \neq k} \frac{\Delta_{1klu}^2}{M_1^3} \left( \frac{y_k^2}{N_{\nu}} + \frac{y_l^2}{N_{\nu}} + 1 \right) \\
\leq \frac{2 M_2}{N_{\nu} e^2 M_1^5} \frac{M_2}{N_{\nu}} \left( 2 \sum_{k \in U_v} \frac{y_k^2}{N_{\nu}} + N_{\nu} \right) \\
\leq \frac{2 M_2}{e^2 M_1^5} \left( 2 M_4 + N_{\nu} \right) \to 0 \text{ as } \nu \to \infty,
\]

by assumptions A1, A2, A3 and A4;
\[
|C_{3\nu}| \leq \frac{8}{N^2_e e^2 \sum_{g \in G} \sum_{k \in U_{g,v}} \sum_{m \neq k \in U_{v}} \frac{\Delta_{kk} y_m}{M_1^4} \sigma_g^2} \\
\leq \frac{8(1 - M_1) M_2}{N\nu e^2 M_1^5} \sum_{k \in U_{v}} \sum_{m \neq k \in U_{v}} \frac{|y_m|}{N \nu} \\
\leq \frac{8(1 - M_1) M_2 M_4}{N\nu e^2 M_1^5} \to 0 \text{ as } \nu \to \infty,
\]

by assumptions A1, A2 and A4;

\[
|C_{4\nu}| \leq \frac{1}{N^2_e e^2} \sum_{g \in G} \sum_{g' \in G} \sum_{g' \in G} \sum_{k \in U_{g,v}} \sum_{m \neq k \in U_{v}} \frac{|\Delta_{kk} y_m y_m \sigma_g^2 \pi_{klmv}|}{M_1^4} \\
\leq \frac{4GM_2}{N^2_e e^2 M_1^5} \left( \sum_{k \in U_{v}} \sum_{m \neq k \in U_{v}} |\Delta_{kk} y_k y_m| \right) + \\
\frac{4GM_2}{N^2_e e^2 M_1^5} \left( \sum_{k \in U_{v}} \sum_{m \neq k \in U_{v}} |\Delta_{kk} y_k y_m| \right) \\
\leq \frac{4GM_2}{N^2_e e^2 M_1^5} \left( (1 - M_1) \sum_{k \in U_{v}} \sum_{m \neq k \in U_{v}} |\Delta_{kk} y_k y_m| \right) + \\
\frac{M_2^2}{N^2} \sum_{k \in U_{v}} \sum_{m \neq k \in U_{v}} |y_k y_m| \\
\leq \frac{4GM_2}{N^2_e e^2 M_1^5} \left( (1 - M_1) M_3 M_4 + \frac{M_2^2}{N\nu} M_4 \right) \to 0 \text{ as } \nu \to \infty,
\]

by assumptions A1-A4;

\[
C_{5\nu} \leq \frac{1}{N^2_e e^2} \sum_{g \in G} \sum_{k \in U_{g,v}} \frac{\Delta_{kk} y_g (1 - \pi_{k\nu}) + \sigma_g^4 (1 + 2\pi_{k\nu})}{M_1^4} \\
\leq \frac{1}{N^2_e e^2 M_1^5} 6M_2 \sum_{k \in U_{v}} \Delta_{kk} y_k \\
\leq \frac{1}{N^2_e e^2 M_1^5} 6M_2 N\nu[(1 - M_1)]^2 \to 0 \text{ as } \nu \to \infty,
\]
by assumptions A1 and A2;

\[
C_{6\nu} \leq \frac{2}{N_0^2 \epsilon^2 M_1^6} \sum_{k \in U_n} \sum_{i \in U_n} \Delta_{1klv}^2 y_k^2 y_i^2 \pi \pi (1 - \pi) \\
\leq \frac{2}{N_0^2 \epsilon^2 M_1^6} (1 - M_1) \sum_{k \in U_n} \sum_{i \in U_n} \Delta_{1klv}^2 y_k^2 y_i^2 \\
= \frac{2}{N_0^2 \epsilon^2 M_1^6} (1 - M_1) \left[ \sum_{k \in U_n} \Delta_{1klv}^2 y_k^4 + \sum_{k \in U_n} \sum_{i \not\in U_n} \Delta_{1klv}^2 y_k^2 y_i^2 \right] \\
\leq \frac{2}{N_0^2 \epsilon^2 M_1^6} (1 - M_1) \left[ (1 - M_1) \sum_{k \in U_n} y_k^4 + \frac{M_2}{N_0^2} \sum_{k \in U_n} \sum_{i \not\in U_n} y_k^2 y_i^2 \right] \\
\leq \frac{2}{N_0^2 \epsilon^2 M_1^6} (1 - M_1) (1 - M_1)^2 M_4 + \frac{M_2^2}{N_0^2} M_4 \rightarrow 0 \text{ as } \nu \rightarrow \infty,
\]

