Linköping Studies in Statistics No. 2 • Doctoral Thesis

# Survey Models for a Vehicle Speed Survey

Annica Isaksson


FACULTY OF ARTS AND SCIENCES
LINKÖPINGS UNIVERSITET

Statistics
Department of Mathematics
Linköpings universitet
SE-581 83 LINKÖPING
Sweden

ISBN: 91-7373-572-8
ISSN: 1651-1700

Cover art by Björn Böke.
Printed in Sweden by UniTryck, Linköping, 2002.

**Abstract**

The impact of some errors associated with a national road traffic survey is examined in this thesis. The current survey aims to evaluate efforts to reduce the speed of traffic on Swedish roads; it covers both state and urban roads, though the thesis considers only urban roads. In the survey, observational sites are selected by a three-stage sampling procedure. A measurement device installed on the road is used to collect data, from which the average speed of traffic on the roads is estimated.

This thesis focuses on errors in the frames used in the final sampling stage, and on errors due to missing data. The impact of these errors on the total error of the survey estimators is investigated. Also explored are possibilities for reducing the total error by weighting adjustments for missing data and by reallocating the sample over the three sampling stages. The problems are approached partly theoretically, by use of various error models; partly empirically, by collecting data on the errors. Throughout, the sampling design of the survey is taken properly into account. Our conclusion is that the frame error under consideration probably does not bias the estimator of average speed, and it only implies a minor increase of its variance. It remains unclear whether the estimator needs to be adjusted for missing data: however, a theoretical framework for further investigations is provided. For unchanged total sample size, the precision of the estimator is likely to improve if the sample sizes in the third stage are increased, and the sampling sizes in the first stage are decreased correspondingly.

**Key words:** Frame error, missing data, error model, optimum allocation, total error.

**Acknowledgements**

The road to a finished thesis is rarely straight, and mine has had its share of detours and deadlocks. Whenever I lost my direction, however, my supervisor, Professor Stig Danielsson, was always approachable and helped me find it again. Someone wrote that the best teacher is the one who suggests rather than dogmatizes, and who inspires his student with the wish to educate herself. Professor Danielsson has been a great teacher to me, and my first and deepest expression of gratitude is for him.

I have enjoyed the privilege of additional guidance by my assistant supervisor, Professor Gösta Forsman, chief statistician at the Swedish National Road Administration (SNRA). Professor Forsman's profound knowledge of survey methodology in general, and road traffic surveys in particular, has made his advice irreplaceable in this work.

I am grateful to the SNRA for financial support of this work. My former superior, the now retired head of the Traffic Division of the SNRA Consulting Services Directorate, Göthe Edlund, originally arranged the support, and I am much obliged to him for his share in the matter. The personal help provided by numerous other SNRA employees during various stages of the work has also been priceless. I wish to express my intense gratitude to all of them, and most notably to Ingela Stenbäck, who has been my undeviating contact at the Traffic Division over the years, and to Magnus Johansson at the SNRA's South Eastern Region, who collected a large part of the experimental data with great care.

The entire staff at the Division of Statistics at Linköping University has contributed to this work by providing a very supportive environment, in which the thesis has been able to develop in serenity. Many thanks to all of them, but in particular to former and present fellow doctoral students for the shared laughter, the welcome coffee-breaks at odd hours, and all the friendly gestures along the way.

Linköping, December 2002
Annica Isaksson

# Contents

# Chapter 1

# Introduction

By decision of the Swedish government, since 1997, the official lodestar for Swedish traffic safety work has been 'Vision Zero': an image of a desirable future society in which no one is killed or seriously injured in road traffic. The overall responsibility for the traffic safety work is held by the Swedish National Road Administration (SNRA). In 1994, a programme of how SNRA and other actors should turn vision into reality was launched: the National Road Traffic Safety Programme for 1995-2000 [31]. In 1999, the program was succeeded by the government's 11-Point Programme for Improving Road Traffic Safety [30]. A common denominator of both programmes has been the emphasis put on road-user responsibility, including compliance with speed limits. Current measures to reduce speeds include physical changes of the traffic environment (for instance, converting intersections into roundabouts) and campaigns directed towards the public. In order to assess the results of these measures, the SNRA has conducted since 1996 an annual survey of vehicle speeds. It is generally wise to evaluate the design and performance of a recurrent survey such as this from time to time, thereby making it possible to maintain and improve the standard of the survey over time. This thesis constitutes the first (but hopefully not the last) evaluation of the speed survey.

In brief, the annual survey is conducted as follows. The roads are thought of as partitioned into one-meter road sites, which are the population elements. In selected sites, during a random 24-hour period, data are collected by use of

a measurement device installed on the road. The device records the number of passing vehicles, and the total time they take to pass the site. From these data, the average speed on the roads is estimated.

Although the survey is conducted on both state and urban roads, in this thesis we restrict our attention to the part of the survey that concerns the urban roads. For these roads, a three-stage sampling design is used for selecting sites. The primary sampling units are population centers, and the secondary sampling units are small areas. For each selected small area, a frame of the road network is used. The frame units are road links, and the frame contains information on the length of each link. From the frame, road sites are randomly selected for observation.

## 1.1 Problems treated in the thesis

The speed survey estimates receive much attention and serve as a basis for decisions on future traffic safety measures. Of course, the estimates need to be reliable. Different users of the results, however, have different needs, and their assessments of an adequate level of reliability are likely to differ accordingly. This raises an urgent need to account for the uncertainty in the results. The sampling error, that is, the uncertainty due to the fact that only a subset of all roads (and of the whole study period) is observed, is presently quantified by conventional 95 percent confidence intervals. Such intervals do not, however, give the full picture if there are additional errors present. In this thesis, the additional uncertainties due to imperfect sampling frames and missing data are investigated. Besides looking at the isolated impact of each type of error, we adopt a comprehensive view towards them, by formulating survey models for the estimators in use. More precisely, we derive the estimators' expectations and variances with respect jointly to the sampling design and to models for errors due to frame imperfections and missing data.

The speed survey is a resource-demanding undertaking, both financially and in terms of personnel. The survey management estimates the cost of the last survey round (in summer 2002) at 5.3 million Swedish kronor (about 0.6 million euros). A large field staff is needed to instal the measurement

equipment in selected sites. Above all, it is in the SNRA's (and the Swedish taxpayers) best interest to render the conduct of the survey more effective, if possible. To support this aim, this thesis also investigates the possibility of reducing the sampling error by reallocating the total sample over sampling stages.

## 1.2 Approaches to the problems

In this section we take a brief look at each identified problem and how we chose to approach them. For a summary of the results of our efforts, the reader is referred to Chapter 7.

### 1.2.1 The frame error problem

Since road sites are selected by a three-stage sampling procedure, several sampling frames are in use in the speed survey. Each of them may suffer from various types of errors. Attention has been restricted in this thesis, however, to a particular error associated with the frames of road links used in the final sampling stage. When these frames were constructed, the link lengths were determined manually from maps. Hence, the lengths may be subject to measurement errors. By use of a simple error model, we examined the impact of this frame error on the bias and variance of employed estimators. In our model, the total frame road length for a small area is viewed as a function of the true length and a random error. Data from an empirical study were used to evaluate the model and estimate the error parameters.

### 1.2.2 The missing data problem

At each selected site, a measurement device is used to collect data. The device consists of two pneumatic tubes stretched across the road and connected to a traffic analyzer. When a wheel of a passing vehicle crosses a tube, this action gives rise to a pulse in the equipment. From these registered pulses, 'vehicles' are created. Typically, a number of the passing vehicles will remain unobserved. The failure to observe some vehicles is indicated on

one hand by imputations automatically created by the device, on the other by the measurement efficiency (ME) – the proportion of registered pulses successfully combined into vehicles – being small. To adjust for missing data, we suggest dividing the traffic passing a site into weighting classes.

Consider the problem of adjusting the observed flow upwards. Within class, one proposal is to add the number of imputed vehicles; another is to weight the observed flow by an estimated probability of registration (assumed to be constant within class). The ME is put forward as estimator of the registration probability. Models for the errors in the number of imputed vehicles, and in the estimated registration probabilities, are used for theoretical evaluations. The models are evaluated, and the suggested estimation strategies compared, by use of some empirical data.

### 1.2.3 The allocation problem

The allocation of the total sample over the three sampling stages was decided at an early stage of the survey and without the benefit of much evidence. Thus, very likely, there is room for improvement. In order to evaluate the allocation, we have estimated the components, arising from each sampling stage, of the total variances of the employed estimators. We also present formulae for determination of optimum sampling sizes in each sampling stage.

In all but the first sampling stage, only one sampling unit per stratum is selected. This makes estimation of the variance contributions from each sampling stage impossible. We have circumvented this problem by using a 'fictitious' sampling design, which resembles the actual design but is conditioned by the set of distinct units selected in stage one, and some experimental data.

### 1.2.4 The development of survey models

Finally, we derive the expectations and variances of the employed estimators, taking into account both the sampling design and our models for errors due to frame errors and missing data.

## 1.3 Thesis goal

The main goal of this thesis is to guide evaluation and improvement of the quality of the speed survey. We do not strive to deliver final solutions on how to make the speed survey better, but rather to aid the SNRA in its revision of the survey. More specifically, we want to

- provide the speed survey management with theoretical tools for assessing the impact of frame errors and missing data on the survey results;
- demonstrate how experiments can be designed to support the theoretical results; and
- make some preliminary statements on the impact and importance of the errors under consideration, and on the possibility of reducing the sampling error by reallocating the total sample over sampling stages.

The time and budget frames of this thesis work have only allowed experiments to be performed on a limited scale. This limitation means that the practical results reported in the thesis must be considered as preliminary.

## 1.4 Contributions of the thesis

The nonsampling errors considered in this thesis are not unfamiliar to the speed survey management. Quite the opposite: The errors have both been recognized and caused uneasiness for quite some time. It is, however, one thing to note the failings of a sampling frame, or the incapacity to obtain complete observational data from selected road sites – but quite another matter to estimate the importance of such findings. This thesis demonstrates ways of dealing statistically with these problems by use of random error models. Despite the simplicity of our models, they do enable us to investigate, both theoretically and practically, the impact of the errors on the bias and variance of the estimators in use. The model assumptions are clearly stated, and we show how to design experiments for evaluating them. Beside this, we also explore the possibility of reducing the sampling error by changing the current sample allocation over sampling stages. In summary, the primary

contribution of the thesis is the support it provides for rational decision making regarding the distribution of available survey resources.

Our ways of attacking the errors due to imperfect sampling frames and missing data draw inspiration mainly from approaches to measurement errors in surveys discussed in [2] and [29, Chapter 16]. Throughout the thesis, we make a point of taking the sampling design of the survey properly into account. We thus illustrate the use (and usefulness) of the statistical theory both on nonsampling errors and on sampling. In our experience, these lines of theory have still not been accepted by traffic surveyors, and we hope that the thesis can contribute to changing this situation.

The statistical journals and conferences devoted to survey methodology are, in our experience, dominated by approaches to methodological problems connected with surveys of human populations. As this thesis illustrates, to fit other fields of applications, available approaches may need modification or expansion. In particular, we find it useful to model both frame errors and missing data problems as 'measurement errors.' Also, although the sampling design of the speed survey is of standard type, due to small sampling sizes, available formulae for estimating variance components do not apply. We thus need to be imaginative, and to try using a 'fictitious' sampling design to estimate the components. This is an approach we have not seen in the literature.

## 1.5 Outline of the thesis

The thesis is outlined as follows. In **Chapter 2**, the main methodological features of the speed survey are described. The impact of erroneous frame road lengths on the survey estimators is analyzed in **Chapter 3**. In **Chapter 4**, we investigate the impact of missing data on the survey estimators. Two strategies for adjusting for missing data in the estimation stage of the survey are also introduced and evaluated. In **Chapter 5**, we turn our attention to the sampling error. We evaluate the current allocation of the total sample over sampling stages, and present formulae for determination of optimum sampling sizes. Survey models for the estimators in use are formulated in **Chapter 6**. A brief summary of our findings, finally, is given in **Chapter 7**.

## 1.5. Outline of the thesis

The thesis is based on earlier work by the author: Chapter 2 is based on [18, Chapter 2] and Chapter 3 on [18, Chapters 3 and 4], Chapter 4 is based on [21], Chapter 5 on [19] and Chapter 6, finally, on [20].

Versions of Chapter 3 were presented at the 3rd Finnish Sampling Symposium, University of Jyväskylä, Finland, 15-18 May 2000, and at the 4th Conference on Methodological Issues in Official Statistics, Statistics Sweden, 12-13 October 2000. Versions of Chapter 4 were presented at the Joint Statistical Meeting, New York City, USA, 11-15 August 2002, and at the International Conference on Improving Surveys, University of Copenhagen, Denmark, 25-28 August 2002. Our efforts to develop survey models for the speed survey have also been shared at the Young Researchers Invited Poster Session of the Conference in Celebration of Wayne A. Fuller's 70th Birthday, Ames, Iowa, USA, 21-22 June 2001.

A revised version of Chapter 4 has been submitted for publication in the Journal of Official Statistics.

# Chapter 2

# The speed survey

## 2.1 Background and survey objectives

The initial model for SNRA's traffic safety work was the National Road Traffic Safety Programme for 1995-2000 [31]. In line with this programme, the work was organized in focus areas, called road traffic safety reforms, such as 'reduction in speeding offences,' 'use of cycle helmets,' and 'use of safety equipment in cars.' Operational goals were stated for each reform, and it was assumed that if a reform goal was reached, this should contribute to a reduction in traffic deaths and injuries. In order to assess whether development was heading toward the reform goals, starting in 1996, the SNRA launched several observational sample surveys; for instance, of helmet usage among cyclists, of usage of luminous tapes or tags among pedestrians, of seat-belt usage among motorists, and of motorists who drive against red lights. The largest initiative, however, was the survey on vehicle speeds.

Currently, the traffic safety work is modeled on the 11-Point Programme for Improving Road Traffic Safety [30]. It has proved difficult to change people's behaviors and attitudes, and therefore more attention is now paid to safety improvements of the physical road environment. The sole surviving observational traffic safety survey conducted by the SNRA is the survey of vehicle speeds.

## 2.2 Population

The target population is the entire Swedish road network except rural private roads. It is divided into two subpopulations of special interest – state roads and 'urban' roads (local authority roads and private roads in built-up areas) – which also serve as strata when the sampling is conducted. In this thesis, we restrict our attention to the part of the survey that concerns the urban roads, and refer throughout to the urban road network as the target population. This road network is considered as partitioned into one-meter road sites, which are the population elements.

From the target population, the following road sections are excluded:

- From major roads: 100 meters before and after each intersection with traffic lights.
- From non-major roads: 100 meters before and after each intersection.

The main reason for excluding road sections close to intersections is to avoid observational difficulties. In the speed survey, observations are carried out by measurement equipment installed on the road (see Section 2.5). Certain traffic situations, such as vehicles lining up, accelerating, or decelerating, have the potential to cause measurement problems. Such situations frequently occur close to intersections.

Survey results are demanded not only for the whole target population, but also for specific subpopulations or 'domains.' One important goal of the survey is to provide results for each SNRA region. The SNRA organization includes seven regional road management directorates (Figure 2.1), which are responsible for the SNRA's regional management, including traffic safety work within their geographic areas.

The definition of the target population is not complete without some restriction in time. The survey is always conducted during the summer months, but the exact period of study changes somewhat from year to year. In the last survey round, in 2002, the study period was May 27 to September 30. The period of study is thought of as a population in time, with 24-hour periods as population elements.

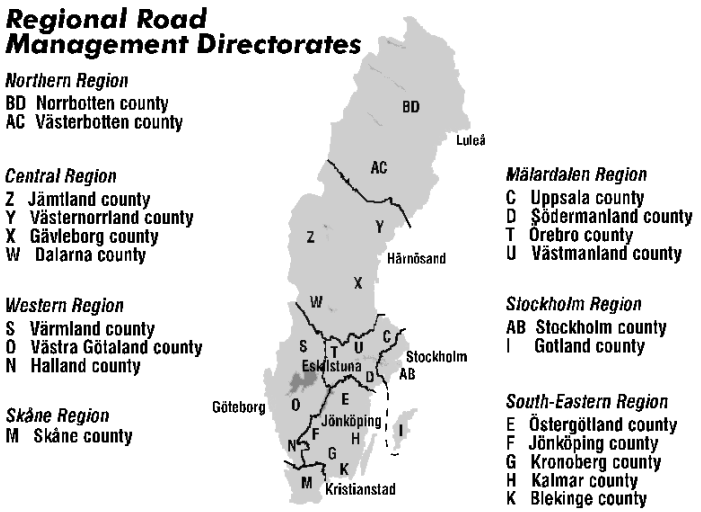


Figure 2.1: The SNRA Regional Road Management Directorates. (Source: SNRA)

## 2.3 Variables and parameters

The primary study variable is the traffic flow, $y$. In general, for a given point on a road and a specified period of time, the traffic flow is defined as the number of passing vehicles. Since in this survey a road is viewed as made up of one-meter sections, a 'point' is interpreted as a one-meter section (a road site). From the total flow, other study variables of interest can be derived; some examples include:

- Flow above a certain speed limit
- Flow with less than a certain headway
- Flow by a certain type of vehicle (e.g., by cars with trailers)

The second main study variable is the travel time, $z$. For a given traffic flow, the travel time is the total time all vehicles take to pass the road point.

Let $U$ denote the target population 'in space' – the set of road sites that make up the urban road network – and $U_\Upsilon$ the target population 'in time'

– the set of 24-hour periods that make up the time period of study. The population total of the study variable $y$, the 'total vehicle mileage' for the road network and time under study, is $\sum_{U_\Upsilon} \sum_U y_k^\upsilon$, where $y_k^\upsilon$ equals the traffic flow in site $k \in U$ during 24-hour period $\upsilon \in U_\Upsilon$. Correspondingly, the population total of $z$, the 'total travel time,' is given by $\sum_{U_\Upsilon} \sum_U z_k^\upsilon$. Since the total vehicle mileage is a measure of distance, and the total travel time a measure of time, their ratio is a measure of speed.

Throughout this thesis, we ignore possible time variability in $y$ and $z$. That is, we consider only the special case when $y_k^\upsilon = y_k$ and $z_k^\upsilon = z_k$ for all $\upsilon \in U_\Upsilon$, $k \in U$. Therefore, we will hereafter drop the time index and refer simply to the parameters $t_y = \sum_U y_k$, $t_z = \sum_U z_k$ and $R = t_y/t_z$. It follows that the sampling in time will not be treated (nor its consequences to the estimation). Some aspects of sampling in time are investigated in [17].

Instead of treating population totals for various study variables as separate cases, we usually find it sufficient to refer to $t_a$: the population total for an arbitrary study variable $a$.