by assumptions A1, A3 and A4;

\[
|C_{7\nu}| \leq \frac{4}{N_0^2 \epsilon^2 M_1^6} \sum_{k \in U_n} \sum_{i \in U_n} \sum_{m \not\in k \in U_n} |\Delta_{1klv} \Delta_{1kmv} y_k^2 y_i y_m (\pi_{1kmv} - \pi_{1klv} \pi_{1kmv})| \\
\leq \frac{4}{N_0^2 \epsilon^2 M_1^6} (1 - M_1^2) \sum_{k \in U_n} \sum_{i \in U_n} \sum_{m \not\in k \in U_n} |\Delta_{1klv} \Delta_{1kmv} y_k^2 y_i y_m| \\
\leq \frac{4}{N_0^2 \epsilon^2 M_1^6} (1 - M_1^2) \left[ \sum_{k \in U_n} \sum_{m \not\in k \in U_n} |\Delta_{1klv} \Delta_{1kmv} y_k^2 y_i y_m| + \sum_{k \in U_n} \sum_{i \not\in k \in U_n} \sum_{m \not\in k \in U_n} |\Delta_{1klv} \Delta_{1kmv} y_k^2 y_i y_m| \right] \\
\leq \frac{4}{N_0^2 \epsilon^2 M_1^6} (1 - M_1^2) (1 - M - 1) \sum_{k \in U_n} \sum_{m \not\in k \in U_n} \frac{|y_k^2 y_i y_m|}{N_0^2} \\
+ \frac{M_2}{N_0^2} \sum_{k \in U_n} \sum_{i \not\in k \in U_n} \sum_{m \not\in k \in U_n} \frac{|y_k^2 y_i y_m|}{N_0^2} \\
\leq \frac{4}{N_0^2 \epsilon^2 M_1^6} (1 - M_1^2) (1 - M_1) M_3 \sum_{k \in U_n} \sum_{m \not\in k \in U_n} \frac{|y_k^2 y_i y_m|}{N_0^2} + \frac{M_2}{N_0^2} M_3 \rightarrow 0 \text{ as } \nu \rightarrow \infty,
\]

by assumptions A1, A3 and A4.

Finally,
\[ |C_{8\nu}| \leq \frac{1}{N^2 \nu^2 M^4_1} \sum_{k \in \mathcal{U}_L \setminus \mathcal{U}_R} \sum_{l \in \mathcal{U}_R \setminus \mathcal{U}_L} \sum_{m \neq k \in \mathcal{U}_L} \sum_{n \neq k \in \mathcal{U}_L} |\Delta_{1klv}\Delta_{1mnv} y_k y_l y_m y_n (\pi_{1klv} - \pi_{1klv}\pi_{1mnv})| \]
\[ \leq \frac{1}{N^2 \nu^2 M^4_1} (1 - M^2_1) \sum_{k \in \mathcal{U}_L \setminus \mathcal{U}_R} \sum_{l \in \mathcal{U}_R \setminus \mathcal{U}_L} \sum_{m \neq k \in \mathcal{U}_L} \sum_{n \neq k \in \mathcal{U}_L} |\Delta_{1klv}\Delta_{1mnv} y_k y_l y_m y_n| \]
\[ \leq \frac{1}{N^2 \nu^2 M^4_1} (1 - M^2_1)(1 - M^2_1) \sum_{k \in \mathcal{U}_L \setminus \mathcal{U}_R} \sum_{m \neq k \in \mathcal{U}_L} |\Delta_{1klv}\Delta_{1mnv} y_k y_l y_m y_n| \]
\[ \quad + \frac{M^2_2}{N^3 \nu} \sum_{k \in \mathcal{U}_L \setminus \mathcal{U}_R} \sum_{m \neq k \in \mathcal{U}_L} \sum_{n \neq k \in \mathcal{U}_L} |\Delta_{1klv}\Delta_{1mnv} y_k y_l y_m y_n| \]
\[ \leq \frac{1}{N^2 \nu^2 M^4_1} (1 - M^2_1)(1 - M^2_1) M_4 + \frac{M^2_2}{N^3 \nu} \to 0 \text{ as } \nu \to \infty, \]

by assumptions A1, A3 and A4.

The last two results have shown the consistency and the rate of convergence of the estimators \( \hat{\theta}_1 \) and \( \hat{V}_2 \), i.e.

\[ \hat{\theta}_1 = \theta_{1\nu} + o_p(N^{1-1}_\nu), \]

\[ \hat{V}_2 = V_{2\nu} + o_p(N^{1-1}_\nu). \]

Let \( X_\nu = (\hat{\theta}_{1\nu}, \hat{V}_{2\nu})' \) and \( a = (\theta_{1\nu}, V_{2\nu})' \). It follows from known large-sample theory (see [Fuller, 1996, p.225]), that for any real valued function \( f \) defined on a \( k \)-dimensional Euclidean space with continuous partial derivatives of first order at \( a \),

\[ f(X_\nu) = f(a) + \frac{\partial f(a)}{\partial \theta_1} \bigg|_{\theta_1 = \theta_{1\nu}} (\hat{\theta}_{1\nu} - \theta_{1\nu}) + \frac{\partial f(a)}{\partial V_2} \bigg|_{V_2 = \hat{V}_{2\nu}} (\hat{V}_{2\nu} - V_{2\nu}) + o_p(N^{2-2}_\nu), \]

(3.6)
given that \( \theta_1 > 0 \) and \( V_2 > 0 \).

In particular, this is true for the function of interest \( \theta_{2\nu} = f(a) = \theta_{1\nu}/(\theta_{1\nu} + V_{2\nu}) \). As a consequence, it is possible to approximate the value of \( \hat{\theta}_2 \) by a value \( \hat{\theta}_2^* \) as follows:
\begin{equation}
\hat{\theta}_2 = \hat{\theta}_2^* = \theta_2 + a_2(\hat{\theta}_1 - \theta_1) + a_1(\hat{V}_2 - V_2),
\end{equation}

where \( a_i = V_i/(V_1 + V_2)^2 \), \( i \in \{1, 2\} \). Based on this linearization, \( \hat{\theta}_2 \) is consistent for \( \theta_2 \) and the variance of its limiting distribution is given by

\begin{equation}
\text{Var}_{p_{\xi_A}}(\hat{\theta}_2^*) = a_2^2 \text{Var}_{p_{\xi_A}}(\hat{V}_1) + a_1^2 \text{Var}_{p_{\xi_A}}(\hat{V}_2) - 2a_1a_2 \text{Cov}_{p_{\xi_A}}(\hat{V}_1, \hat{V}_2),
\end{equation}

where the explicit forms of \( \text{Var}_{p_{\xi_A}}(\hat{V}_1) \), \( \text{Var}_{p_{\xi_A}}(\hat{V}_2) \) and \( \text{Cov}_{p_{\xi_A}}(\hat{V}_1, \hat{V}_2) \) are given in previous results.

### 3.4 The Subsample Design

Assume the quantities \( \eta_g \) and \( \sigma_g^4 \) are estimated using estimators \( \hat{\eta}_g \) and \( \hat{\sigma}_g^4 \), based on previous quality evaluation studies.

Consider \( p_2 \) as a stratified sample design where, due to the presence of measurement error, the set of groups \( G \) is used as strata. In each stratum, a simple random sample is taken. Under this context, \( \mathcal{R} = \bigcup_{g \in G} \mathcal{R}_g \). The subsample design \( p_2 \) is described as

\[
p_2 : \quad \pi_{2k} = \frac{n_{2g}}{n_{1g}} \quad \forall k \in \mathcal{R}_g \quad \pi_{2kl} = \frac{n_{2g}(n_{2g} - 1)}{n_{1g}(n_{1g} - 1)} \quad \forall k, l \in \mathcal{R}_g,
\]

where \( n_{1g} \) and \( n_{2g} \) denote the sample and the subsample size of stratum \( g \in G \). In order to estimate the desired parameters \( \theta_1 \) and \( \theta_2 \), the problem of sample size allocation should be addressed. A result concerning optimum allocation under a linear cost function, for the case where \( \theta_1 \) is estimated, is given next.