## 2.4 Sampling design

### 2.4.1 Use of a master frame

The speed survey was launched together with a number of other observational traffic safety surveys (see Section 2.1). To keep the money and time spent on frame construction down, the surveys shared the same sampling in the first and second stages. In each survey, the final sample then consisted of locations selected from a master frame of roads. Depending on the nature of the survey, the locations could be, for instance, intersections with traffic lights (suitable for observing motorists who drive through red lights) or, as in the speed survey, one-meter sections of the road. Today, the only traffic safety survey still utilizing the master frame is the vehicle speed survey. The frame was used for a survey on the conditions of various types of road equipment (road signs, road fences, and lamp posts) in 2001, and other fields of application may be brought to the fore in the future.

The method of using a master frame is discussed in the literature by,

among others, Kish [23, pp. 478-480], and can briefly be described as follows. Initially, a 'master sample' of sampling units is selected. For each sampled unit, a frame is prepared. The sample for a particular survey is then selected from these frames, which serve for a longer time period.

The SNRA's master sample was selected during 1995-96 by a two-stage sampling design. The primary sampling units (PSUs) are population centers, and the secondary sampling units (SSUs) are small areas.

### 2.4.2 Frames

When the master sample was selected, the frame used in the first stage was a list, supplied by Statistics Sweden (SCB), of the Swedish population centers in 1990. The list contained auxiliary information on the number of inhabitants in each population center, which served as a size measure for probability-proportional-to-size sampling with replacement (pps). The frames used in the second stage were lists of the small areas within selected population centers. In all essentials, these small areas correspond to SCB's small area market statistics (SAMS) regions. Various population statistics collected by SCB are tied to developed properties. In co-operation with the local governments, SCB has grouped similar adjacent properties. By a special technique called 'register generated borders,' geographic borders between the groups have been fixed. The resulting nationwide area division is called SAMS. There are about 9,200 SAMS regions; their main use is for statistical presentations.

For each selected small area, a list frame of road links was prepared at the SNRA from city maps. Using the intersections as breakpoints, the road-map network was partitioned into links, and the link lengths were determined manually by the use of map measurers.

### 2.4.3 Sample selection

In the first stage of sample selection, the population centers are stratified according to *SNRA region* (see Figure 2.1) and three *size classes*:

- Large major population center of a municipality

- Other major population center of a municipality
- Other population center

In the second stage, the small areas within a selected population center are stratified according to four *development types*:

- City
- Industrial
- Residential
- Other type

In the final stage, the road sites within a selected small area are stratified according to three *road types*:

- Major roads with a speed limit of 70 km/h (M70)
- Major roads with a speed limit of 50 km/h (M50)
- Other roads

To simplify, the stratification in each stage is hereafter generally ignored, and all stated sample sizes will refer to one stratum. We also ignore the fact that in stage one, the three largest PSUs (Stockholm, Göteborg, and Malmö) define a take-all stratum [29, p. 465]. The subsequent sampling stages in the take-all stratum differ somewhat from those described below.

### Selection of the master sample

The PSUs are the $N_I$ population centers in Sweden, labeled $i = 1, ..., N_I$. For simplicity, we represent the $i$th PSU by its label $i$. Thus, we denote the set of PSUs as $U_I = \{1, ..., i, ..., N_I\}$. Population center $i \in U_I$ is partitioned into $N_{IIi}$ small areas, labeled $q = 1, ..., N_{IIi}$, that represent the SSUs. Again we represent the sampling units by their labels; hence, the set of SSUs formed by the partitioning of $i$ is denoted $U_{IIi} = \{1, ..., q, ..., N_{IIi}\}$.

The master sample of small areas was selected in the following way:

**Stage I** A pps sample of PSUs was drawn with probability proportional to the number of inhabitants. At every draw, $p_i$ was the probability of selecting the $i$th PSU. Let $i_\nu$ denote the PSU selected in the $\nu$th draw, $\nu = 1, ..., m_I$, where $m_I$ is the number of draws. The probability of selecting $i_\nu$ is denoted $p_{i_\nu}$. For $\nu = 1, ..., m_I$ and $i \in U_I$, if the $i$th PSU was selected in the $\nu$th draw, then $p_{i_\nu} = p_i$. The vector of selected PSUs, $(i_1, ..., i_\nu, ..., i_{m_I})$, is the resulting ordered sample $os_I$.

**Stage II** For every $i_\nu$ that is a component of $os_I$, simple random sampling without replacement (SI) was used to draw a sample $s_{II i_\nu}$ of SSUs of size $n_{II i_\nu}$.[1]

In practice, the sample sizes in each stage were $m_I = 10$ and $n_{II i_\nu} = 1$. The resulting sample of SSUs is the master sample.

In general, the use of a multi-stage sampling design may imply some quite laborious variance estimation formulae. Sampling with replacement in the first sampling stage, however, makes variance estimation an easy matter. Possibly, the simplicity is gained at the cost of a larger variance.

**Selection of the final-stage sample for the speed survey**

The road network in small area $q$ in population center $i$ is viewed as partitioned into $N_{iq}$ one-meter road sites which represent the tertiary sampling units (TSUs). This set of TSUs is denoted by $U_{iq}$. In the speed survey, the final sampling stage is as follows:

**Stage III** An SI sample $s_{i_\nu q}$ of TSUs of size $n_{i_\nu q}$ is drawn for every small area $q \in s_{II i_\nu}$.

In practice, the sample sizes in the third sampling stage are $n_{i_\nu q} = 1$. The sample of road sites finally obtained is denoted $s$.

[1] SSUs were actually selected with pps within one stratum (Residential areas). In this thesis, for simplicity, this exception is generally ignored.

## 2.5 Data collection and processing

A sampled road site is positioned a certain number of meters into a road link. The field staff search out the site and install measurement equipment to collect data during a selected 24-hour period. The equipment consists of two pneumatic tubes stretched across the road in parallel, a fixed distance apart, and connected to a traffic analyzer. When a wheel crosses a tube, this changes the air pressure in the tube. The times of such events, or *pulses*, are registered by the traffic analyzer. From the resulting pulse stream, the analyzer creates vehicles and assigns speeds to them. The variables of interest are later calculated from the vehicle data produced by the traffic analyzer.

## 2.6 Estimation in the absence of nonsampling errors

Here, an estimator $\hat{t}_a$ of the population total

$$t_a = \sum\nolimits_U a_k = \sum\nolimits_{U_I} t_{ai} = \sum\nolimits_{U_I} \sum\nolimits_{U_{IIi}} t_{aiq},$$

where $t_{ai} = \sum_{U_{IIi}} t_{aiq}$ and $t_{aiq} = \sum_{U_{iq}} a_k$, will be presented, along with an estimator $\hat{R}$ of $R$. The variances of $\hat{t}_a$ and $\hat{R}$, and the components of these variances due to each sampling stage, will also be given.

In the speed survey, PSUs are selected *with* replacement and SSUs and TSUs *without* replacement. In order to construct estimators which are unbiased with respect to all three sampling stages, the 'p-expanded with replacement' estimation principle (first used by Hansen and Hurwitz [15]; treated for instance in [29, Section 2.9]), and the Horvitz-Thompson estimation principle (usually ascribed to Horvitz and Thompson [16]; treated extensively in [29]) are combined. Throughout this thesis, estimators of population entities are denoted by a hat, and the subscripts 'pwr' and '$\pi$' used to indicate 'p-expanded with replacement' estimators and Horvitz-Thompson estimators, respectively.

In the ideal situation, in which $a_k$ are known for all $k \in s$, the parameter

$t_a$ would be estimated by

$$\hat{t}_a = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{\hat{t}_{\pi a i_\nu}}{p_{i_\nu}} \tag{2.1}$$

where $\hat{t}_{\pi a i_\nu} = (N_{II i_\nu}/n_{II i_\nu}) \sum_{s_{II i_\nu}} \hat{t}_{\pi a i_\nu q}$ and $\hat{t}_{\pi a i_\nu q} = (N_{i_\nu q}/n_{i_\nu q}) \sum_{s_{i_\nu q}} a_k$. If $i \in U_I$ was selected in the $\nu$th draw, then $\hat{t}_{\pi a i_\nu} = \hat{t}_{\pi a i}$ and $\hat{t}_{\pi a i_\nu q} = \hat{t}_{\pi a i q}$. The estimator of $R$ would be

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}. \tag{2.2}$$

We call $\hat{t}_a$ and $\hat{R}$ the 'prototype' estimators of $t_a$ and $R$, respectively. The randomness in these estimators stems solely from the sample selection.

Let $E_p$ and $V_p$ denote expectation and variance with respect to the sampling design $p$ described in Section 2.4. For nonlinear estimators, such as the ratio of two estimated population totals, it is the practice to use the variance of a linearized statistic as an approximation to the exact variance. Let $AV_p$ denote such an approximative variance, again with respect to $p$. (For details on the linearization technique, see [29, Section 5.5].) In some parts of the thesis, we need to refine the notation regarding the sampling design. Expectations and variances are then indicated by subscript $I$ if taken with respect to the design used in stage one; by $II$ if taken with respect to the design used in stage two (given $os_I$); and by $III$ if taken with respect to the design used in stage three (given $os_I$ and $s_{II i_\nu}$).

The properties of $\hat{t}_a$ and $\hat{R}$ will now be investigated.

From [29, Result 4.5.1], $\hat{t}_a$ is unbiased for $t_a$. The variance of $\hat{t}_a$ is given by

$$V_p(\hat{t}_a) = \frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left( \frac{t_{ai}}{p_i} - t_a \right)^2 + \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_{ai}}{p_i} \tag{2.3}$$

where $V_{ai}$ is the variance of $\hat{t}_{\pi a i}$ with respect to the last two sampling stages:

$$V_{ai} = V_{aIIi} + \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} V_{aiq} \tag{2.4}$$

where

$$V_{aIIi} = N_{IIi}^2 \frac{1 - f_{IIi}}{n_{IIi}} S_{t_a U_i}^2; \quad f_{IIi} = n_{IIi}/N_{IIi};$$
$$S_{t_a U_i}^2 = \sum\nolimits_{U_{IIi}} (t_{aiq} - t_{ai}/N_{IIi})^2 / (N_{IIi} - 1)$$

for $i \in U_I$, and

$$V_{aiq} = N_{iq}^2 \frac{1 - f_{iq}}{n_{iq}} S_{aU_{iq}}^2; \quad f_{iq} = n_{iq}/N_{iq};$$
$$S_{aU_{iq}}^2 = \sum\nolimits_{U_{iq}} \left(a_k - t_{aiq}/N_{iq}\right)^2 / \left(N_{iq} - 1\right)$$

for $q \in U_{IIi}; i \in U_I$.

Similarly as in [29, Result 4.4.11], the variance $V_p\big(\hat{t}_a\big)$ can be written as the sum of three components, mirroring the variation arising from each sampling stage:

$$V_p\big(\hat{t}_a\big) = V_{\text{PSU}}\big(\hat{t}_a\big) + V_{\text{SSU}}\big(\hat{t}_a\big) + V_{\text{TSU}}\big(\hat{t}_a\big) \tag{2.5}$$

where

$$V_{\text{TSU}}\big(\hat{t}_a\big) = \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} V_{aiq} \tag{2.6}$$

expresses the variance due to the third-stage sampling of TSUs,

$$V_{\text{SSU}}\big(\hat{t}_a\big) = \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_{aIIi}}{p_i} \tag{2.7}$$

the variance due to the second-stage sampling of SSUs, and

$$V_{\text{PSU}}\big(\hat{t}_a\big) = \frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left(\frac{t_{ai}}{p_i} - t_a\right)^2 \tag{2.8}$$

the variance due to the first-stage sampling of PSUs.

Now consider the estimator $\hat{R}$. From [27, Section 6.8.2], $\hat{R}$ is approximately unbiased for $R$. Define the new study variable $E = y - Rz$. The approximate variance of $\hat{R}$ is given by

$$\begin{aligned} AV_p\Big(\hat{R}\Big) &= \frac{1}{t_z^2} V_p\big(\hat{t}_E\big) \\ &= \frac{1}{t_z^2} \left[\frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left(\frac{t_{Ei}}{p_i} - t_E\right)^2 + \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_{Ei}}{p_i}\right] \\ &= \frac{1}{t_z^2} \left(\frac{1}{m_I} \sum_{i=1}^{N_I} \frac{t_{Ei}^2}{p_i} + \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_{Ei}}{p_i}\right) \end{aligned} \tag{2.9}$$

where the last equality holds since $t_E = 0$. The variance $V_{Ei}$ is obtained from Equation (2.4) by letting the variable $a$ equal $E$.

Like $V_p\big(\hat{t}_a\big)$, the approximate variance of $\hat{R}$ can be decomposed by sampling stage:

$$\begin{aligned} AV_p\Big(\hat{R}\Big) &= AV_{\text{PSU}}\Big(\hat{R}\Big) + AV_{\text{SSU}}\Big(\hat{R}\Big) + AV_{\text{TSU}}\Big(\hat{R}\Big) \\ &= \frac{V_{\text{TSU}}\big(\hat{t}_E\big)}{t_z^2} + \frac{V_{\text{SSU}}\big(\hat{t}_E\big)}{t_z^2} + \frac{V_{\text{PSU}}\big(\hat{t}_E\big)}{t_z^2} \end{aligned} \tag{2.10}$$

where $V_{\text{TSU}}\big(\hat{t}_E\big)$, $V_{\text{SSU}}\big(\hat{t}_E\big)$ and $V_{\text{PSU}}\big(\hat{t}_E\big)$ are obtained from Equations (2.6) to (2.8) by (again) letting $a$ equal $E$.

# Chapter 3

# Frame errors

## 3.1 Introduction

The construction of small area frames was described in Section 2.4.2. This chapter is concerned with the link lengths being subject to measurement errors. By use of a simple error model, the impact of erroneous frame link lengths on the bias and variance of the survey estimators is examined. In our model, the total frame road length for a small area is viewed as a function of the true length and a random error. Data from an empirical study of the errors in the frame are used to evaluate the model and estimate the error parameters.

All possible nonsampling errors aside from erroneous frame link lengths (including all other possible frame imperfections) are here ignored. In particular, we assume that the small area frames list all the links in the areas correctly.

## 3.2 The frame error problem

It is not quite obvious how the problem of erroneous frame link lengths should be examined. One may say that the frame suffers from coverage errors, or that, due to faulty auxiliary information in the frame, incorrect element-inclusion probabilities are used.

**The coverage error view** Apart from rounding errors, a link length cor-

responds to a geographically ordered vector of population elements. If a road link is shorter in the frame than in reality, this corresponds to an undercoverage of target elements. Correspondingly, if a frame link is too long, the frame suffers from overcoverage.

**The incorrect inclusion probabilities view** The length is known for each frame unit (road link) prior to sampling; thus, length can be thought of as an auxiliary variable. If a road link is shorter or longer in the frame than in reality, this corresponds to an incorrect auxiliary variable value. For a given small area, the sum of all link lengths in the frame is supposed to be the number of road sites that make up the road network (the population). If this summed length is in error, but the sample of road sites actually is selected from the target population, the inclusion probabilities that are used for sampled road sites are incorrect.

The latter view is somewhat more general, since incorrect inclusion probabilities may arise for other reasons in other types of surveys. However, for our purposes, it does not really matter how we decide to entitle the problem.

Discrepancies between measured and actual link lengths have implications on the data collection stage of the survey. This follows since, in the presence of erroneous frame link lengths, the instructions to the field staff may no longer hold. Field staff are told to seek out a sampled road site located a certain number of meters into a specified link. In reality, the site may simply not exist if the link is shorter than what the frame indicates. If the link in reality is much longer than what the frame indicates, the site will indeed exist, but at different places depending on the direction from which the link is entered. In each case, the field staff adjust to real-life conditions by observing the traffic 'somewhere' along the designated link.

As far as we know, most studies of traffic characteristics are based on nonprobability samples. Instead of choosing road sites at random from a frame, efforts are made (by visual inspection of the road) to pick 'representative' sites for observation. It is therefore not very surprising that we have not seen this frame problem treated in the traffic research literature. The statistical literature on frame errors, on the other hand, mainly deals with errors in sampling frames used in surveys of individuals or households. The

conditions of such surveys differ substantially from those in the speed survey, so the methods suggested for evaluating the impact of coverage errors are not quite applicable to our problem.

Our work is, however, inspired by the approaches to measurement errors discussed by, among others, Biemer and Stokes [2] and Särndal et al. [29, Chapter 16]. In this field of research, a survey is viewed as a two-stage process such that each stage contributes with randomness to the estimators. The first stage, the sample selection, determines what part of the population to observe. The second stage is the measurement procedure, which generates an observation for each element in the sample. Unlike traditional sampling theory, the observations are not presupposed to coincide with the true values, but assumed to be subject to random errors. In order to evaluate the impact of measurement errors on the estimators, the relation between observed and true values is modeled.

## 3.3 A frame error model

If the road lengths in the final-stage frames are in error, the actual sampling procedure differs from the one described in Section 2.4.3. Let $U_{Fiq}$ denote the set of road sites (of size $N_{Fiq}$) in $(i, q)$ *according to the frame*. For every small area $q \in s_{II i_\nu}$, an SI sample $s_{Fi_\nu q}$ of sites (of size $n_{i_\nu q}$) is drawn from $U_{Fiq}$. In the data-collection stage, the field staff adjust to the real road network when installing the measurement equipment. Consequently, the set of sites actually observed may differ from $s_{Fi_\nu q}$. We do not, however, introduce any special notation to distinguish between these sets. The sample of sites finally obtained (as well as the sample finally observed) is denoted by $s_F$.

Our frame error model, which we denote by $\mathrm{m}_1$, is formulated as follows:

(1) The sample $s_{Fiq}$ is an SI sample from $U_{iq}$. In mathematical terms, we assume that $s_{Fiq} = s_{iq}$.

(2) The frame road length $N_{Fiq}$ is a function of the true length $N_{iq}$ and a random error $\zeta_{iq}$.

(3) All $N_{Fiq}$'s are independent random variables with expected values $\mu_{iq}$ and variances $\sigma^2_{iq}$.

In cases of unclear instructions due to frame errors, the field staff place the measurement equipment 'somewhere' along designated links. Then, Assumption (1) holds if the road sections within the link can be considered 'randomly ordered,' or if the field staff randomly choose a road section within the designated link for measurement. The field staff's choice of a road section within the link is probably more adequately described as haphazard than as random. This follows since, when deciding upon a location, they pay regard to the road environment (e.g., by avoiding locations where cars parked by the roadside may obstruct the installation of the equipment). A 'random ordering' of the road sections is however, for the following reason, quite likely. As described in Section 2.2, only road sections located more than one hundred meters from an intersection are included in the target population. Results from a pilot study [3] suggest that, within a link, the remaining road sections are reasonably similar with respect to the study variables. Consequently, it is not crucial which road section within a link is actually measured – the result will be about the same anyway.

## 3.4 Estimation with frame errors

We now investigate the influence of the road length error on the estimation of $t_a$ and $R$.

### 3.4.1 The estimators $\hat{t}_{Fa}$ and $\hat{R}_F$

Define the population totals $t_{Faiq} = \sum_{U_{Fiq}} a_k$, $t_{Fai} = \sum_{U_{IIi}} t_{Faiq}$ and $t_{Fa} = \sum_{U_I} t_{Fai}$. Further, let $R_F = t_{Fy}/t_{Fz}$, $t_{FEiq} = t_{Fyiq} - R_F t_{Fziq}$ and $t_{FEi} = t_{Fyi} - R_F t_{Fzi}$.

The estimator of $t_a$ obtained by replacing $N_{iq}$ with $N_{Fiq}$ in Equation (2.1) is given by

$$\hat{t}_{Fa} = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{\hat{t}_{F\pi ai_\nu}}{p_{i_\nu}} \tag{3.1}$$

where $\hat{t}_{F\pi ai_\nu} = (N_{IIi_\nu}/n_{IIi_\nu}) \sum_{s_{IIi_\nu}} \hat{t}_{F\pi ai_\nu q}$ and $\hat{t}_{F\pi ai_\nu q} = (N_{Fi_\nu q}/n_{i_\nu q}) \sum_{s_{Fi_\nu q}} a_k$. If $i \in U_I$ was selected in the $\nu$th draw, then $\hat{t}_{F\pi ai_\nu} = \hat{t}_{F\pi ai}$ and $\hat{t}_{F\pi ai_\nu q} = \hat{t}_{F\pi aiq}$. The corresponding estimator of $R$ is

$$\hat{R}_F = \frac{\hat{t}_{Fy}}{\hat{t}_{Fz}}. \tag{3.2}$$

Consider the special case when, for every small area $q$ included in the master sample, the frame road length $N_{Fiq}$ equals the true length $N_{iq}$, and $s_{Fiq}$ is an SI sample from $U_{iq}$. Then, the estimators $\hat{t}_{F\pi aiq}$, $\hat{t}_{F\pi ai}$ and $\hat{t}_{Fa}$ are design-unbiased for $t_{aiq}$, $t_{ai}$ and $t_a$, respectively, and the index $F$ is no longer needed.

Let expectations and variances be indicated by subscript $\mathrm{m}_1$ if taken with respect to model $\mathrm{m}_1$; by subscript $p\mathrm{m}_1$ if taken with respect jointly to the sampling design $p$ and model $\mathrm{m}_1$. The statistical properties of $\hat{t}_{Fa}$ and $\hat{R}_F$ are investigated in the following theorem:

**Theorem 3.4.1** *Jointly under sampling design $p$ in Section 2.4 and model $m_1$, the expected value of $\hat{t}_{Fa}$ is given by*

$$E_{pm_1}(\hat{t}_{Fa}) = \sum\nolimits_{i=1}^{N_I} E_{pm_1}(\hat{t}_{F\pi ai}) \tag{3.3}$$

*where $E_{pm_1}(\hat{t}_{F\pi ai}) = \sum_{U_{IIi}} (\mu_{iq}/N_{iq})\, t_{aiq}$. The variance of $\hat{t}_{Fa}$ is given by*

$$\begin{aligned} V_{pm_1}(\hat{t}_{Fa}) = {} & \frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left( \frac{1}{p_i} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{aiq} - \sum\nolimits_{i=1}^{N_I} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{aiq} \right)^2 \\ & + \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_{pm_1}(\hat{t}_{F\pi ai})}{p_i} \end{aligned} \tag{3.4}$$

*where*

$$\begin{aligned} V_{pm_1}(\hat{t}_{F\pi ai}) = {} & N_{IIi}^2 \frac{1 - f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi} - 1} \sum\nolimits_{U_{IIi}} \left( \frac{\mu_{iq}}{N_{iq}} t_{aiq} - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{aiq} \right)^2 \\ & + \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left( \frac{\mu_{iq}}{N_{iq}} \right)^2 V_{aiq} \\ & + \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left( \frac{\sigma_{iq}}{N_{iq}} \right)^2 (V_{aiq} + t_{aiq}^2)\,. \end{aligned}$$

*The estimator $\hat{R}_F$ is approximately unbiased for $E_{pm_1}(\hat{t}_{Fy})/E_{pm_1}(\hat{t}_{Fz})$, with the approximate variance*

$$AV_{pm_1}\left(\hat{R}_F\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\sum\nolimits_{U_{IIi}}\left(\frac{\mu_{iq}}{N_{iq}}\right)^2 t_{Eiq}^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{pm_1}(\hat{t}_{F\pi Ei})}{p_i}\right\} \tag{3.5}$$

*where $V_{pm_1}(\hat{t}_{F\pi Ei})$ is obtained from $V_{pm_1}(\hat{t}_{F\pi ai})$ by replacing the variable $a$ with $E$.*

The proof is given in Appendix B.1.2.

From Theorem 3.4.1, results can be derived for various situations of interest. An important special case is when the frame road lengths $N_{Fiq}$ are 'unbiased' – that is, if in a (hypothetical) long run of repeated length measurements on the same small area road network, the average of the obtained values will equal the true value $N_{iq}$. This seems to us a quite likely situation. The major sources of errors in the measurements are probably the map measurer tool producing 'shaky' results and the haste under which the measurements were performed. We have no reason to believe that these errors have a systematic influence on the measurement values.

The case of unbiased frame road lengths is treated in the following corollary (which is easily derived from Theorem 3.4.1 by replacing $\mu_{iq}$ with $N_{iq}$):

**Corollary 3.4.1** *If the frame road lengths $N_{Fiq}$ have expected value $N_{iq}$, the estimator $\hat{t}_{Fa}$ is unbiased for $t_a$, and $\hat{R}_F$ is approximately unbiased for $R$. The use of $\hat{t}_{Fa}$ instead of $\hat{t}_a$ as estimator of $t_a$ increases the variance by*

$$\begin{aligned} &V_{pm_1}(\hat{t}_{Fa}) - V_p(\hat{t}_a) \\ &= \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\left(\frac{\sigma_{iq}}{N_{iq}}\right)^2\left(V_{aiq}+t_{aiq}^2\right). \end{aligned} \tag{3.6}$$

*The variance increase due to the use of $\hat{R}_F$ instead of $\hat{R}$ as estimator of $R$ is obtained from Equation (3.6) by multiplying by $t_z^{-2}$ and letting $a = E$.*

### 3.4.2 Results for specific error structures

Under Assumption (1) of the frame error model (see Section 3.3), the only remaining difference between $\hat{t}_{\pi aiq}$ and the error-prone estimator $\hat{t}_{F\pi aiq}$ is that the latter is weighted by $N_{Fiq}$ instead of $N_{iq}$. A random error model for $N_{Fiq}$ is stated in Assumptions (2) and (3), but to be really useful the model needs further specification. The two most simple error structures are the additive error model,

$$N_{Fiq} = N_{iq} + \zeta_{iq} \tag{3.7}$$

and the multiplicative error model,

$$N_{Fiq} = N_{iq}\zeta_{iq}. \tag{3.8}$$

We denote the expected value and variance of the random error $\zeta_{iq}$ with $\theta_{iq}$ and $\tau_{iq}^2$, respectively. Then, under the additive error model, $\mu_{iq} = N_{iq} + \theta_{iq}$ and $\sigma_{iq}^2 = \tau_{iq}^2$, whereas under the multiplicative error model, $\mu_{iq} = N_{iq}\theta_{iq}$ and $\sigma_{iq}^2 = N_{iq}^2\tau_{iq}^2$. Note that, depending on the assumed error structure, $\theta_{iq}$ and $\tau_{iq}^2$ are expected to take quite different numerical values. Consider, for instance, the case when the road length measurements are rather accurate, so that $\mu_{iq}$ approximately equals $N_{iq}$. Under the additive error model, this occurs when $\theta_{iq}$ is close to zero; under the multiplicative error model when $\theta_{iq}$ is close to one.

It is straightforward to adapt Theorem 3.4.1 to various error structures of interest. By replacing $\mu_{iq}$ with $N_{iq}+\theta_{iq}$ and $\sigma_{iq}^2$ with $\tau_{iq}^2$, results are obtained for the additive error model in (3.7), whereas by replacing $\mu_{iq}$ with $N_{iq}\theta_{iq}$ and $\sigma_{iq}^2$ with $N_{iq}^2\tau_{iq}^2$, we get results for the multiplicative error model in (3.8).

In the remainder of this section, we will only look at the model we a priori believe to be the most realistic: the multiplicative error model with equal error expectations $\theta$ and variances $\tau^2$. The multiplicative error model means that the error associated with $N_{Fiq}$ depends on the true length $N_{iq}$ – a view we regard as intuitively appealing. For example, it is probably harder to obtain accurate measurements for areas with extensive road networks, since such networks usually are partitioned into a large number of links. (Remember that each link length was measured separately.) Further, we have no reason

to believe the error expectations and variances to differ between population centers or small areas. The same tool, a map measurer, was used everywhere, and the staff performing the measurements were given the same training. An important objective for the multiplicative model is that it states that the variances of the frame road lengths, $\sigma_{iq}^2$, are proportional to the squared true lengths. It is not obvious that this assumption holds; an equally natural assumption is that the variances are proportional to the (unsquared) lengths.

For the assumed model, the following corollary applies:

**Corollary 3.4.2** *If the frame road lengths $N_{Fiq}$ have expected value $N_{iq}\theta$, the bias of $\hat{t}_{Fa}$ as estimator of $t_a$ is given by $t_a(\theta - 1)$.*

Corollary 3.4.2 is easily derived from Theorem 3.4.1 by replacing $\mu_{iq}$ with $N_{iq}\theta$. Note that if $\theta$ equals one, $\hat{t}_{Fy}$ and $\hat{R}_F$ are unbiased and approximately unbiased, respectively, for their true counterparts.

Let us proceed by investigating the variances when $\theta$ equals one.

**Corollary 3.4.3** *If the frame road lengths $N_{Fiq}$ have expected value $N_{iq}$ and variance $N_{iq}^2\tau^2$, the use of $\hat{t}_{Fa}$ instead of $\hat{t}_a$ as estimator of $t_a$ increases the variance by*

$$V_{pm_1}(\hat{t}_{Fa}) - V_p(\hat{t}_a) = \tau^2 \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left(V_{aiq} + t_{aiq}^2\right). \tag{3.9}$$

*The approximate variance increase due to the use of $\hat{R}_F$ instead of $\hat{R}$ as estimator of $R$ is given by*

$$AV_{pm_1}\left(\hat{R}_F\right) - AV_p\left(\hat{R}\right) = \tau^2 AV_p\left(\hat{R}\right). \tag{3.10}$$

The proof of Corollary 3.4.3 is given in Appendix B.2.

### 3.4.3 Summary of theoretical findings

We have argued for our belief that the measured link lengths are unbiased for the corresponding true lengths. If our expectation is correct, the length error does not introduce bias in the estimators – a very encouraging result.

Of course, there will still be a loss of precision due to the variability of the frame lengths.

The length of a small area road network may be viewed as a measure of the degree of difficulty of the measurement task. With this view, the multiplicative error model makes sense. Analytically, things get especially simple if these errors have the same expectations and variances; this also seems like a realistic assumption. For this case, 'unbiased' road length measurements corresponds to an error expectation equal to one. If this is fulfilled, the length error implies a relative variance increase in the estimator of average speed that is simply equal to the error variance. This variance is likely to be numerically small, since the multiplicative errors are 'relative.'

## 3.5 Empirical study

### 3.5.1 Study objectives

In Chapter 3.4, by use of an error model, we investigated theoretically the impact of erroneous frame road lengths on the estimators $\hat{t}_{Fy}$ and $\hat{R}_F$. Hence a theoretical foundation is laid, but it needs to be complemented by knowledge about the real road length errors in the frame. Then, a choice of a realistic error structure can be made, the constant error expectations and variances assumption can be evaluated and, if proved to hold, $\theta$ and $\tau^2$ can be estimated. To gain this knowledge, we conducted an experiment, the design and analysis of which we now present.

### 3.5.2 Design of the study

Data on the frame road length errors were collected in the following way. From the 469 small areas included in the master sample, 70 small areas were selected. A controller measured all the links in selected areas and fed the results into computer files. In the course of the work, the controller had access only to the originally used maps with the intersections numbered. Hence, for a small area, she started by making a list of all the links found on the map (using the existing numbering) and then measured them one after

the other.

In the selection of small areas for the experiment, we wanted areas from different SNRA regions and from population centers of various sizes. Furthermore, we wanted the areas to represent different development types.

Note that SNRA region, population center size class, and development type were all used as stratification variables in the sample selection (see Section 2.4.3). An SNRA region effect was possible since, when the frame was constructed, each regional office was responsible for the work in its region, including the length measurements. Population center size and development type may correlate with the quality of available maps. To accomplish the desired dispersion of small areas, they were randomly selected within SNRA region, population center size class, and small area stratum.

For at least two reasons, the measurement values obtained in the study are probably more accurate than the frame values. Above all, when the frame was constructed, the road length measurements were made hurriedly (the entire construction work was behind schedule). Our controller was not put under time pressure; on the contrary, she was encouraged to give priority to carefulness and to take her time. Also, when the frame was constructed, the road lengths were determined by use of a digital map measurer. This tool is convenient to use, since it can be programmed to produce length data in meters for a map with a specified scale. In our experience, however, the tool is over-sensitive to the user's hand movements. The controller used a less sophisticated instrument, a common ruler, which we believe is less subject to measurement errors.

### 3.5.3 Data processing

For five of the chosen areas, the available maps were of such poor quality that the links could not be identified or measured properly. Therefore, those areas were entirely omitted from the analysis. From each remaining area, we excluded the links known to be administered by the state, as well as road links that did not occur both in the frame and in the controller's list. In practice, we applied (in turn) the following rules for excluding road links:

1. Road links, found in control, that are missing in the frame.

| | Number | Percent of original no. |
|---|---|---|
| Links in original data set | 4123 | 100 |
| Left after rule 1 applied | 4013 | 97.3 |
| Left after rule 2 applied | 3762 | 91.2 |
| Left after rule 3 applied | 3618 | 87.8 |

Table 3.1: Exclusion of road links.

| Population center size class | 1 | | | | 2 | | | | 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Small area development type | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Central Region | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| Mälardalen Region | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | - | 1 | 1 | 1 |
| Northern Region | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | - | 1 | 2 | 1 |
| Skåne Region | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | - | - | 2 | 1 |
| Stockholm Region | 0 | 2 | 0 | 0 | 1 | 1 | 1 | 1 | - | 1 | 1 | 1 |
| South-Eastern Region | 2 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | - | - | 4 | - |
| Western Region | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | - | - | 1 | - |

Table 3.2: Number of small areas included in the analysis, by SNRA region, population center size class and small area development type. Non-existing strata are indicated by hyphens.

2. Road links that, according to the frame, are state authority roads.

3. Road links included in the frame that, according to the control, do not exist.

The resulting gradual reduction of the original data set (the set of all links occurring either in the frames or in the controller's lists) is shown in Table 3.1. The allocation of the 65 small areas over SNRA regions, population center size classes and small area development types is shown in Table 3.2. In the table, the following numbering of size classes is used: '1' for large major population center of a municipality, '2' for other major population center of a municipality, and '3' for other population center. Also, the small area development types are assigned the numbers '1' for city, '2' for industrial, '3' for residential, and '4' for other areas.

In the data processing, we encountered several frame quality problems other than erroneous road lengths. First, remember that we had to give up five chosen areas because of bad maps. Most likely, the frames in use for these areas are not, in general, very reliable. Second, we see in Table 3.1 that 110 links turned out to be missing in the frame and that 144 urban road links that were included in the frame could not be found by the controller. We take these figures as a warning signal that the frame may suffer from some serious coverage errors regarding road links. Finally, as a result of incorrect frame link lengths, some links may erroneously be excluded from or included in the target population. Among the non-major road links in our reduced data set, 42 links were shorter than 200 meters in the frame but longer than 200 meters in the control, while 34 links were longer than 200 meters in the frame but shorter than 200 meters in the control.

Like erroneous frame road lengths, all the frame imperfections discussed above may lower the quality of the survey estimates. In this thesis, we restrict our attention solely to the length problem. An expanded study would be needed in order to judge the influence and relative importance of all frame imperfections on the total error of the estimates.

### 3.5.4 Analysis

Assume that the road link lengths according to the control are the true lengths. Then, by summing the frame link lengths for a small area $(i, q)$, we get an observation on $N_{Fiq}$, and by summing the link lengths according to the control, we get $N_{iq}$. Under the additive error model in Equation (3.7), the error in the frame road length $N_{Fiq}$ is given by $\zeta_{iq} = N_{Fiq} - N_{iq}$, whereas under the multiplicative error model in Equation (3.8), the error is given by $\zeta_{iq} = N_{Fiq}/N_{iq}$. For the 65 small areas comprised by our analysis, the errors were calculated under both the additive and the multiplicative error model (see Figure 3.1). We see that in the additive case, the points scatter around an imaginary horizontal line placed at a level close to zero, whereas in the multiplicative case, the scatter is around a line at a level close to one. Hence, under both error models, data suggest that the frame road lengths, on the average, are correct. In the additive case, the variance for the scatter

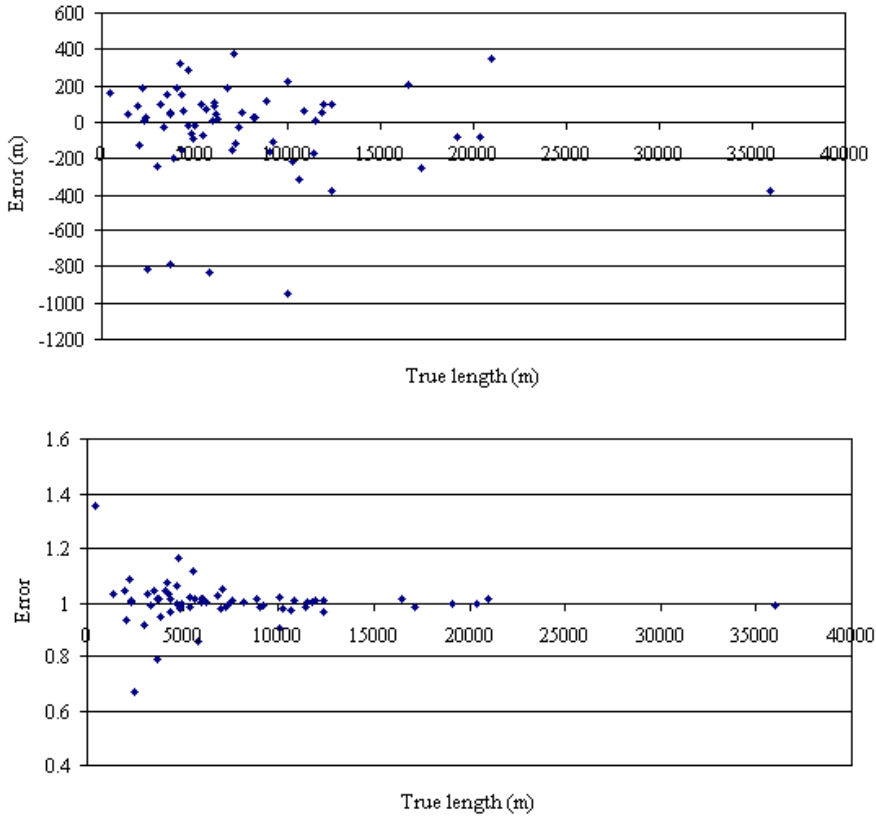


Figure 3.1: Observed errors under additive error model (top) and multiplicative error model (bottom).

of points seems constant, exactly as we had hoped for. In the multiplicative case however, the point scatter shows a tendency to narrow with the true length. This is a sign that the variance of the frame road lengths rather is proportional to the true length than to the squared true length (as the model states). However, due to the shortage of observations for large values of the true length, it is hard to draw any certain conclusions.