**Result 3.7** Consider estimating \( \theta_1 \). Assume the cost of implementing the subsample \( \mathcal{R} \) is a linear function of the form \( C = c_0 + \sum_{g \in G} n_{2g}c_g \), where \( C \) is the total cost, \( c_0 \) is a given overhead cost and \( c_g \) is the cost of collecting data at each psu belonging to stratum \( g \). Then, \( \text{Var}_{p_{\xi_A}}(\hat{\theta}_1) \) is minimized when choosing \( n_{2g} \) proportional to \( (\frac{A_g}{c_g})^{1/2} \), where
\[ A_g = (2N^4)^{-1} \left[ (\eta_g + \sigma_g^4)n_{1g} \left( \sum_{k \in \mathcal{U}_g} \pi_{1k}^{-3}\right) - 2\sigma_g^4 \frac{n_{1g}}{n_{1g} - 1} \left( \sum_{k,l \in \mathcal{U}_g} \frac{\pi_{1kl}}{\pi_{1l}^2} \right) \right]. \] (3.9)

Proof:

It is sufficient to show that, under the proposed stratified sample design \( p_2 \),

\[
\text{Var}_{p_2}(\hat{\theta}_1) = (2N^4)^{-1} \sum_{g \in \mathcal{G}} \left( \frac{\eta_g + \sigma_g^4}{n_{2g}} \left( \sum_{k \in \mathcal{U}_g} \pi_{1k}^{-3}\right) \right)
+ N^{-4} \sum_{g \in \mathcal{G}} \sigma_g^4 \left[ \sum_{k,l \in \mathcal{U}_g} \frac{1}{\pi_{1k}^2 \pi_{1l}^2} \left( \frac{\pi_{1kl}}{n_{1g}(n_{1g} - 1)} - \frac{\pi_{1k} \pi_{1l}}{n_{1g}^2} \right) \right]
- N^{-4} \sum_{g \in \mathcal{G}} \sigma_g^4 \left[ \sum_{k,l \in \mathcal{U}_g} \frac{1}{\pi_{1k}^2 \pi_{1l}^2} \left( \frac{\pi_{1kl}}{n_{1g}(n_{1g} - 1)} \right) \right]
= \sum_{g \in \mathcal{G}} \frac{A_g}{n_{2g}} + B,
\]

where \( A_g \) is given above and

\[ B = N^{-4} \sum_{g \in \mathcal{G}} \sigma_g^4 \left[ \sum_{k,l \in \mathcal{U}_g} \frac{1}{\pi_{1k}^2 \pi_{1l}^2} \left( \frac{\pi_{1kl}}{n_{1g}(n_{1g} - 1)} - \frac{\pi_{1k} \pi_{1l}}{n_{1g}^2} \right) \right]. \]

The desired result follows from known stratified sample design theory (see [Särndal et al., 1992, p.105].

In addition to the optimum allocation result, three different options of sample allocation are also investigated. Define the observable standard finite population variance of the variable of interest \( y \) on stratum \( g \) as

\[ S_{Y_g}^2 = (N - 1)^{-1} \sum_{k \in \mathcal{U}_g} (Y_{1k} - \bar{Y}_1)^2. \] (3.10)

Then, the following sample size allocation schemes are considered:

**Allocation i:** Proportional to measurement error variance,

\[ n_{2g} = \frac{n_{1g} \sigma_g^2}{\sum_{g \in \mathcal{G}} n_{1g} \sigma_g^2}. \] (3.11)
**Allocation ii:** Standard Neyman allocation for estimating the mean,

\[ n_{2g} = n_2 \frac{n_{1g} Sy_g}{\sum_{g \in G} n_{1g} Sy_g} \]  

(3.12)

**Allocation iii:** Proportional to the stratum size,

\[ n_{2h} = n_2 \frac{n_{1g}}{n_1} \]  

(3.13)

Table 3.1 summarizes the allocation types.