In our study design, population center size classes are nested under the SNRA regions, and small area strata are nested under the size class levels. Thus, it is a three-stage nested design (see, e.g., [25]). To account for the design, we introduce some new notation. Consider again the length error for small area $(i, q)$, $\zeta_{iq}$. Let $\zeta_{iq} = \zeta_{rstq}$ if population center $i$ is included in SNRA region $r$ and size class $s$, and small area $q$ is included in small area stratum (development type) $t$. The analysis of variance (ANOVA) model for our design is

$$\zeta_{rstq} = \alpha + \beta_r + \gamma_{s(r)} + \delta_{t(rs)} + \epsilon_{(rst)q} \tag{3.11}$$

where $\alpha$ is an overall mean, $\beta_r$ is the random effect of the $r$th region, $\gamma_{s(r)}$ is the random effect of size class $s$ within the $r$th region, $\delta_{t(rs)}$ is the random effect of small area stratum $t$ within size class $s$ within the $r$th region, and $\epsilon_{(rst)q}$ is a random error.

Each of the factors – region, size class, and small area stratum – has a small number of possible levels (7, 3, and 4, respectively). Nevertheless we consider these factors as random. Regarding the regional factor, we are not interested in the administrative division in itself, but rather in potential differences in the behavior of the staff. Hence, we view the seven SNRA regions as a selection of levels from a population of behavior levels. Correspondingly, we are not interested in the divisions in size classes or small area strata, but in potential differences in the quality of the maps.

Assume that $\beta_r$, $\gamma_{s(r)}$, $\delta_{t(rs)}$ and $\epsilon_{(rst)q}$ are independent, with variances $\sigma^2_\beta$, $\sigma^2_\gamma$, $\sigma^2_\delta$ and $\sigma^2_\epsilon$, respectively. We would like to test if these variances are zero. That is, we want to know whether variability exists in the length errors that is due to SNRA region, population center size class, or small area stratum. We do not have enough data to perform such tests 'by the book,' but use instead a simplified (approximative) test procedure. To put it briefly, we

| Error structure | Sample statistics $\bar{\zeta}$ | Sample statistics $s_\zeta^2$ | 95% c.i. for $\theta$ | 95% c.i. for $\tau^2$ |
|---|---|---|---|---|
| Additive | $-18.262$ | 85481.915 | $[-89.340, 52.816]$ | $[0, 115296.998]$ |
| Multiplicative | 1.00209 | 0.00629 | $[0.98281, 1.02137]$ | $[0, 0.00848]$ |

Table 3.3: Sample statistics and confidence intervals (c.i.). The intervals for $\tau^2$ are upper-bounded.

look only at one effect at the time and ignore the nesting. This was done for each of the three effects and for the observed errors under both the additive and the multiplicative error model. The relevant ANOVA tables are given in Appendix E.1. In no case is the hypothesis of zero variance rejected. We take this as an indication that the variances $\sigma_\beta^2$, $\sigma_\gamma^2$, and $\sigma_\delta^2$ are all zero.

We proceed by viewing the observed length errors $\{\zeta_{iq}\}$ simply as independent and identically distributed (iid) random variables with mean $\theta = \alpha$ and variance $\tau^2 = \sigma_\zeta^2$. As unbiased estimators of $\theta$ and $\tau^2$, we use the sample mean $\bar{\zeta}$ and the sample variance $s_\zeta^2$, respectively. The resulting estimates are given in Table 3.3. The confidence intervals shown in the table hold under the added assumption of normally distributed errors. We see that under the additive error model, the hypothesis of $\theta = 0$ cannot be rejected. If in fact the hypothesis is true, Corollary 3.4.1 applies and the length error does not bias $\hat{t}_{Fy}$ or $\hat{R}_F$. Under the multiplicative error model, the hypothesis of $\theta = 1$ cannot be rejected. If this hypothesis is true, Corollary 3.4.2 tells us that the length error does not bias the estimators. We conclude that irrespective of which error structure we look at, our data do not suggest that the length error will cause bias in the estimators.

We are also interested in the possible variance increase due to the length error. Although the additive error model with equal error expectations and variances seems to fit the data somewhat better than the multiplicative counterpart (according to Figure 3.1), we choose the multiplicative model. The reason for this is simply that if the errors are multiplicative, Corollary 3.4.3 applies and we can easily estimate the approximate variance increase due to the use of $\hat{R}_F$ instead of $\hat{R}$. Since the observed errors are numerically quite small, the choice of model is not so crucial. If our point estimate of $\tau^2$ in

Table 3.3 coincides with the true parameter value, the relative variance increase is only about 0.6 percent. Hence, at least for the ratio, the variance increase seems negligible.

### 3.5.5 Summary of empirical findings

For the 65 small areas comprised by the analysis, we calculated the errors in $N_{Fiq}$ under the additive and the multiplicative error model. Under both models, data suggested that the frame road lengths, on the average, were correct. In the additive case, the error variance seemed constant as function of $N_{iq}$ (as the model states). In the multiplicative case, we saw some tendency of the error variance to decrease as $N_{iq}$ increased: a sign that the variance is proportional to $N_{iq}$ rather than to $N_{iq}^2$ (as the model states). Our tests did not suggest variability in the length error due to any stratification variables used in stages I and II. Irrespective of error model, our data did not suggest that the length error would bias the estimators of $t_a$ or $R$. Under the multiplicative error model, for the estimator of the ratio, the variance increase due to the length error seemed negligible.

## 3.6 Summary

If the road lengths in the frames used in the final stage of sample selection are in error, how are the statistical properties of the estimators affected? Our theoretical derivations, supported by an error model, resulted in expressions for the effects of the length error on the bias and variance of the estimators. In particular, we showed that if the errors were multiplicative with expectation of one and constant variance, the length error had no bias effect on the estimator of average speed, and the relative (approximate) variance increase for this estimator simply equalled the error variance. We also collected some data on the real errors in the frames. The observed errors were found to be quite small, and for simplicity we chose the multiplicative model, although the additive model actually had a slightly better fit. The multiplicative errors were found to have an expectation close to one, and their variance was estimated to less than one percent. Putting all this together, our investigation

led us to reach the following conclusions. First, neither the estimator of average speed nor the estimator of a total seems to be biased by the length error. Second, the variance increase due to the length error, for the average speed estimator, seems to be negligible.

It should be noted that our results are useful only if one trusts our model, since the entire investigation relies heavily upon it. The employed error model includes a very strong assumption: that the actual final-stage samples are selected by simple random sampling from the true road networks. For the future, we recommend that the data-collection instructions be given an overhaul. Improved instructions would increase the chances that the model assumption really holds. It should also be noted that the only frame imperfection considered in this study was the length error. The empirical study exposed several other imperfections associated with the last-stage frames which need to be addressed.

# Chapter 4

# Missing data

## 4.1 Introduction

The data collection procedure described in Section 2.5 is not unproblematic. In fact, a number of the vehicles passing a chosen site typically remain unobserved. If a large amount of data is missing, the site is likely to be remeasured at a later date. The most common situation, and the one of interest here, however, is that some data are missing but not to the extent that the measurement is disqualified. Currently, data are subsequently used in the estimation without any special action being taken. Attention is restricted here to incompleteness due to occasional loss of vehicles, thus ignoring cases of lost time periods. Possible nonsampling errors aside from missing data are also ignored.

In this chapter, two strategies for adjusting for missing data in the estimation stage of the survey are introduced. Both are designed for easy implementation: they do not require simulations, or collection of new auxiliary data, but only minor modifications of the computer programs presently used for estimation. The failure to observe some vehicles is indicated on one hand by imputations automatically created by the measurement device, on the other by the ME being small. One suggested strategy uses the imputations for adjustments; the other uses the ME for the same purpose. The two strategies rest, however, on a common model for the vehicle registration mechanism. By use of empirical data, the models are evaluated and the

suggested strategies compared.

## 4.2 The missing data problem

Missing data arise when the traffic analyzer fails to translate arrived pulses unambiguously into vehicles. The ambiguity may be caused, for instance, by vehicles simultaneously crossing the tubes (due to meetings or passings) or by a dense traffic situation. An undercount of vehicles is bound to bias the estimators of the totals, whereas the impact on the estimator of $R$ is unclear. It is possible, but far from certain, that (in practice) the biases in the estimators of the totals 'cancel out' when their ratio is taken.

The notation from Section 2.3 will now be slightly expanded. Let the set of vehicles passing road site $k$ (during a given time period) consist of $y_k$ vehicles labeled $v = 1, ..., y_k$. For simplicity, the $v$th vehicle is represented by its label $v$. Hence, the (finite) population of passing vehicles is denoted as $U_k = \{1, ..., v, ..., y_k\}$. The travel time $z_k$ for site $k$ is given by $z_k = \sum_{U_k} x_v$ where $x_v$ is the time vehicle $v$ takes to travel the site.[1] The successfully observed subset of $U_k$ is denoted $r_k$ of size $n_{r_k}$.

The present 'do nothing' approach to missing data is henceforth referred to as *Strategy 0*. The estimators of $y_k$ and $z_k$ under Strategy 0 are $\hat{y}_k^{(0)} = n_{r_k}$ and $\hat{z}_k^{(0)} = \sum_{r_k} x_v$, respectively.

## 4.3 Proposals for missing data adjustments

Both our proposals for missing data adjustments involve the estimation of registration probabilities for vehicles passing an observational site. We see no practicable way of estimating the probabilities for individual vehicles, but need some simplifying assumptions. It would be unreasonable to assume a constant registration probability for all vehicles, but it may make sense for groups of vehicles. This motivates the use of the registration model in Section 4.3.1 as a common starting point for our proposals.

[1] In practice, the $x_v$'s are calculated as the inverses of the registered vehicle speeds.

### 4.3.1 A model of the registration mechanism

The true registration distribution, which generates the set of registered vehicles $r_k$ for an observed road site $k$, is of course unknown. Our ambition here is only to formulate some reasonable model assumptions about this distribution. If we succeed, we will possess a useful tool for constructing (and evaluating) estimators that adjust for unregistered vehicles.

In all essentials, our registration model coincides with the response homogeneity group (RHG) model formulated in [28, Eq. (8.1)] or [29, Eq. (15.6.6)]. In brief, the model states that a realized sample can be partitioned into groups such that, *conditional on the sample*, the individual response probabilities are the same for all group members. The conditioning is motivated by the fact that elements of a given sample are exposed to a specific set of survey operations. The RHG model has a quite general formulation, and many weighting class adjustment methods rely on special cases of this model (for an overview of adjustment methods, see [24, Chapter 8]).

Our model has three special features, when contrasted with the RHG model. First, in the speed survey, data are collected by observing (registering) vehicles. Hence, instead of probability of response, our concern is about probability of *registration*. Second, road sites are selected for observation by a multi-stage procedure. Hence, our model is conditioned on the final-stage samples $s_{iq}$ of sites. Finally, we are not interested in observing a sample of the vehicles passing the site, but rather all of them.

Our registration model, which we denote by $\mathrm{m}_r$, is summarized below.

**The registration model, $\mathbf{m_r}$**
Assume that the vehicles passing road site $k \in s_{iq}$ during a selected day are partitioned into $H_k$ groups $U_{kh}$ $(h = 1, ..., H_k)$ such that, given $s_{iq}$,

- all vehicles in group $U_{kh}$ have the same (unknown) probability $\theta_{kh} > 0$ of being registered, and
- the registration of one vehicle is independent of all others.

The independent registrations assumption is made solely to simplify the model. In reality, dependencies in the registrations of successive vehicles are likely to occur.

The theoretical part of this chapter is applicable to any groups of traffic. In our experiment, however, we presume partitioning of the traffic by time intervals (see Section 4.5). This partitioning is easy to make and corresponds roughly to a partitioning by flow level. The shortest time unit considered is watch-hour. One reason for this is the common advice (see, e.g., [22]) to avoid too small weighting classes when estimating (response) probabilities $\theta_{kh}$ by class response rates. The response rates (in our case, the registration rates) for small classes tend to be unstable, and this may produce large variation in the weights. A second reason for our choice of smallest time unit is that we also try to estimate $\theta_{kh}$ by use of the ME, which is only known at watch-hour level.

The set of registered vehicles in group $U_{kh}$ is denoted $r_{kh}$ of size $n_{r_{kh}}$, and the vector of all $n_{r_{kh}}$'s is denoted $\mathbf{n}_{r_k} = \left(n_{r_{k1}}, ..., n_{r_{kh}}, ...n_{r_{kH_k}}\right)$. Expectation and variance taken with respect to the registration distribution $\mathrm{m}_r$, conditional on $s_{iq}$, is denoted $E_{\mathrm{m}_r}(\cdot\,|s_{iq})$ and $V_{\mathrm{m}_r}(\cdot\,|s_{iq})$, respectively. In Section 4.4.2, we make use also of the conditional expectation and variance with respect to all realizations $\mathbf{n}_{r_k}$ obeying $\sum_{h=1}^{H_k} n_{r_{kh}} = n_{r_k}$; $E_{\mathbf{n}_{r_k}}(\cdot\,|s_{iq})$ and $V_{\mathbf{n}_{r_k}}(\cdot\,|s_{iq})$. Then,

$$
\begin{aligned}
E_{\mathrm{m}_r}(\cdot\,|s_{iq}) &= E_{\mathbf{n}_{r_k}} E_{\mathrm{m}_r}(\cdot\,|s_{iq}, \mathbf{n}_{r_k}) \\
V_{\mathrm{m}_r}(\cdot\,|s_{iq}) &= E_{\mathbf{n}_{r_k}} V_{\mathrm{m}_r}(\cdot\,|s_{iq}, \mathbf{n}_{r_k}) + V_{\mathbf{n}_{r_k}} E_{\mathrm{m}_r}(\cdot\,|s_{iq}, \mathbf{n}_{r_k})\,.
\end{aligned}
$$

The conditional mean and variance of $n_{r_{kh}}$ given $s_{iq}$ are denoted $\mu_{r_{kh}} = E_{\mathrm{m}_r}(n_{r_{kh}}\,|s_{iq})$ and $\sigma^2_{r_{kh}} = V_{\mathrm{m}_r}(n_{r_{kh}}\,|s_{iq})$, respectively.

For future reference, some implications of the registration model will be stated:

1. Under model $\mathrm{m}_r$, given $s_{iq}$,

   $$(n_{r_{kh}}\,|s_{iq}) \sim \text{binomial}(y_{kh}, \theta_{kh})$$

   where $y_{kh}$ is the true number of vehicles in group $U_{kh}$. Hence, $\mu_{r_{kh}} = y_{kh}\theta_{kh}$ and $\sigma^2_{r_{kh}} = y_{kh}\theta_{kh}\,(1 - \theta_{kh})$.

2. If the vector $\mathbf{n}_{r_k}$ is conditioned upon as well, the set $r_k$ behaves as selected by stratified sampling with SI sampling in each stratum (STSI) from $U_k$.

### 4.3.2 Strategy 1

We are now ready for our first proposal for missing data adjustments. The idea here is to make use of the procedure for handling missing data which is already built into the traffic analyzer. From excess pulses, vehicles are created or *imputed*. The imputed vehicles are also assigned speeds, based on those of previously registered vehicles. For details on the stepwise, basically non-random imputation procedure, see [1].

At present, the survey management chooses to discard all imputed vehicles in the estimation. Why? The traffic analyzer, including its imputation algorithm, was developed back in the 1970s in order to meet the demands of that time: flow measurements on state roads. Today's speed survey is conducted on urban roads, where the traffic situation (and hence the 'patterns' of arriving pulses) is far more complicated. The performance of the imputation procedure under the new conditions has not yet been completely evaluated and is therefore distrusted. In particular, the imputed speeds are believed to be undependable.

The *Strategy 1* estimators, now to be presented, put some trust in the *number* of imputed vehicles, but none in the imputed speeds.

#### Estimator of flow

As estimator of the flow in site $k$, $y_k$, we propose

$$\hat{y}_k^{(1)} = \sum_{h=1}^{H_k} (n_{r_{kh}} + n_{I_{kh}}) = \sum_{h=1}^{H_k} \hat{y}_{kh}^{(1)} \tag{4.1}$$

where $n_{I_{kh}}$ is the number of imputed vehicles in homogeneity group $U_{kh}$, and $n_{I_k} = \sum_{h=1}^{H_k} n_{I_{kh}}$.

The estimator $\hat{y}_k^{(1)}$ is a function of the $n_{r_{kh}}$'s, whose stochastic properties are regulated by model $\mathrm{m}_r$, and of the $n_{I_{kh}}$'s, which in principle are fix entities. To simplify, we will treat the latter also as random variables. A random model for $n_{I_{kh}}$ is stated in Section 4.3.2.

**Estimator of travel time**

As estimator of the travel time in site $k$, $z_k$, we suggest using

$$\hat{z}_k^{(1)} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{\hat{\theta}_{kh}^{(1)}} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{n_{r_{kh}}/\hat{y}_{kh}^{(1)}} = \sum_{h=1}^{H_k} (n_{r_{kh}} + n_{I_{kh}})\, \bar{x}_{r_{kh}} \tag{4.2}$$

where $\bar{x}_{r_{kh}} = \sum_{r_{kh}} x_v / n_{r_{kh}}$. In words, the registered travel times are simply weighted by the corresponding inverse estimated registration probabilities.

If we had a choice, we would estimate $\theta_{kh}$ by the true registration rate $n_{r_{kh}}/y_{kh}$ instead of $\hat{\theta}_{kh}^{(1)}$. Then, the estimator $\hat{z}_k^{(1)}$ would be the census version (the special case when the ambition is to observe all members of the population, and thus missing data is the sole source of randomness) of the *direct weighting estimator* ( [28, Equation (4.10)], [29, Equation (15.6.8)] ) of $z_k$. Conditional on $s_{iq}$, and provided that the probability of an empty homogeneity group is negligible, $\hat{z}_k^{(1)}$ would then be unbiased for $z_k$ under model $\mathrm{m}_r$.

However, we do not know the denominator $y_{kh}$ of the registration rate, but use $\hat{y}_{kh}^{(1)}$. Since the $\hat{y}_{kh}^{(1)}$'s are random, the statistical properties of $\hat{z}_k^{(1)}$ remain to be investigated.

**A model of the imputation mechanism**

The traffic analyzer's procedure for creating imputed vehicles is not easy to penetrate or describe. A flow chart facilitates the understanding, but the procedure is still hard to handle formally. We therefore enter upon an easier (simplified) course and treat the number of imputed vehicles, $n_{I_{kh}}$, as random.