<table>
<thead>
<tr>
<th>Allocation type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Optimum allocation (Result 3.6)</td>
</tr>
<tr>
<td>2</td>
<td>Proportional to measurement error variance (allocation i)</td>
</tr>
<tr>
<td>3</td>
<td>Proportional to ( Sy_g ) (allocation ii)</td>
</tr>
<tr>
<td>4</td>
<td>Proportional to the stratum size (allocation iii)</td>
</tr>
</tbody>
</table>

### 3.5 A Sample Allocation Simulation Study

In order to investigate the efficiency of the considered sample allocation rules, a simulation study was conducted based on the 2001 NRI sample information. The available data set included the following variables:

- Segment identification;
- First-order inclusion probabilities \((\pi_{1k})\);
- Rotation or core sample indicator;
• Cropland indicator for a point;

• Point identification;

• Soil Erosion (in tons/acre) based on universal soil loss equation (USLE).

An area segment was classified as cropland whenever at least one of its points was classified as cropland. Based on these variables, a further variable group was created as follows: Group 1 indicates segments classified as cropland that belong to the core panel; Group 2 indicates segments with no cropland that belong to the core panel; Group 3 indicates segments classified as cropland that belong to the rotation panel; and Group 4 indicates segments with no cropland that belong to the rotation panel. The group sizes are summarized in Table 3.2. The final data set contained observations associated with 39,208 area segments.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>Size (n_{1g})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>core panel and cropland</td>
<td>13,844</td>
</tr>
<tr>
<td>2</td>
<td>core panel and no cropland</td>
<td>8,750</td>
</tr>
<tr>
<td>3</td>
<td>rotation panel and cropland</td>
<td>10,127</td>
</tr>
<tr>
<td>4</td>
<td>rotation panel and no cropland</td>
<td>6,487</td>
</tr>
</tbody>
</table>

The simulation study compared the four allocation types with respect to the variance of \( \hat{\theta}_1 \). In this simulation, \( \hat{\theta}_1 \) is an estimator of the contribution to the variance of the mean estimator of the soil erosion variable due to measurement error. The comparison was carried out for each one of five different sample fractions (sample sizes): 15\% (5882), 12.5\% (4902), 10\% (3922), 7.5\% (2942), and 5\% (1961).

Estimates of the proportion of the variance of soil erosion loss due to measurement error for groups 1 and 2, denoted as \( v_1 \) and \( v_2 \), were provided by the fitting of a first-order autoregressive model, as suggested in [Breidt and Fuller, 1999]. The estimates were obtained using information for the NRI samples of 1982, 1987, 1992, 1997, 2000, 2001 and 2002 (Legg, J., unpublished estimates). Based on these values, the \( v_g \) values for groups 3 and 4 were obtained as follows:
$v_3 = 1.3v_1;

v_4 = 1.3v_2.$

These values reflect the belief that rotation panel units have more relative measurement error contribution than core panel units.

Let $S_{Y_g}^2$ denote the variance of the soil erosion loss for group $g$. Using the $v_g$ values, the measurement error variances $\sigma_g^2$ were calculated for each group as follows:

$$\sigma_g^2 = v_g S_{Y_g}^2. \tag{3.14}$$

The estimated values $v_g$, and the $S_{Y_g}^2$ and $\sigma_g^2$ values are shown in Table 3.3 for each group.

Table 3.3 Values of Proportion of Measurement Error, Variance of Soil Erosion Loss and Measurement Error Variance by Group Used in the Simulation

<table>
<thead>
<tr>
<th>Group</th>
<th>$v_g$</th>
<th>$S_{Y_g}^2$</th>
<th>$\sigma_g^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.19</td>
<td>15.25</td>
<td>2.90</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>5.28</td>
<td>1.11</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>15.66</td>
<td>3.92</td>
</tr>
<tr>
<td>4</td>
<td>0.27</td>
<td>6.00</td>
<td>1.62</td>
</tr>
</tbody>
</table>

The value of $\eta_g$ was set to be the same as $\sigma_g^4$, under the assumption of normality of the measurement errors. The cost of sampling within each stratum (group) was assumed constant.