In [29, Section 16.3], a *simple measurement model* is formulated, in which measurements on elements of a sample are modeled as random variables. An observed value is viewed as composed of the true value and a random measurement error. The model is 'simple' since the model moments do not depend on the realized sample. Our imputation model, denoted $\mathrm{m}_i$, is formulated in the same spirit as the simple measurement model. The observations considered are the imputed numbers $n_{I_{kh}}$. An $n_{I_{kh}}$ is viewed as composed of the true number of unregistered vehicles, $y_{kh} - n_{r_{kh}}$, and a random error $\varepsilon_{kh}$.

The model moments are assumed to be independent of the sample. The moments are, however, allowed to depend on the number of registered vehicles, $n_{r_{kh}}$. This makes sense since the imputed vehicles are created from surplus pulses.

**The imputation model, $\mathbf{m_i}$**
Given $s_{iq}$ and $\mathbf{n}_{r_k}$,

- the number $n_{I_{kh}}$ of imputed vehicles in homogeneity group $U_{kh}$ ($h = 1, ..., H_k, k \in s_{iq}$), has the mean $\mu_{(I|r)_{kh}} = E_{\mathrm{m}_i}(n_{I_{kh}} | s_{iq}, n_{r_{kh}})$ and variance $\sigma^2_{(I|r)_{kh}} = V_{\mathrm{m}_i}(n_{I_{kh}} | s_{iq}, n_{r_{kh}})$,
- the $n_{I_{kh}}$'s are independent, and
- the model moments $\mu_{(I|r)_{kh}}$ and $\sigma^2_{(I|r)_{kh}}$ are independent of $s_{iq}$.

The conditional expectation and variance of $n_{I_{kh}}$ given $s_{iq}$, with respect jointly to model $\mathrm{m}_r$ and $m$, are, respectively,

$$\mu_{I_{kh}} = E_{\mathrm{m}_r\mathrm{m}_i}(n_{I_{kh}} | s_{iq}) = E_{\mathrm{m}_r} E_{\mathrm{m}_i}(n_{I_{kh}} | s_{iq}, n_{r_{kh}})$$
$$\sigma^2_{I_{kh}} = V_{\mathrm{m}_r\mathrm{m}_i}(n_{I_{kh}} | s_{iq}) = E_{\mathrm{m}_r} V_{\mathrm{m}_i}(n_{I_{kh}} | s_{iq}, n_{r_{kh}}) + V_{\mathrm{m}_r} E_{\mathrm{m}_i}(n_{I_{kh}} | s_{iq}, n_{r_{kh}}).$$

In its present form, the imputation model is quite vague: it does not say how $n_{I_{kh}}$ is connected with $y_{kh} - n_{r_{kh}}$ and $\varepsilon_{kh}$. The model is further specified in Section 4.4.2.

### 4.3.3 Strategy 2

Our second proposal for missing data adjustments, *Strategy 2,* rests on the use of the auxiliary variable ME for estimating registration probabilities.

**Estimator of flow**

If we do not use the imputed vehicles, we have few options left for adjusting the flow for missing data. One remaining possibility, however, is to weight the numbers of registered vehicles in a suitable manner. The (estimated) registration rates used in Equation (4.2) are no longer an option, but other estimates of the registration probabilities are needed.

The possibility of estimating (response) probabilities from auxiliary data is quite sparsely discussed in the literature. The idea is put forward in [5, Section 9]; other references include [8], [9] and [6, Section 3.5]. In [10], response probabilities are modeled by logistic regression and estimated from the fitted model. Nonparametric estimation methods are discussed for instance in [14].

We do not want to introduce model parameters into our adjusted estimator (we do not know how to estimate them from sample data), and therefore choose a very simple approach: we try to find an auxiliary variable with roughly a one-to-one relationship with the unknown registration probability. Within our limited supply of variables, the ME is the one we hope fits the description best. Thus, our second proposal for estimator of the flow in site $k$ relies on the use of $(ME)_{kh}$, the ME for homogeneity group $U_{kh}$, as estimator of $\theta_{kh}$:

$$\hat{y}_k^{(2)} = \sum_{h=1}^{H_k} \frac{n_{r_{kh}}}{\hat{\theta}_{kh}^{(2)}} = \sum_{h=1}^{H_k} \frac{n_{r_{kh}}}{(ME)_{kh}} = \sum_{h=1}^{H_k} \hat{y}_{kh}^{(2)}. \tag{4.3}$$

In order to evaluate the statistical properties of $\hat{y}_k^{(2)}$, we need to specify the relationship between $\theta_{kh}$ and $\hat{\theta}_{kh}^{(2)}$. A model for this relationship is stated in Section 4.3.3.

### Estimator of travel time

As estimator of the travel time in site $k$, $z_k$, we suggest using

$$\hat{z}_k^{(2)} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{\hat{\theta}_{kh}^{(2)}} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{(ME)_{kh}}. \tag{4.4}$$

The estimator $\hat{z}_k^{(2)}$ is constructed according to the same principles as $\hat{z}_k^{(1)}$ in Equation (4.2), only with $\theta_{kh}$ estimated by $\hat{\theta}_{kh}^{(2)}$ instead of $\hat{\theta}_{kh}^{(1)}$.

### An error model for $\hat{\theta}^{(2)}$

Our error model for $\hat{\theta}_{kh}^{(2)} = (ME)_{kh}$ as estimator of $\theta_{kh}$ has very much in common with the imputation model in Section 4.3.2 (and thus also with the

simple measurement model in [29, Section 16.3]). Again, an observed value is viewed as composed of the true value and a random measurement error, and the model is 'simple.' The observations considered here are the measurement efficiencies $(ME)_{kh}$. *In the role as estimator of* $\theta_{kh}$, the $(ME)_{kh}$ is viewed as random; or, more precisely, as composed of the true registration probability, $\theta_{kh}$, and a random error $\epsilon_{kh}$. The model moments are assumed to be independent of the sample *and* of the number of registered vehicles, $n_{r_{kh}}$.

**The error model for $\hat{\theta}_{kh}^{(2)}$, $\mathbf{m_t}$**

- The estimator $\hat{\theta}_{kh}^{(2)} = (ME)_{kh}$ of $\theta_{kh}$ $(h = 1, ..., H_k, k \in s_{iq})$, has the mean $\mu_{\hat{\theta}_{kh}^{(2)}}$ and variance $\sigma^2_{\hat{\theta}_{kh}^{(2)}}$,
- the $\hat{\theta}_{kh}^{(2)}$'s are independent, and
- the model moments $\mu_{\hat{\theta}_{kh}^{(2)}}$ and $\sigma^2_{\hat{\theta}_{kh}^{(2)}}$ are independent of $s_{iq}$ and $n_{r_{kh}}$.

The error model does not specify how $\hat{\theta}_{kh}^{(2)}$ is connected with $\theta_{kh}$ and $\epsilon_{kh}$. Two possible relationships, the additive and the multiplicative, are considered in Section 4.4.2.

## 4.4 Estimation with missing data

A more realistic situation than the one dealt with in Section 2.6 is that some observational data are missing. Then, the true $a_k$'s are unknown.

### 4.4.1 The estimators $\hat{t}_{\hat{a}^{(c)}}$ and $\hat{R}^{(c)}$

Let $\hat{a}_k^{(c)}$, $k \in s_{iq}$, be the estimator of $a_k$ under Strategy $c$ $(c = 0, 1, 2)$. The joint probability distribution (conditional on $s_{iq}$) of the random variables $\hat{a}_k^{(c)}$ is called model $\mathrm{m}_2$. The estimator of $t_a$ obtained by replacing $a$ by $\hat{a}^{(c)}$ in Equation (2.1) is denoted $\hat{t}_{\hat{a}^{(c)}}$; the corresponding estimator of $R$ is $\hat{R}^{(c)} = \hat{t}_{\hat{y}^{(c)}}/\hat{t}_{\hat{z}^{(c)}}$.

Let expectations and variances be indicated by subscript $\mathrm{m}_2$ if taken with respect to model $\mathrm{m}_2$; by subscript $p\mathrm{m}_2$ if taken with respect jointly to the sampling design $p$ and model $\mathrm{m}_2$. In order to shorten the formulae, we denote $E_{\mathrm{m}_2}\left(\hat{a}_k^{(c)} \,|s_{iq}\right)$ and $V_{\mathrm{m}_2}\left(\hat{a}_k^{(c)} \,|s_{iq}\right)$ by $\gamma\big(\hat{a}^{(c)}\big)_k$ and $\delta\big(\hat{a}^{(c)}\big)_k$, respectively. The population entities $t_{\gamma\left(\hat{a}^{(c)}\right)iq}$, $t_{\gamma\left(\hat{a}^{(c)}\right)i}$, $t_{\gamma\left(\hat{a}^{(c)}\right)}$, $S^2_{t_{\gamma\left(\hat{a}^{(c)}\right)}U_i}$, $S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{iq}}$, $V_{\gamma\left(\hat{a}^{(c)}\right)i}$, $V_{\gamma\left(\hat{a}^{(c)}\right)IIi}$ and $V_{\gamma\left(\hat{a}^{(c)}\right)iq}$ for $\gamma\big(\hat{a}^{(c)}\big)$ are defined in the same manner as the corresponding entities for $a$ in Section 2.6.

The statistical properties of $\hat{t}_{\hat{a}^{(c)}}$ are investigated in the following theorem:

**Theorem 4.4.1** *Jointly under the sampling design $p$ in Section 2.4 and model $m_2$, the expected value of $\hat{t}_{\hat{a}^{(c)}}$ is given by*

$$E_{pm_2}\big(\hat{t}_{\hat{a}^{(c)}}\big) = \sum_{i=1}^{N_I} E_{pm_2}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big) = t_{\gamma\left(\hat{a}^{(c)}\right)} \tag{4.5}$$

*where $E_{pm_2}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big) = t_{\gamma\left(\hat{a}^{(c)}\right)i}$. The variance of $\hat{t}_{\hat{a}^{(c)}}$ is given by*

$$V_{pm_2}\big(\hat{t}_{\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{t_{\gamma\left(\hat{a}^{(c)}\right)i}}{p_i} - t_{\gamma\left(\hat{a}^{(c)}\right)}\right)^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{pm_2}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big)}{p_i} \tag{4.6}$$

*where*

$$V_{pm_2}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big) = V_{\gamma\left(\hat{a}^{(c)}\right)i} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\,.$$

The proof of Theorem 4.4.1 is given in Appendix B.1.3.

From Theorem 4.4.1, the bias of $\hat{t}_{\hat{a}^{(c)}}$ as estimator of $t_a$ is

$$E_{p\mathrm{m}_2}\big(\hat{t}_{\hat{a}^{(c)}}\big) - t_a = \sum\nolimits_{i=1}^{N_I}\sum\nolimits_{U_{IIi}}\sum\nolimits_{U_{iq}}\left(\gamma\big(\hat{a}^{(c)}\big)_k - a_k\right). \tag{4.7}$$

In general, the sign of the bias is unknown. This is also true of the sign of the variance change due to the use of $\hat{t}_{\hat{a}^{(c)}}$ instead of $\hat{t}_a$, $V_{p\xi}\big(\hat{t}_{\hat{a}^{(c)}}\big) - V_p\big(\hat{t}_{\pi a}\big)$.

If the estimators $\hat{a}_k^{(c)}$ are unbiased for $a_k$, the following corollary applies:

**Corollary 4.4.1** *Assume that $\gamma\big(\hat{a}^{(c)}\big)_k$ equals $a_k$ $(k \in s)$. Then, the estimator $\hat{t}_{\hat{a}^{(c)}}$ is unbiased for $t_a$. The use of $\hat{t}_{\hat{a}^{(c)}}$ instead of $\hat{t}_a$ as estimator of $t_a$ increases the variance by*

$$V_{pm_2}\big(\hat{t}_{\hat{a}^{(c)}}\big) - V_p\big(\hat{t}_a\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\,. \tag{4.8}$$

The statistical properties of $\hat{R}^{(c)}$ are investigated in Theorem 4.4.2:

**Theorem 4.4.2** *Jointly under the sampling design $p$ in Section 2.4 and model $m_2$, the estimator $\hat{R}^{(c)}$ is approximately unbiased for*

$$R^{(c)} = \frac{E_{pm_2}\left(\hat{t}_{\hat{y}^{(c)}}\right)}{E_{pm_2}\left(\hat{t}_{\hat{z}^{(c)}}\right)} = \frac{t_{\gamma\left(\hat{y}^{(c)}\right)}}{t_{\gamma\left(\hat{z}^{(c)}\right)}}. \tag{4.9}$$

*The approximate variance of $\hat{R}^{(c)}$ is given by*

$$AV_{pm_2}\left(\hat{R}^{(c)}\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{t^2_{\gamma\left(\hat{E}^{(c)}\right)i}}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{pm_2}\left(\hat{t}_{\pi\hat{E}^{(c)}i}\right)}{p_i}\right\} \tag{4.10}$$

*where $t_{\gamma\left(\hat{E}^{(c)}\right)i}$ and $V_{pm_2}\left(\hat{t}_{\pi\hat{E}^{(c)}i}\right)$ correspond to $t_{\gamma\left(\hat{a}^{(c)}\right)i}$ and $V_{pm_2}\left(\hat{t}_{\pi\hat{a}^{(c)}i}\right)$, respectively; $\gamma$ and $\delta$ are however functions of $\hat{E}^{(c)} = \hat{y}^{(c)} - R^{(c)}\hat{z}^{(c)}$ instead of $\hat{a}^{(c)}$.*

The proof of Theorem 4.4.2 follows by a slight generalization of the results in [27, Section 6.8.2.].

From Theorem 4.4.2, the sign of the bias of $\hat{R}^{(c)}$ as estimator of $R$, as well as the sign of the variance change due to using $\hat{R}^{(c)}$ instead of the prototype estimator $\hat{R}$, is in general unknown.

The following corollary applies if $\hat{y}_k^{(c)}$ and $\hat{z}_k^{(c)}$ are unbiased for $y_k$ and $z_k$, respectively:

**Corollary 4.4.2** *Assume that $\gamma\left(\hat{y}^{(c)}\right)_k = y_k$ and $\gamma\left(\hat{z}^{(c)}\right)_k = z_k$ $(k \in s)$. Then, the estimator $\hat{R}^{(c)}$ is approximately unbiased for $R$. The approximate variance increase due to the use of $\hat{R}^{(c)}$ instead of $\hat{R}$ as estimator of $R$ is given by*

$$\begin{aligned} &AV_{pm_2}\left(\hat{R}^{(c)}\right) - AV_p\left(\hat{R}\right) \\ &= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{E}^{(c)}\right)_k \end{aligned} \tag{4.11}$$

*where $\hat{E}^{(c)} = \hat{y}^{(c)} - R\hat{z}^{(c)}$.*

### 4.4.2 Results for present and proposed estimators

In Section 4.4.1, the general statistical properties of $\hat{t}_{\hat{a}^{(c)}}$ and $\hat{R}^{(c)}$ were derived. Here, specific results for the estimators under Strategies 0, 1, and 2 are derived. This implies presenting the explicit $\gamma$ and $\delta$ expressions.

When dealing with the Strategy 1 and 2 estimators, only the special case with a single homogeneity group is considered. Subscript $h$ is then no longer needed. The sole reason for this demarcation is to keep the notation simple; expansion of the results to the case $H_k > 1$ is straightforward.

**Strategy** 0

Model $\mathrm{m}_2$ is here interpreted as the registration model $\mathrm{m}_r$. When applying Theorem 4.4.1 on the estimators $\hat{t}_{\hat{y}^{(0)}}$ and $\hat{t}_{\hat{z}^{(0)}}$, we use

$$\left(a_k, \hat{a}_k^{(0)}\right) = \left(y_k, \hat{y}_k^{(0)}\right) = (y_k, n_{r_k})$$

for $\hat{t}_{\hat{y}^{(0)}}$, and

$$\left(a_k, \hat{a}_k^{(0)}\right) = \left(z_k, \hat{z}_k^{(0)}\right) = (z_k, n_{r_k}\bar{x}_{r_k})$$

for $\hat{t}_{\hat{z}^{(0)}}$.

For $\hat{t}_{\hat{y}^{(0)}}$, the model moments are simply $\gamma\big(\hat{y}^{(0)}\big)_k = \mu_{r_k}$ and $\delta\big(\hat{y}^{(0)}\big)_k = \sigma^2_{r_k}$, whereas for $\hat{t}_{\hat{z}^{(0)}}$, they are

$$\gamma\big(\hat{z}^{(0)}\big)_k = \frac{z_k}{y_k}\mu_{r_k} \tag{4.12}$$

$$\delta\big(\hat{z}^{(0)}\big)_k = E_{\mathrm{m}_r}\left(n^2_{r_k}\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}\,|s_{iq}\right) + \left(\frac{z_k}{y_k}\right)^2\sigma^2_{r_k} \tag{4.13}$$

where

$$S^2_{xU_k} = \frac{1}{y_k - 1}\sum\nolimits_{U_k}\left(x_v - \frac{z_k}{y_k}\right)^2.$$

The first term on the right-hand side of Equation (4.13) simplifies to

$$E_{\mathrm{m}_r}\left(n^2_{r_k}\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}\,|s_{iq}\right) = \left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right]S^2_{xU_k}. \tag{4.14}$$

Equations (4.12) and (4.13) are derived by use of Proposition C.1 with $(A, B) = (n_{r_k}, \bar{x}_{r_k})$. We also use the fact that

$$E_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k}, s_{iq}) = \frac{z_k}{y_k} \tag{4.15}$$

$$V_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k}, s_{iq}) = \frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k} \tag{4.16}$$

which follows from implication number 2 of the registration model – see Section 4.3.1. (The moments in Equations (4.15) and (4.16) are in fact conditional also on the event that $n_{r_k} \geq 1$. For details, see [29, Section 7.10.1].)

When applying Theorem 4.4.2 on the estimator $\hat{R}^{(0)}$, we use

$$\hat{E}^{(0)}_k = n_{r_k} - R^{(0)} n_{r_k} \bar{x}_{r_k} = n_{r_k} \left(1 - R^{(0)} \bar{x}_{r_k}\right)$$

where $R^{(0)} = \sum_U \mu_{r_k} / \sum_U (z_k/y_k)\, \mu_{r_k}$. The model moments are

$$\gamma\Big(\hat{E}^{(0)}\Big)_k = \left(1 - R^{(0)} \frac{z_k}{y_k}\right) \mu_{r_k} \tag{4.17}$$

$$\begin{aligned} \delta\Big(\hat{E}^{(0)}\Big)_k &= \left(R^{(0)}\right)^2 \left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right] S^2_{xU_k} \\ &\quad + \left(1 - R^{(0)} \frac{z_k}{y_k}\right)^2 \sigma^2_{r_k}. \end{aligned} \tag{4.18}$$

Equations (4.17) and (4.18) are derived by use of Proposition C. 1 with $(A, B) = \left(n_{r_k}, 1 - R^{(0)} \bar{x}_{r_k}\right)$ and by applying Equation (4.14).