The value of the variance of $\hat{\sigma}$ was simulated considering $\pi_1^{kl} = \pi_1^{kl} \pi_{11}$ and $\pi_2^{kl} = \pi_2^{kl} \pi_{2l}$, which is equivalent to consider sampling units with replacement. Under this set of assumptions, the $A_g$ expression for optimum allocation (Result 3.7) assumes the form

$$A_g = \frac{n_g \sigma_g^4}{N^4} \sum_{k \in U_g} \pi_k^{-3}.$$

Based on this last expression, it is expected that the performances of allocation types 1 and 2 will be similar.

Figure 3.1 shows a graphical comparison of the allocation types using the data shown at Table 3.4.
Figure 3.1 Graphical Analysis of Allocation Types
Table 3.4 Simulation Values of the Variance of $\hat{\theta}_1$ by Allocation Type and Sample Size

<table>
<thead>
<tr>
<th>Sample Size (Sample Fraction)</th>
<th>Allocation Type</th>
<th>Simulate $\hat{\theta}_1$ Variance x 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>5882 (0.150)</td>
<td>1</td>
<td>0.04906</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.04989</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.05141</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.05567</td>
</tr>
<tr>
<td>4902 (0.125)</td>
<td>1</td>
<td>0.05888</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.05987</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.06171</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.06683</td>
</tr>
<tr>
<td>3922 (0.100)</td>
<td>1</td>
<td>0.07359</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.07482</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.07714</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.08352</td>
</tr>
<tr>
<td>2942 (0.075)</td>
<td>1</td>
<td>0.09809</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.09972</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.10281</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.11132</td>
</tr>
<tr>
<td>1961 (0.050)</td>
<td>1</td>
<td>0.14706</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.14949</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.15417</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.16688</td>
</tr>
</tbody>
</table>

The graphical analysis shows that allocation type 1 (optimum allocation) minimizes the simulated variance of $\hat{\theta}_1$ for every considered sample fraction. Allocation type 2 (proportional to measurement error variance) gives values close to the approximated variances produced by allocation type 1, as expected. On the other hand, the performance of allocation type 3 is also close to the performances of allocation types 1 and 2. The performance of allocation type 4 is slightly poorer than allocation type 3 performance. Based on this simulation, given the small difference in performances, allocation type 2 should be chosen over allocation type 1 for its simplicity. In situations where the assumption $\pi_{1kl} = \pi_{1k}\pi_{1l}$ and $\pi_{2kl} = \pi_{2k}\pi_{2l}$ are not reasonable, allocation type 1 may be preferable. However, it should be noted that allocations type 1 and 2 require prior information about the $\sigma^2_g$ values, which in practice may not be
available. Allocations type 3 and 4 do not require any prior information about $\sigma_g^2$, and have performances close to allocations type 1 and 2.

3.5.1 A Sensitivity Analysis

The simulated values of the variance of $\hat{\theta}_1$ used to compare allocation types are local in the sense that they were calculated based on the estimated $\nu_g$ values shown in Table 3.3. In order to study the sensitivity of the allocation performances to different values of $\nu_g$, the same graphical analysis was produced considering lower and higher values of $\nu_g$ than the ones shown in Table 3.3. The assumptions for lower values considered the values shown in Table 3.5. Figure 3.2 shows a graphical comparison of the allocation types for these lower values, using the results shown in Table 3.6.

<table>
<thead>
<tr>
<th>Group</th>
<th>$\nu_g$</th>
<th>$\sigma_g^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>2.59</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.23</td>
<td>3.60</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>1.50</td>
</tr>
</tbody>
</table>

The assumptions for higher values considered the values shown at Table 3.7. Figure 3.3 shows a graphical comparison of the allocation types for these higher values, using the results shown at Table 3.8.