We now make an attempt to simplify the results for the Strategy 0 estimators. From implication number 1 of the registration model in Section 4.3.1, $\mu_{r_k} = y_k \theta_k$ and $\sigma^2_{r_k} = y_k \theta_k (1 - \theta_k)$. It follows that for $\hat{t}_{\hat{y}^{(0)}}$, the model moments are $\gamma\big(\hat{y}^{(0)}\big)_k = y_k \theta_k$ and $\delta\big(\hat{y}^{(0)}\big)_k = y_k \theta_k (1 - \theta_k)$; for $\hat{t}_{\hat{z}^{(0)}}$, they are

$$\gamma\big(\hat{z}^{(0)}\big)_k = z_k \theta_k \tag{4.19}$$

$$\begin{aligned} \delta\big(\hat{z}^{(0)}\big)_k &= \theta_k (1 - \theta_k) \left[\frac{z_k^2}{y_k} + (y_k - 1)\, S^2_{xU_k}\right] \\ &= \theta_k (1 - \theta_k) \sum\nolimits_{U_k} x_v^2 \end{aligned} \tag{4.20}$$

and for $\hat{R}^{(0)}$, they are

$$\gamma\left(\hat{E}^{(0)}\right)_k = \theta_k \left(y_k - R^{(0)} z_k\right) \tag{4.21}$$

$$\begin{aligned}\delta\left(\hat{E}^{(0)}\right)_k &= \theta_k \left(1-\theta_k\right)\left[\left(R^{(0)}\right)^2 \left(y_k - 1\right) S^2_{xU_k} + \left(1 - R^{(0)}\frac{z_k}{y_k}\right)^2 y_k\right] \\ &= \theta_k \left(1-\theta_k\right)\left[\left(R^{(0)}\right)^2 \sum\nolimits_{U_k} x_v^2 - 2R^{(0)} z_k + y_k\right]\end{aligned} \tag{4.22}$$

where $R^{(0)} = \sum_U y_k\theta_k / \sum_U z_k\theta_k$.

Assume that the registration probabilities $\theta_k = \theta$ for all $k \in U$. Then, $R^{(0)}$ coincides with $R$, and $\gamma\left(\hat{E}^{(0)}\right)_k = \theta E_k$. It follows that

$$t^2_{\gamma\left(\hat{E}^{(0)}\right)i} = \theta^2 t^2_{Ei} \tag{4.23}$$

$$\begin{aligned}V_{\mathrm{pm}_r}\left(\hat{t}_{\pi\hat{E}^{(0)}i}\right) &= N^2_{IIi}\frac{1-f_{IIi}}{n_{IIi}}\theta^2 S^2_{t_E U_i} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}} N^2_{iq}\frac{1-f_{iq}}{n_{iq}}\theta^2 S^2_{EU_{iq}} \\ &\quad + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{E}^{(0)}\right)_k \\ &= \theta^2 V_p\left(\hat{t}_{\pi Ei}\right) + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{E}^{(0)}\right)_k .\end{aligned} \tag{4.24}$$

By insertion of Equations (4.23) and (4.24) into Equation (4.10), and comparison of the resulting variance expression with the one in Equation (2.9), we see that

$$\begin{aligned}AV_{\mathrm{pm}_r}\left(\hat{R}^{(0)}\right) &= \theta^2 AV_p\left(\hat{R}\right) \\ &\quad + \theta\left(1-\theta\right)\frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}} \\ &\quad \times \sum\nolimits_{U_{iq}}\left(R^2\sum\nolimits_{U_k} x_v^2 - 2Rz_k + y_k\right).\end{aligned} \tag{4.25}$$

**Strategy** 1

Expectations and variances with respect to model $\mathrm{m}_2$ are here interpreted as taken with respect jointly to model $\mathrm{m}_r$ and $\mathrm{m}_i$, and are indicated by subscript $\mathrm{m}_r\mathrm{m}_i$. When applying Theorem 4.4.1 on $\hat{t}_{\hat{y}^{(1)}}$ and $\hat{t}_{\hat{z}^{(1)}}$, we use

$$\left(a_k, \hat{a}_k^{(1)}\right) = \left(y_k, \hat{y}_k^{(1)}\right) = \left(y_k, n_{r_k} + n_{I_k}\right)$$

for $\hat{t}_{\hat{y}^{(1)}}$, and

$$\left(a_k, \hat{a}_k^{(1)}\right) = \left(z_k, \hat{z}_k^{(1)}\right) = \left(z_k, \left(n_{r_k} + n_{I_k}\right) \bar{x}_{r_k}\right)$$

for $\hat{t}_{\hat{z}^{(1)}}$. When applying Theorem 4.4.2 on $\hat{R}^{(1)}$, we use

$$\hat{E}_k^{(1)} = \left(n_{r_k} + n_{I_k}\right) - R^{(1)} \left(n_{r_k} + n_{I_k}\right) \bar{x}_{r_k} = \left(n_{r_k} + n_{I_k}\right) \left(1 - R^{(1)} \bar{x}_{r_k}\right).$$

The $\gamma$ and $\delta$ expressions will first be presented for the general imputation model $m$ in Section 4.3.2 ('general' in the sense that it does not say how $n_{I_k}$ is connected with $y_k - n_{r_k}$ and $\varepsilon_k$); then for a more specified model.

**Under general imputation model assumptions** The model moments for $\hat{t}_{\hat{y}^{(1)}}$ are

$$\gamma\left(\hat{y}^{(1)}\right)_k = \mu_{r_k} + \mu_{I_k} \tag{4.26}$$

$$\delta\left(\hat{y}^{(1)}\right)_k = \sigma_{r_k}^2 + \sigma_{I_k}^2 + 2Cov_{\mathrm{m}_r}\left(n_{r_k}, \mu_{(I|r)_k} | s_{iq}\right) \tag{4.27}$$

where $Cov_{\mathrm{m}_r}\left(n_{r_k}, \mu_{(I|r)_k} | s_{iq}\right)$ is the conditional covariance of $n_{r_k}$ and $\mu_{(I|r)_k}$, given $s_{iq}$, with respect to model $\mathrm{m}_r$. For $\hat{t}_{\hat{z}^{(1)}}$, the moments are

$$\gamma\left(\hat{z}^{(1)}\right)_k = \frac{z_k}{y_k} \gamma\left(\hat{y}^{(1)}\right)_k \tag{4.28}$$

and

$$\begin{aligned}
\delta\left(\hat{z}^{(1)}\right)_k &= E_{\mathrm{m}_r \mathrm{m}_i}\left[\left(n_{r_k} + n_{I_k}\right)^2 \frac{1 - n_{r_k}/y_k}{n_{r_k}} S_{xU_k}^2 | s_{iq}\right] \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(1)}\right)_k \\
&= E_{\mathrm{m}_r}\left\{\left[\left(n_{r_k} + \mu_{(I|r)_k}\right)^2 + \sigma_{(I|r)_k}^2\right] \frac{1 - n_{r_k}/y_k}{n_{r_k}} S_{xU_k}^2 | s_{iq}\right\} \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(1)}\right)_k .
\end{aligned} \tag{4.29}$$

Equations (4.28) and (4.29) are derived by use of Proposition C.1 with $(A, B) = (n_{r_k} + n_{I_k}, \bar{x}_{r_k})$, and the equalities

$$\begin{aligned}
E_{\mathrm{m}_r\mathrm{m}_i}(\bar{x}_{r_k} \,|n_{r_k} + n_{I_k}, s_{iq}) &= E_{\mathrm{m}_i} E_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k} + n_{I_k}, n_{I_k}, s_{iq}) \\
&= E_m E_r(\bar{x}_{r_k} \,|n_{r_k}, s_{iq}) = E_r(\bar{x}_{r_k} \,|n_{r_k}, s_{iq}) \\
V_{\mathrm{m}_r\mathrm{m}_i}(\bar{x}_{r_k} \,|n_{r_k} + n_{I_k}, s_{iq}) &= E_{\mathrm{m}_i} V_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k} + n_{I_k}, n_{I_k}, s_{iq}) \\
&\quad + V_{\mathrm{m}_i} E_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k} + n_{I_k}, n_{I_k}, s_{iq}) \\
&= E_{\mathrm{m}_i} V_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k}, s_{iq}) + V_{\mathrm{m}_i} E_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k}, s_{iq}) \\
&= V_{\mathrm{m}_r}(\bar{x}_{r_k} \,|n_{r_k}, s_{iq})
\end{aligned}$$

which hold since $\bar{x}_{r_k}$ and $n_{I_k}$ are independent.

Finally, the model moments for $\hat{R}^{(1)}$ are

$$\gamma\Big(\hat{E}^{(1)}\Big)_k = \left(1 - R^{(1)}\frac{z_k}{y_k}\right)\gamma\big(\hat{y}^{(1)}\big)_k \tag{4.30}$$

$$\begin{aligned}
\delta\Big(\hat{E}^{(1)}\Big)_k &= \big(R^{(1)}\big)^2 E_{\mathrm{m}_r\mathrm{m}_i}\left[\left(n_{r_k} + n_{I_k}\right)^2 \frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k} \,|s_{iq}\right] \\
&\quad + \left(1 - R^{(1)}\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(1)}\big)_k \\
&= \big(R^{(1)}\big)^2 E_{\mathrm{m}_r}\left\{\left[\left(n_{r_k} + \mu_{(I|r)_k}\right)^2 + \sigma^2_{(I|r)_k}\right]\frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k} \,|s_{iq}\right\} \\
&\quad + \left(1 - R^{(1)}\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(1)}\big)_k .
\end{aligned} \tag{4.31}$$

Equations (4.30) and (4.31) are obtained by use of Proposition C.1 with $(A, B) = \big(n_{r_k} + n_{I_k}, 1 - R^{(1)}\bar{x}_{r_k}\big)$.

Consider the desirable case where, given $n_{r_k}$, the number of imputed vehicles 'on the average' equals the number of unregistered vehicles (that is, $\mu_{(I|r)_k} = y_k - n_{r_k}$.) Then, $\mu_{I_k} = y_k - \mu_{r_k}$, and consequently, $\gamma\big(\hat{y}^{(1)}\big)_k = y_k$ and $\gamma\big(\hat{z}^{(1)}\big)_k = z_k$. From Corollary 4.4.1, $\hat{t}_{\hat{y}^{(1)}}$ and $\hat{t}_{\hat{z}^{(1)}}$ are unbiased for $t_y$ and $t_z$, respectively, and from Corollary 4.4.2, $\hat{R}^{(1)}$ is approximately unbiased for $R$.

**Under a multiplicative imputation error model** The size of the error associated with $n_{I_k}$ is likely to depend on the number of unregistered vehicles. The more vehicles that are not registered, the more complicated the imputation task apparently is, and the higher the risk of large errors arising. For

this reason, let us assume that the number of imputed vehicles $n_{I_k}$ consists of the number of unregistered vehicle *times* a random error:

$$n_{I_k} = (y_k - n_{r_k})\,\varepsilon_k. \tag{4.32}$$

Let the conditional mean and variance of $\varepsilon_k$ given $s_{iq}$ and $n_{r_k}$ be denoted $\mu_\varepsilon = E_{\mathrm{m}_i}(\varepsilon_k\,|s_{iq}, n_{r_k})$ and $\sigma^2_\varepsilon = V_{\mathrm{m}_i}(\varepsilon_k\,|s_{iq}, n_{r_k})$, respectively. As the notation suggests, the conditional moments $\mu_\varepsilon$ and $\sigma^2_\varepsilon$ are assumed to depend neither on $s_{iq}$ or $n_{r_k}$ nor on the road site $k$. This makes sense since the same imputation software is used throughout the survey.

Under the multiplicative error model,

$$\mu_{(I|r)_k} = (y_k - n_{r_k})\,\mu_\varepsilon \tag{4.33}$$

$$\sigma^2_{(I|r)_k} = (y_k - n_{r_k})^2\,\sigma^2_\varepsilon \tag{4.34}$$

and

$$\mu_{I_k} = E_{\mathrm{m}_r}[(y_k - n_{r_k})\,\mu_\varepsilon\,|s_{iq}] = \big(y_k - \mu_{r_k}\big)\,\mu_\varepsilon \tag{4.35}$$

$$\begin{aligned}\sigma^2_{I_k} &= E_{\mathrm{m}_r}\big[(y_k - n_{r_k})^2\,\sigma^2_\varepsilon\,|s_{iq}\big] + V_{\mathrm{m}_r}[(y_k - n_{r_k})\,\mu_\varepsilon\,|s_{iq}] \\ &= \Big[\big(y_k - \mu_{r_k}\big)^2 + \sigma^2_{r_k}\Big]\,\sigma^2_\varepsilon + \sigma^2_{r_k}\mu^2_\varepsilon.\end{aligned} \tag{4.36}$$

We now modify the $\gamma$ and $\delta$ expressions presented earlier (Equations (4.26) to (4.31)) in compliance with Equations (4.33) to (4.36). The resulting model moments for $\hat{t}_{\hat{y}^{(1)}}$ are

$$\gamma\big(\hat{y}^{(1)}\big)_k = \mu_{r_k}\,(1 - \mu_\varepsilon) + y_k\mu_\varepsilon \tag{4.37}$$

$$\delta\big(\hat{y}^{(1)}\big)_k = \sigma^2_{r_k}\,\big[(1 - \mu_\varepsilon)^2 + \sigma^2_\varepsilon\big] + \big(y_k - \mu_{r_k}\big)^2\,\sigma^2_\varepsilon. \tag{4.38}$$

In the derivation of Equation (4.38), we use the fact that

$$Cov_{\mathrm{m}_r}[n_{r_k}, (y_k - n_{r_k})\,\mu_\varepsilon\,|s_{iq}] = -\mu_\varepsilon V_{\mathrm{m}_r}(n_{r_k}\,|s_{iq}) = -\mu_\varepsilon\sigma^2_{r_k}.$$

For $\hat{t}_{\hat{z}^{(1)}}$, the moments are

$$\gamma\big(\hat{z}^{(1)}\big)_k = \frac{z_k}{y_k}\gamma\big(\hat{y}^{(1)}\big)_k \tag{4.39}$$

$$\begin{aligned}\delta\big(\hat{z}^{(1)}\big)_k = E_{\mathrm{m}_r}\Big\{&\big[(n_{r_k}\,(1 - \mu_\varepsilon) + y_k\mu_\varepsilon)^2 + (y_k - n_{r_k})^2\,\sigma^2_\varepsilon\big] \\ &\times \frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k}\,|s_{iq}\Big\} + \left(\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(1)}\big)_k\end{aligned} \tag{4.40}$$

and for $\hat{R}^{(1)}$,

$$\gamma\Big(\hat{E}^{(1)}\Big)_k = \left(1 - R^{(1)}\frac{z_k}{y_k}\right)\gamma\big(\hat{y}^{(1)}\big)_k \tag{4.41}$$

$$\begin{aligned}\delta\Big(\hat{E}^{(1)}\Big)_k = {} & \big(R^{(1)}\big)^2 E_{\mathrm{m}_r}\Big\{\big[\left(n_{r_k}\left(1-\mu_\varepsilon\right)+y_k\mu_\varepsilon\right)^2+\left(y_k-n_{r_k}\right)^2\sigma_\varepsilon^2\big] \\ & \times\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\Big\}+\left(1-R^{(1)}\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k\end{aligned} \tag{4.42}$$

with

$$R^{(1)} = \frac{\sum_U\left[\mu_{r_k}\left(1-\mu_\varepsilon\right)+y_k\mu_\varepsilon\right]}{\sum_U\left\{\frac{z_k}{y_k}\left[\mu_{r_k}\left(1-\mu_\varepsilon\right)+y_k\mu_\varepsilon\right]\right\}} = \frac{\left(1-\mu_\varepsilon\right)\sum_U\mu_{r_k}+\mu_\varepsilon t_y}{\left(1-\mu_\varepsilon\right)\sum_U\frac{z_k}{y_k}\mu_{r_k}+\mu_\varepsilon t_z}.$$

Let us now revisit the favorable case of $\mu_{(I|r)_k} = y_k - n_{r_k}$. We have already concluded that in this case, the estimators $\hat{t}_{\hat{y}^{(1)}}$, $\hat{t}_{\hat{z}^{(1)}}$ and $\hat{R}^{(1)}$ are unbiased, or approximately unbiased, for their true counterparts. But how about the variance increases due to not using the complete-data estimators? For the multiplicative imputation error model, this case corresponds to a conditional error mean equal to unity ($\mu_\varepsilon = 1$). The associated $\delta$ expressions for $\hat{t}_{\hat{y}^{(1)}}$ and $\hat{t}_{\hat{z}^{(1)}}$ (to be inserted in Equation (4.8)) are

$$\delta\big(\hat{y}^{(1)}\big)_k = \Big[\big(y_k-\mu_{r_k}\big)^2+\sigma_{r_k}^2\Big]\sigma_\varepsilon^2 \tag{4.43}$$

and

$$\begin{aligned}\delta\big(\hat{z}^{(1)}\big)_k = {} & E_{\mathrm{m}_r}\left\{\left[y_k^2+\left(y_k-n_{r_k}\right)^2\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\} \\ & +\left(\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k\end{aligned} \tag{4.44}$$

respectively. The $\delta$ expression for $\hat{R}^{(1)}$ (to be inserted in Equation (4.11)) is

$$\begin{aligned}\delta\Big(\hat{E}^{(1)}\Big)_k = {} & R^2E_{\mathrm{m}_r}\left\{\left[y_k^2+\left(y_k-n_{r_k}\right)^2\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\} \\ & +\left(1-R\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k.\end{aligned} \tag{4.45}$$

The expectation occurring in Equations (4.44) and (4.45) can be worked out.

Some straightforward algebra gives

$$
\begin{aligned}
&E_{\mathrm{m}_r}\left\{\left[y_k^2+\left(y_k-n_{r_k}\right)^2\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\}\\
&=\left\{y_k\left(1+\sigma_\varepsilon^2\right)\left[y_kE_{\mathrm{m}_r}\left(\frac{1}{n_{r_k}}\,|s_{iq}\right)-1\right]+\sigma_\varepsilon^2\left(3\mu_{r_k}-2y_k-\frac{\sigma_{r_k}^2+\mu_{r_k}^2}{y_k}\right)\right\}S_{xU_k}^2\\
&\approx\left\{y_k\left(1+\sigma_\varepsilon^2\right)\left(\frac{y_k}{\mu_{r_k}}-1\right)+\sigma_\varepsilon^2\left(3\mu_{r_k}-2y_k-\frac{\sigma_{r_k}^2+\mu_{r_k}^2}{y_k}\right)\right\}S_{xU_k}^2 \qquad (4.46)
\end{aligned}
$$

where the approximate equality arises from the (first order) Taylor approximation $E_{\mathrm{m}_r}(1/n_{r_k}\,|s_{iq})\approx 1/E_{\mathrm{m}_r}(n_{r_k}\,|s_{iq})$.

Can Equations (4.43) and (4.46) be additionally simplified? From implication number 1 of the registration model in Section 4.3.1, $\mu_{r_k}=y_k\theta_k$ and $\sigma_{r_k}^2=y_k\theta_k\left(1-\theta_k\right)$. Insertion in Equation (4.43) gives

$$
\delta\big(\hat{y}^{(1)}\big)_k=y_k\left(1-\theta_k\right)\left[\theta_k\left(1-y_k\right)+y_k\right]\sigma_\varepsilon^2 \qquad (4.47)
$$

and in Equation (4.46)

$$
\begin{aligned}
&E_{\mathrm{m}_r}\left\{\left[y_k^2+\left(y_k-n_{r_k}\right)^2\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\}\\
&\approx\left\{y_k\left[\left(\frac{1}{\theta_k}-1\right)+\sigma_\varepsilon^2\left(\frac{1}{\theta_k}-3+3\theta_k-\theta_k^2\right)\right]-\sigma_\varepsilon^2\theta_k\left(1-\theta_k\right)\right\}S_{xU_k}^2. \qquad (4.48)
\end{aligned}
$$

**Strategy** 2

Expectations and variances with respect to model $\mathrm{m}_2$ are here interpreted as taken with respect jointly to model $\mathrm{m}_r$ and $\mathrm{m}_t$, and are indicated by subscript $\mathrm{m}_r\mathrm{m}_t$. When applying Theorem 4.4.1 on the estimators $\hat{t}_{\hat{y}^{(2)}}$ and $\hat{t}_{\hat{z}^{(2)}}$, we use

$$
\left(a_k,\hat{a}_k^{(2)}\right)=\left(y_k,\hat{y}_k^{(2)}\right)=\left(y_k,\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\right)
$$

for $\hat{t}_{\hat{y}^{(2)}}$, and

$$
\left(a_k,\hat{a}_k^{(2)}\right)=\left(z_k,\hat{z}_k^{(2)}\right)=\left(z_k,\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\bar{x}_{r_k}\right)
$$

for $\hat{t}_{\hat{z}^{(2)}}$. Finally, when applying Theorem 4.4.2 on $\hat{R}^{(2)}$, we use

$$\hat{E}_k^{(2)} = \frac{n_{r_k}}{\hat{\theta}_k^{(2)}} - R^{(2)}\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\bar{x}_{r_k} = \frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\left(1 - R^{(2)}\bar{x}_{r_k}\right).$$

The $\gamma$ and $\delta$ expressions will first be presented by use of the general error model $q$ in Section 4.3.3 ('general' in the sense that it does not say how $\hat{\theta}_k^{(2)}$ is connected with $\theta_k$ and $\epsilon_k$), then two special cases will be treated.

**Under general error model assumptions** The estimator $\hat{y}_k^{(2)}$ is theoretically complicated, being a ratio of random variables. By use of Taylor's theorem (see, e.g., [4, Theorem 7.4.1]), we are however able to approximate its moments. The first-order Taylor approximations of the model moments for $\hat{t}_{\hat{y}^{(2)}}$ are given by

$$\gamma\big(\hat{y}^{(2)}\big)_k \approx \frac{\mu_{r_k}}{\mu_{\hat{\theta}_k^{(2)}}} \tag{4.49}$$

$$\delta\big(\hat{y}^{(2)}\big)_k \approx \left(\frac{\mu_{r_k}}{\mu_{\hat{\theta}_k^{(2)}}}\right)^2\left(\frac{\sigma_{r_k}^2}{\mu_{r_k}^2} + \frac{\sigma_{\hat{\theta}_k^{(2)}}^2}{\mu_{\hat{\theta}_k^{(2)}}^2}\right). \tag{4.50}$$

The model moments for $\hat{t}_{\hat{y}^{(2)}}$ are obtained by also using Proposition C.1 with $(A, B) = \left(n_{r_k}/\hat{\theta}_k^{(2)}, \bar{x}_{r_k}\right)$, and the equalities

$$\begin{aligned} E_{\mathrm{m}_r\mathrm{m}_t}\left(\bar{x}_{r_k}\,\middle|n_{r_k}/\hat{\theta}_k^{(2)}, s_{iq}\right) &= E_{\mathrm{m}_t}E_{\mathrm{m}_r}\left(\bar{x}_{r_k}\,\middle|n_{r_k}/\hat{\theta}_k^{(2)}, \hat{\theta}_k^{(2)}, s_{iq}\right) \\ &= E_{\mathrm{m}_t}E_{\mathrm{m}_r}(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) = E_{\mathrm{m}_r}(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) \end{aligned}$$

$$\begin{aligned} V_{\mathrm{m}_r\mathrm{m}_t}\left(\bar{x}_{r_k}\,\middle|n_{r_k}/\hat{\theta}_k^{(2)}, s_{iq}\right) &= E_{\mathrm{m}_t}V_{\mathrm{m}_r}\left(\bar{x}_{r_k}\,\middle|n_{r_k}/\hat{\theta}_k^{(2)}, \hat{\theta}_k^{(2)}, s_{iq}\right) \\ &\quad + V_{\mathrm{m}_t}E_{\mathrm{m}_r}\left(\bar{x}_{r_k}\,\middle|n_{r_k}/\hat{\theta}_k^{(2)}, \hat{\theta}_k^{(2)}, s_{iq}\right) \\ &= E_{\mathrm{m}_t}V_{\mathrm{m}_r}(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) + V_{\mathrm{m}_t}E_{\mathrm{m}_r}(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) \\ &= V_{\mathrm{m}_r}(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) \end{aligned}$$

which hold since $\bar{x}_{r_k}$ and $\hat{\theta}_k^{(2)}$ are independent. The resulting moments are

$$\gamma\big(\hat{z}^{(2)}\big)_k \approx \frac{z_k}{y_k}\frac{\mu_{r_k}}{\mu_{\hat{\theta}_k^{(2)}}} \tag{4.51}$$

$$\begin{aligned}
\delta\left(\hat{z}^{(2)}\right)_k &\approx E_{\mathrm{m}_r\mathrm{m}_t}\left[\left(\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\right)^2 \frac{1-n_{r_k}/y_k}{n_{r_k}} S_{xU_k}^2 \,|s_{iq}\right] \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\
&= E_{\mathrm{m}_t}\left[\frac{1}{\left(\hat{\theta}_k^{(2)}\right)^2}\right]\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu_{r_k}^2 + \sigma_{r_k}^2\right)\right] S_{xU_k}^2 \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\
&\approx \frac{1}{\sigma_{\hat{\theta}_k^{(2)}}^2 + \mu_{\hat{\theta}_k^{(2)}}^2}\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu_{r_k}^2 + \sigma_{r_k}^2\right)\right] S_{xU_k}^2 \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k . \qquad (4.52)
\end{aligned}$$

The second equality for $\delta$ is derived using the independency of $n_{r_k}$ and $\hat{\theta}_k^{(2)}$, and Equation (4.14). The final equality arises from the (first-order) Taylor approximation

$$E_{\mathrm{m}_t}\left[\frac{1}{\left(\hat{\theta}_k^{(2)}\right)^2}\right] \approx \frac{1}{E_{\mathrm{m}_t}\left[\left(\hat{\theta}_k^{(2)}\right)^2\right]}. \qquad (4.53)$$

The model moments for $\hat{R}^{(2)}$ are

$$\gamma\left(\hat{E}^{(2)}\right)_k \approx \left(1 - R^{(2)}\frac{z_k}{y_k}\right)\gamma\left(\hat{y}^{(2)}\right)_k \qquad (4.54)$$

$$\begin{aligned}
\delta\left(\hat{E}^{(2)}\right)_k &\approx \left(R^{(2)}\right)^2 E_{\mathrm{m}_r\mathrm{m}_t}\left[\left(\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\right)^2 \frac{1-n_{r_k}/y_k}{n_{r_k}} S_{xU_k}^2 \,|s_{iq}\right] \\
&\quad + \left(1 - R^{(2)}\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\
&= \left(R^{(2)}\right)^2 E_{\mathrm{m}_t}\left[\frac{1}{\left(\hat{\theta}_k^{(2)}\right)^2}\right]\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu_{r_k}^2 + \sigma_{r_k}^2\right)\right] S_{xU_k}^2 \\
&\quad + \left(1 - R^{(2)}\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k
\end{aligned}$$

$$\approx \frac{\left(R^{(2)}\right)^2}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right] S^2_{xU_k}$$
$$+ \left(1 - R^{(2)}\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k . \tag{4.55}$$

Equations (4.54) and (4.55) are derived by use of Proposition C. 1 with $(A, B) = \left(n_{r_k}/\hat{\theta}_k^{(2)}, 1 - R^{(2)}\bar{x}_{r_k}\right)$, and (for $\delta$) the independency of $n_{r_k}$ and $\hat{\theta}_k^{(2)}$, Equation (4.14), and the approximation in Equation (4.53).

Under the registration model, $\mu_{r_k} = y_k\theta_k$ and $\sigma^2_{r_k} = y_k\theta_k(1-\theta_k)$. It follows that Equations (4.49) and (4.50) simplify to

$$\gamma\left(\hat{y}^{(2)}\right)_k = y_k \frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}} \tag{4.56}$$

$$\delta\left(\hat{y}^{(2)}\right)_k = y_k^2 \left(\frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}}\right)^2 \left(\frac{1-\theta_k}{y_k\theta_k} + \frac{\sigma^2_{\hat{\theta}_k^{(2)}}}{\mu^2_{\hat{\theta}_k^{(2)}}}\right); \tag{4.57}$$

Equations (4.51) and (4.52) to

$$\gamma\left(\hat{z}^{(2)}\right)_k \approx z_k \frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}} \tag{4.58}$$

$$\begin{aligned}\delta\left(\hat{z}^{(2)}\right)_k &\approx \frac{1}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \theta_k(1-\theta_k)(y_k - 1) S^2_{xU_k} \\ &\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\ &= \frac{1}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \theta_k(1-\theta_k)\left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) \\ &\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k ;\end{aligned} \tag{4.59}$$

and Equations (4.54) and (4.55) to

$$\gamma\left(\hat{E}^{(2)}\right)_k \approx y_k \left(1 - R^{(2)}\frac{z_k}{y_k}\right) \frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}} \tag{4.60}$$

$$\delta\left(\hat{E}^{(2)}\right)_k \approx \frac{\left(R^{(2)}\right)^2}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \theta_k \left(1-\theta_k\right) \left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) + \left(1 - R^{(2)} \frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \tag{4.61}$$

where $R^{(2)} = \sum_U y_k \left(\theta_k / \mu_{\hat{\theta}_k^{(2)}}\right) / \sum_U z \left(\theta_k / \mu_{\hat{\theta}_k^{(2)}}\right)$.

Assume that $\hat{\theta}_k^{(2)}$ is an unbiased estimator of the true registration probability ($\mu_{\hat{\theta}_k^{(2)}} = \theta_k$). Then, from Equations (4.56) and (4.58), $\gamma\left(\hat{y}^{(2)}\right)_k = y_k$ and $\gamma\left(\hat{z}^{(2)}\right)_k = z_k$. It follows from Corollary 4.4.1 that $\hat{t}_{\hat{y}^{(2)}}$ and $\hat{t}_{\hat{z}^{(2)}}$ then are unbiased for $t_y$ and $t_z$, respectively. Furthermore, from Corollary 4.4.2, $\hat{R}^{(2)}$ is approximately unbiased for $R$.

**Under some special cases of the error model for $\hat{\theta}_k^{(2)}$** Consider the error model $\mathrm{m}_t$ for $\hat{\theta}_k^{(2)}$ stated in Section 4.3.3. Two possible functional relationships between $\hat{\theta}_k^{(2)}$ and $\theta_k$ are the additive error model,

$$\hat{\theta}_k^{(2)} = \theta_k + \epsilon_k \tag{4.62}$$

and the multiplicative error model,

$$\hat{\theta}_k^{(2)} = \theta_k \epsilon_k. \tag{4.63}$$

Let the mean and variance of $\epsilon_k$ be denoted $\mu_\epsilon$ and $\sigma^2_\epsilon$, respectively. According to model $\mathrm{m}_t$, these moments are independent of $s_{iq}$ and $n_{r_k}$. In addition, we now assume that the error moments are independent of the site $k$ as well. In Equations (4.62) and (4.63), by letting $\mu_{\hat{\theta}_k^{(2)}} = \theta_k + \mu_\epsilon$ and $\sigma^2_{\hat{\theta}_k^{(2)}} = \sigma^2_\epsilon$, results are obtained for the additive model. In the same manner, by letting $\mu_{\hat{\theta}_k^{(2)}} = \theta_k \mu_\epsilon$ and $\sigma^2_{\hat{\theta}_k^{(2)}} = \theta_k^2 \sigma^2_\epsilon$, we get results for the multiplicative model. Consider in particular the latter model. For this, Equations (4.56) and (4.57) modify to

$$\gamma\left(\hat{y}^{(2)}\right)_k = \frac{y_k}{\mu_\epsilon} \tag{4.64}$$

$$\delta\left(\hat{y}^{(2)}\right)_k = \left(\frac{y_k}{\mu_\epsilon}\right)^2 \left(\frac{1-\theta_k}{y_k \theta_k} + \frac{\sigma^2_\epsilon}{\mu^2_\epsilon}\right); \tag{4.65}$$

Equations (4.58) and (4.59) to

$$\gamma\big(\hat{z}^{(2)}\big)_k \approx \frac{z_k}{\mu_\epsilon} \tag{4.66}$$

$$\delta\big(\hat{z}^{(2)}\big)_k \approx \frac{1}{\sigma_\epsilon^2+\mu_\epsilon^2}\frac{1-\theta_k}{\theta_k}\left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) + \left(\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(2)}\big)_k\,; \tag{4.67}$$

and Equations (4.60) and (4.61) to

$$\gamma\Big(\hat{E}^{(2)}\Big)_k \approx \frac{y_k}{\mu_\epsilon}\left(1 - R\frac{z_k}{y_k}\right) \tag{4.68}$$

$$\delta\Big(\hat{E}^{(2)}\Big)_k \approx \frac{R^2}{\sigma_\epsilon^2+\mu_\epsilon^2}\frac{1-\theta_k}{\theta_k}\left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) + \left(1 - R^2\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(2)}\big)_k\,. \tag{4.69}$$

For the multiplicative model, the propitious case $\mu_{\theta_k^{(2)}} = \theta_k$, for which $\hat{t}_{\hat{y}^{(2)}}$ and $\hat{t}_{\hat{z}^{(2)}}$ are unbiased and $\hat{R}^{(2)}$ approximately unbiased, corresponds to an error mean equal to unity ($\mu_\epsilon = 1$).

### 4.4.3 Summary of theoretical findings

We investigated the statistical properties of various estimators of the parameters $t_y$, $t_z$ and $R$. The estimators are all based on estimates, rather than the true values, of $y$ and $z$ for sampled sites. In general, for each estimator, the sign of its possible bias (as estimator of the true population entity) is unknown. Further, the sign of the difference between the estimator's variance and that of the corresponding prototype estimator is unknown. A key issue is whether the estimators of the values of $y$ and $z$ are unbiased or not. If they are, the estimators of $t_y$ and $t_z$ are unbiased as well, and the estimator of $R$ approximately unbiased. The variances of the estimators of $t_y$, $t_z$ and $R$ are then surely larger than those of the corresponding prototype estimators.

Under Strategy 0, the values of both $y$ and $z$ are always underestimated, and so are $t_y$ and $t_z$. In what direction (if any) missing data bias the estimator of $R$ remains unknown. If, by chance, the registration probabilities $\theta_k$

are equal for all sites, the Strategy 0 estimator of $R$ is, however, not biased by missing data. Under Strategy 1, the estimators of the $y$ and $z$ values are unbiased if the (conditional) expected number of imputed vehicles, and the number of missing vehicles, coincide. If the error in the number of imputed vehicles is multiplicative, the variance expressions slightly simplify. Still, they contain a number of unknown model parameters: the registration probabilities $\theta_k$ as well as the error variance. Finally, under Strategy 2, the estimators of $y$ and $z$ are (approximately) unbiased if the estimator of $\theta_k$ is. If the error in the estimator of $\theta_k$ is multiplicative, as under Strategy 1, this allows us to simplify the variance expressions somewhat.

The investigation provided us with formulae for a number of cases and special cases. The expressions are, however, typically quite complicated and include several unknown entities, which make the expressions hard to evaluate theoretically. Detailed experiments would be needed to complete the picture.

## 4.5 Empirical study

### 4.5.1 Study objectives

In Section 4.4, the statistical properties of the various estimators were investigated. The results, however, relied on model assumptions which may not reflect reality. Also, the results did not allow us to draw general conclusions on which strategy that is preferable. The need for model evaluations and further guidance in the choice of estimation strategy motivated the collection of empirical data.

The main objectives of this study were to investigate:

- The **forming of registration homogeneity groups**. For reasons stated in Section 4.3.1, the smallest groups considered are watch-hours. We would, however, like to evaluate the option to join several hours into larger groups. Can unnecessarily large variation in group registration rates be avoided this way?

- The assumptions of the **multiplicative imputation error model**. Is

| Site no. | Street name | Street characteristics |
|---|---|---|
| 1 | Nygårdsvägen | Feeding lane for suburban area |
| 2 | G:a Tanneforsvägen | Part of major route encircling the city |
| 3 | Drottninggatan | Inner city street |
| 4 | Kaserngatan | Part of major route encircling central city |
| 5 | Bergsvägen | Throughfare |

Table 4.1: Selected road sites in the city of Linköping, Sweden.

the error in the number of imputed vehicles multiplicative (as suggested in Section 4.4.2)? Is the number of imputed vehicles conditionally unbiased for the true number of missing vehicles (conditional on the number of registered vehicles)?

- The assumptions of the **error model for** $\hat{\theta}^{(2)}$. Is the functional relationship between $\hat{\theta}_k^{(2)}$ and $\theta_k$ additive or multiplicative (or neither of them)? Is the estimator $\hat{\theta}_k^{(2)}$ unbiased for the true registration probability?

Finally, we are interested in the empirical behavior of the proposed estimators of flow and travel time for a road site.

Note the limited scope of the study. We did not attempt to perform an experiment detailed enough to estimate all unknown entities included in the formulae in Section 4.4, and additional work would be needed to make full use of our theoretical results.

### 4.5.2 Design of the study

Data were collected for five road sites in the city of Linköping, Sweden. The sites were purposively chosen to represent different types of traffic environments. However, to simplify, the study was limited to two-way, two-lane streets with a speed limit of 50 kilometers per hour – a typical road design and speed limit for Swedish urban roads. For details on selected sites, see Table 4.1.