In both situations, under lower and higher values of $\nu_g$, the simulate variance was uniformly minimum for allocation type 1. Allocation type 2 had a similar performance, with simulate variance values slightly higher than allocation type 1. Showing performance close to allocation types 1 and 2, allocation type 3 was uniformly better than allocation type 4. A sensitivity analysis to evaluate the impact of changing the multiplicative term 1.3 (when setting the $\nu_g$ values for groups 3 and 4) did show no change in the allocation performances.
Figure 3.2 Sensitivity Analysis of Allocation Types for Lower Values
Figure 3.3 Sensitivity Analysis of Allocation Types for Higher Values
Table 3.6 Simulation values of the Variance of \( \hat{\theta}_1 \) by Allocation Type and Sample Size for Lower Values

<table>
<thead>
<tr>
<th>Sample Size (Sample Fraction)</th>
<th>Allocation Type</th>
<th>Simulated ( \hat{\theta}_1 ) Variance x 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>5882 (0.150)</td>
<td>1</td>
<td>0.04064</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.04133</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.04280</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.04629</td>
</tr>
<tr>
<td>4902 (0.125)</td>
<td>1</td>
<td>0.04876</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.04961</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.05137</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.05557</td>
</tr>
<tr>
<td>3922 (0.100)</td>
<td>1</td>
<td>0.06095</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.06200</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.06422</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.06945</td>
</tr>
<tr>
<td>2942 (0.075)</td>
<td>1</td>
<td>0.08125</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.08263</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.08559</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.09257</td>
</tr>
<tr>
<td>1961 (0.050)</td>
<td>1</td>
<td>0.12181</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.12377</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.12835</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.13877</td>
</tr>
</tbody>
</table>

3.6 Simulation Conclusions

Theoretical development of a stratified sample design for measurement error evaluation of large-scale longitudinal surveys lead to the proposal of an optimal sample size allocation rule. A simulation study, using NRI data, investigated the performance of such allocation (type 1) compared to other three allocation types: proportional to measurement error variance (type 2), standard Neyman allocation for estimating the mean (type 3), and proportional to stratum size allocation (type 4). The results have shown that the optimal allocation rule has uniformly better performance than the other allocation types for different sample sizes corresponding to different sample fractions. Allocations type 2 and 3 had performances very close to the shown by allocation type 1.
Table 3.7  Higher Values of $v_g$ and $\sigma_g^2$ for Soil Erosion Loss

<table>
<thead>
<tr>
<th>Group</th>
<th>$v_g$</th>
<th>$\sigma_g^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.21</td>
<td>3.20</td>
</tr>
<tr>
<td>2</td>
<td>0.23</td>
<td>1.21</td>
</tr>
<tr>
<td>3</td>
<td>0.27</td>
<td>4.23</td>
</tr>
<tr>
<td>4</td>
<td>0.29</td>
<td>1.74</td>
</tr>
</tbody>
</table>

The results have also shown the robustness of the four types of allocation to lower and larger values of the percentage of the variable of interest variance due to measurement error ($v_g$).

It should be noted that only allocation types 3 and 4 can be implemented with no prior information regarding $\sigma_g^2$ values. Therefore, if no reasonable information about $\sigma_g^2$ is available, allocations type 3 and 4 are alternatives that can be implemented in practice with reasonable performances when compared to allocation types 1 and 2.
Table 3.8 Simulation Values of the Variance of $\hat{\theta}_1$ by Allocation Type and Sample Size for Higher Values

<table>
<thead>
<tr>
<th>Sample Size (Sample Fraction)</th>
<th>Allocation Type</th>
<th>Simulated $\hat{\theta}_1$ Variance x 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>5882 (0.150)</td>
<td>1</td>
<td>0.05829</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.05926</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.06083</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.06593</td>
</tr>
<tr>
<td>4902 (0.125)</td>
<td>1</td>
<td>0.06994</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.07111</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.07301</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.07914</td>
</tr>
<tr>
<td>3922 (0.100)</td>
<td>1</td>
<td>0.08741</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.08887</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.09127</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.09891</td>
</tr>
<tr>
<td>2942 (0.075)</td>
<td>1</td>
<td>0.11656</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.11846</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.12164</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.13184</td>
</tr>
<tr>
<td>1961 (0.050)</td>
<td>1</td>
<td>0.17468</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.17762</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.18240</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.19764</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


ACKNOWLEDGEMENTS

I would like to express my thanks to Dr. Sarah M. Nusser and Dr. Jean D. Opsomer, my major professors, for their support throughout this research.

I would like to thank my wife, Carla and my son, Rafael without whose support, I would not have been able to complete this work.

I would also like to thank my parents Euclides Ferraz Filho and Maria Margarida de Assis Ferraz, and my siblings, Andreza, Eugenio and Daniela, for their support throughout my life.

I want to express my thanks to the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior - CAPES, Brasil for their financial support.