In each site, data collection went on for 24 successive hours by use of two pairs of pneumatic tubes and three traffic analyzers. The installation of the

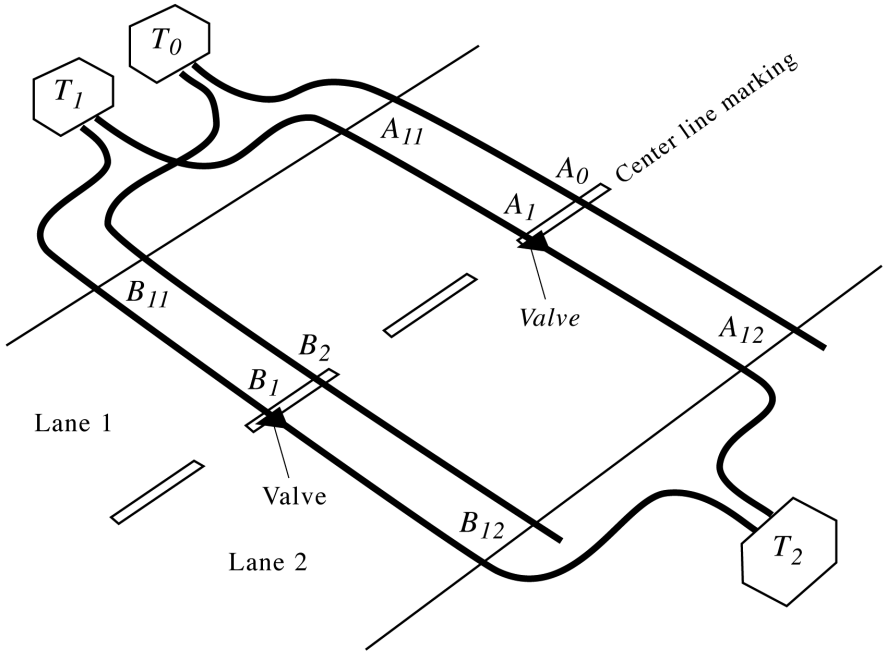


Figure 4.1: Installation of the measurement equipment. (Illustration: Björn Böke)

equipment is outlined in Figure 4.1. One pair of tubes $(A_0, B_0)$ connected to a traffic analyzer $T_0$ was used for simultaneous observation of vehicles on both street lanes. The second pair of tubes $(A_1, B_1)$ was installed in parallel with the first, only with a slight lateral displacement. The length of the displacement, about 30 centimeters, was chosen to satisfy two criteria: (1) sufficiently long to prevent the tubes from disturbing each other, yet (2) sufficiently short to ensure that passing vehicles keep the same speed as while passing $(A_0, B_0)$. By use of valves, the tubes $(A_1, B_1)$ were *plugged* at the centerline marking of the street. This procedure enables separate measurement of the traffic on each lane. The tube ends on each side of the valves were connected to a traffic analyzer. In Figure 4.1, lane 1 is measured by the tube parts $(A_{11}, B_{11})$ connected to traffic analyzer $T_1$; lane 2

by $(A_{12}, B_{12})$ connected to traffic analyzer $T_2$.

The plugging method has been developed at the SNRA as a means of improving data quality. The registration task facing $T_1$ and $T_2$ is much easier (and hence less subject to measurement errors) than that of $T_0$: vehicles do not meet while passing the tubes, fewer vehicles pass, and their direction is known beforehand. Despite this, the method is rarely used in the speed survey. The main reason is that it is more time-consuming to use than the unplugged alternative; the valves need to be mounted in the tubes, and the laying out of the tubes demands greater care. Another drawback of the method is the vulnerability of the valves. If a valve, for instance, becomes filled with rain water, or squeezed by a vehicle wheel, it may quit working. According to plan, all experimental data were to be collected August 21-22, 2001. Due to valve malfunctioning, however, sites 1, 3 and 4 were remeasured September 24-25, 2001.

In the experiment, the data set produced by $T_0$ is intended to represent the output one would expect from a measurement performed within the regular survey. The data set produced jointly by $T_1$ and $T_2$, on the other hand, is intended to represent the 'truth.'

### 4.5.3 Data processing

We start by introducing some notation. Consider site $k$ during hour $h$ as measured by traffic analyzer $T_d; k = 1, ..., 5; h = 1, ..., 24; d = 0, 1, 2$. For $(k, h, T_d)$, let $n_{r_{kh(d)}}$ and $n_{I_{kh(d)}}$ denote the number of registered and imputed vehicles, respectively, and $(ME)_{kh(d)}$ the measurement efficiency. The corresponding numbers of vehicles for a 24-hour period are $n_{r_{k(d)}} = \sum_{h=1}^{24} n_{r_{kh(d)}}$ and $n_{I_{k(d)}} = \sum_{h=1}^{24} n_{I_{kh(d)}}$. The measurement efficiency for a 24-hour period is not available from the analyzer, but we calculate an approximate value as $(ME)_{k(d)} = \sum_{h=1}^{24} n_{r_{kh(d)}} (ME)_{kh(d)} / n_{r_{k(d)}}$.

The observational data from $M_0$ are to be compared with the joint data from $M_1$ and $M_2$. To simplify, let $n_{r_{kh(1+2)}} = n_{r_{kh(1)}} + n_{r_{kh(2)}}$ and $n_{r_{k(1+2)}} = \sum_{h=1}^{24} n_{r_{kh(1+2)}}$, and let the entities $n_{I_{kh(1+2)}}$ and $n_{I_{k(1+2)}}$ be defined correspondingly. For the set $r_{kh(1+2)}$ of size $n_{r_{kh(1+2)}}$ of vehicles registered during hour $h$ by $M_1$ or $M_2$, the total travel time is $\sum_{r_{kh(1+2)}} x_v$. The total travel time for

| | $T_0$ | | | $T_1$ | | | $T_2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Site | $n_r$ | $n_I$ | $ME$ | $n_r$ | $n_I$ | $ME$ | $n_r$ | $n_I$ | $ME$ |
| 1 | 5690 | 69 | 98.3 | 2976 | 13 | 99.5 | 2763 | 13 | 99.3 |
| 2 | 14314 | 747 | 94.7 | 7924 | 249 | 96.9 | 6989 | 58 | 98.6 |
| 3 | 10850 | 2856 | 83.9 | 5772 | 2038 | 81.0 | 6546 | 527 | 93.2 |
| 4 | 10948 | 181 | 97.8 | 5363 | 47 | 98.6 | 5730 | 80 | 98.3 |
| 5 | 11259 | 338 | 96.6 | 5660 | 8 | 99.8 | 5907 | 66 | 98.3 |

Table 4.2: The number of registered vehicles ($n_r$) and imputed vehicles ($n_I$), and the measurement efficiency ($ME$) in percent, by site and traffic analyzer.

a 24-hour period is $\sum_{r_{k(1+2)}} x_v = \sum_{h=1}^{24} \sum_{r_{kh(1+2)}} x_v$.

A summary of the outcome of the measurements is given in Table 4.2. If the data collection had turned out perfectly, the table would have contained nothing but zeroes in the $n_I$ columns for analyzer $T_1$ and $T_2$ (the MEs for $T_1$ and $T_2$ had then also been 100 percent.) Table 4.2 exposes, however, that even though the use of valves reduced the need for imputations, it did not succeed in eliminating it. Site 3 is our real 'problem child'; on this busy inner city street, all three analyzers encountered difficulties. In particular, on lane 1, the traffic approaches a traffic signal. The signal causes the vehicles to either move slowly with short time gaps or to stand in line – an especially difficult measurement situation. We judge that the resulting large number of imputations, and low ME, makes the $T_1$ data useless for our purposes. For this reason, only the lane 2 part of the $T_0$ data, and the $T_2$ data, are used in the coming analysis of site 3.

In certain cases, imputations in the $T_1$ or $T_2$ data can be matched with vehicles properly registered by $T_0$. These situations are most likely to occur when passing vehicles straddle the valves. For each site, we compared the data files from $T_0$, $T_1$ and $T_2$, looking for imputations in $T_1$ and $T_2$ which, with reasonable certainty, could be matched with registered vehicles in $T_0$. These imputations were then substituted by the registered vehicles. Table 4.3 shows the number of imputed vehicles that were substituted, how many registered vehicles they were substituted by, and how many unsubstituted vehicles were left in the adjusted data files. We see that the number of substituted vehi-

| | $T_1$ | | | $T_2$ | | |
|---|---|---|---|---|---|---|
| Site | $n_I^*$ | $n_S$ | $n_I - n_I^*$ | $n_I^*$ | $n_S$ | $n_I - n_I^*$ |
| 1 | 4 | 3 | 9 | 2 | 1 | 11 |
| 2 | 46 | 30 | 203 | 23 | 20 | 35 |
| 3 | – | – | – | 62 | 50 | 465 |
| 4 | 34 | 23 | 13 | 47 | 29 | 33 |
| 5 | 0 | 0 | 8 | 42 | 33 | 24 |

Table 4.3: The number of imputed vehicles that were substituted ($n_I^*$), how many registered vehicles they were substituted by ($n_S$), and the remaining number of imputations ($n_I - n_I^*$), by site and analyzer. For site 3, only data from $M_2$ were examined.

cles is consistently larger than the number of substitutes. This makes sense since a vehicle straddling the valves typically produces two ore more imputed vehicles, distributed among $T_1$ and $T_2$.

### 4.5.4 Estimation

In the estimation, for $T_1$ and $T_2$, the number of registered vehicles $n_{r_{kh(1+2)}}$, and their associated total travel time $\sum_{r_{kh(1+2)}} x_v$, is calculated from the *adjusted* data set $r_{kh(1+2)}$ (see Section 4.5.3) with no distinction made between 'truly registered' and 'substitute' vehicles. From Table 4.3, after adjustments, the sets $r_{kh(1+2)}$ still contain imputed vehicles. Some of these imputations are probably correct, whereas others ought to be removed. For each selected site and each measured hour, to form a basis of later analysis, we calculate a number of estimates. Since there is no way for us to know how to treat each imputation case, our estimates are calculated both with the imputations in $r_{kh(1+2)}$ retained and removed. Estimates for which the imputations are retained are indexed by 'wi'; estimates for which they are removed by 'woi'.

For site $k$ and hour $h$, the following estimates are calculated.

**Estimates of registration probability** The registration probability $\theta_{kh}$ for $(k, h)$ is estimated by

$$\hat{\theta}_{kh,\text{woi}} = \frac{n_{r_{kh(0)}}}{n_{r_{kh(1+2)}}} \tag{4.70}$$

$$\hat{\theta}_{kh,\text{wi}} = \frac{n_{r_{kh(0)}}}{n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}}}. \tag{4.71}$$

In both Equation (4.70) and (4.71), the denominator is intended to represent the true flow.

**Estimates of multiplicative imputation error** Consider the multiplicative imputation error model in Section 4.4.2. For $(k, h)$, the multiplicative error $\varepsilon_{kh}$ is estimated by

$$\hat{\varepsilon}_{kh,\text{woi}} = \frac{n_{I_{kh(0)}}}{n_{r_{kh(1+2)}} - n_{r_{kh(0)}}} \tag{4.72}$$

$$\hat{\varepsilon}_{kh,\text{wi}} = \frac{n_{I_{kh(0)}}}{n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}} - n_{r_{kh(0)}}}. \tag{4.73}$$

In both Equation (4.72) and (4.73), the denominator is intended to represent the number of vehicles missing in the $T_0$ data.

**Estimates of error in $\hat{\theta}^{(2)}$** Consider the additive error model for $\hat{\theta}^{(2)}$ in Equation (4.62). For $(k, h)$, the error $\epsilon_{kh}$ is estimated by

$$\hat{\epsilon}_{kh,\text{woi}} = (ME)_{kh(0)} - \hat{\theta}_{kh,\text{woi}} \tag{4.74}$$

$$\hat{\epsilon}_{kh,\text{wi}} = (ME)_{kh(0)} - \hat{\theta}_{kh,\text{wi}}. \tag{4.75}$$

Further consider the multiplicative model for $\hat{\theta}^{(2)}$ in Equation (4.63). Under this model, the error $\epsilon_{kh}$ is estimated by

$$\hat{\epsilon}_{kh,\text{woi}} = \frac{(ME)_{kh(0)}}{\hat{\theta}_{kh,\text{woi}}} \tag{4.76}$$

$$\hat{\epsilon}_{kh,\text{wi}} = \frac{(ME)_{kh(0)}}{\hat{\theta}_{kh,\text{wi}}}. \tag{4.77}$$

The resulting estimates are presented, by site, in graphs in [21, Appendix C]. In the next section (Section 4.5.5), our analysis is illustrated with selected graphs from site 4.

For each selected site, we also calculate the following estimates:

**Estimates of flow and travel time** For site $k$, the traffic flow $y_k$ and travel time $z_k$ are estimated by use of the formulae in Section 4.3. The resulting estimates under Strategy $c$ ($c = 0, 1, 2$) are denoted $\hat{y}^{(c)}_{k(0)}$ and $\hat{z}^{(c)}_{k(0)}$, respectively, where subscript (0) indicates that only $T_0$ data are used for the calculations. For easy evaluation of the estimates, we continue by *standardizing* them. For site $k$ and Strategy $c$, the standardized flow estimates without and with imputations are

$$\tilde{y}^{(c)}_{k,\mathrm{woi}} = \frac{\hat{y}^{(c)}_{k(0)}}{n_{r_{k(1+2)}}} \tag{4.78}$$

$$\tilde{y}^{(c)}_{k,\mathrm{wi}} = \frac{\hat{y}^{(c)}_{k(0)}}{n_{r_{k(1+2)}} + n_{I_{kh(1+2)}}} \tag{4.79}$$

whereas the standardized estimate of travel time is

$$\tilde{z}^{(c)}_{k,\mathrm{woi}} = \frac{\hat{z}^{(c)}_{k(0)}}{\sum_{r_{k(1+2)}} x_v}. \tag{4.80}$$

We choose to standardize the travel time estimates only by the sum of travel times for vehicles registered in the valve measurements (that is, the imputations in the latter are ignored). The reason is that we do not trust the travel times of imputed vehicles.

**Estimates of average speed** For site $k$, define the average speed (also known as the *space mean speed* or *harmonic mean speed* [13, Section 2.2.2])

$$u_k = \frac{1}{\frac{1}{y_k}\sum_{v=1}^{y_k}\frac{1}{u_v}} = \frac{y_k}{z_k} \tag{4.81}$$

where $u_v$ is the speed at which vehicle $v$ passes the site. Under Strategy $c$ ($c = 0, 1, 2$), $u_k$ is estimated by the ratio

$$\hat{u}^{(c)}_k = \frac{\hat{y}^{(c)}_{k(0)}}{\hat{z}^{(c)}_{k(0)}}. \tag{4.82}$$

Again for easy evaluation, the estimates are standardized. For site $k$ and Strategy $c$, the standardized average speed estimates without and

with imputations are

$$\tilde{u}^{(c)}_{k,\text{woi}} = \frac{\hat{u}^{(c)}_k}{n_{r_{k(1+2)}} / \sum_{r_{k(1+2)}} x_v} \tag{4.83}$$

$$\tilde{u}^{(c)}_{k,\text{wi}} = \frac{\hat{u}^{(c)}_k}{\left(n_{r_{k(1+2)}} + n_{I_{kh(1+2)}}\right) / \sum_{r_{k(1+2)}} x_v}. \tag{4.84}$$

In both Equation (4.83) and (4.84), the denominator is intended to represent the true average speed.

The standardized estimates of $y_k$ and $z_k$ are presented in Table 4.6 and 4.7, respectively, whereas the standardized estimates of $u_k$ are given in Table 4.8.

### 4.5.5 Analysis

**The forming of registration homogeneity groups**

When the estimated registration probabilities $\hat{\theta}_{kh}$ are plotted against the 'true' flows, the probability estimates are often fairly constant for adjacent flow levels – see Figure 4.2 – which speaks in favor of merging hours into larger homogeneity groups by flow. It is, however, not obvious where to draw the lines between groups: in Figure 4.2, the relationship between registration probability and flow is quite smooth. (In practice, the true flows are obviously not available, but a grouping of hours would need to be based on registered flows.)

**Evaluation of the multiplicative imputation error model**

If the multiplicative imputation error model is correct, the estimated errors $\hat{\varepsilon}_{kh,\text{woi}}$ and $\hat{\varepsilon}_{kh,\text{wi}}$ should not reveal any obvious patterns if plotted against other variables. However, when we plot the errors against the number of missing vehicles, for some sites, we discern a tendency of the error variance to decrease as the number of missing vehicles increases (see Figure 4.3). Due to the scarcity of observations for large numbers of missing vehicles, it is nevertheless hard to draw any certain conclusions. When the errors are plotted against the number of registered vehicles, on the other hand, no unusual structures are apparent.

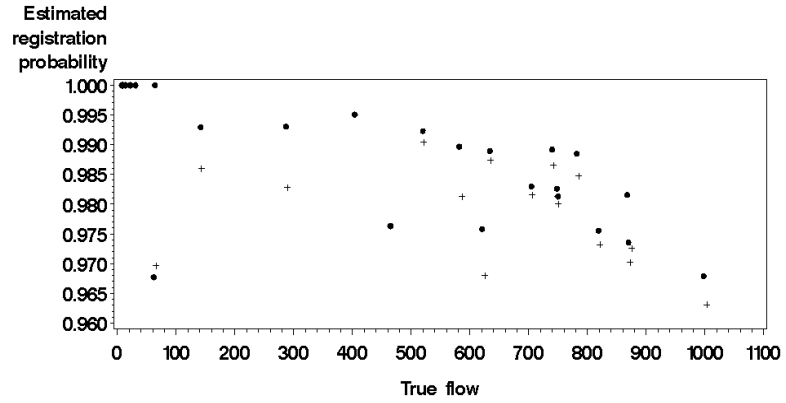


Figure 4.2: Estimated registration probabilities versus 'true' flows for site 4. One data point corresponds to one hour. Dots indicate $\hat{\theta}_{kh,\mathrm{woi}}$ plotted against $n_{r_{kh(1+2)}}$; plus signs indicate $\hat{\theta}_{kh,\mathrm{wi}}$ plotted against $n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}}$.

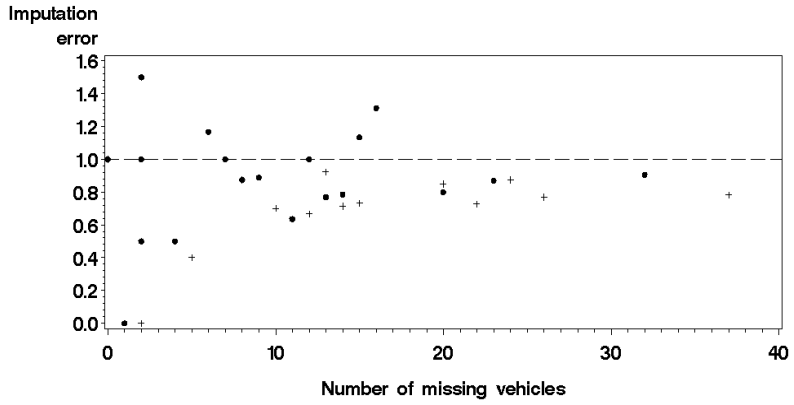


Figure 4.3: Estimated imputation errors versus 'true' numbers of missing vehicles for site 4. One data point corresponds to one hour. Dots indicate $\hat{\varepsilon}_{kh,\mathrm{woi}}$ plotted against $n_{r_{kh(1+2)}} - n_{r_{kh(0)}}$; plus signs indicate $\hat{\varepsilon}_{kh,\mathrm{wi}}$ plotted against $n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}} - n_{r_{kh(0)}}$.

To investigate whether the variance of the errors is independent of the site (as the model states), we formulate an ANOVA model:

$$\hat{\varepsilon}_{kh} = \alpha + \beta_k + e_{kh} \begin{cases} k = 1, 2, ..., b \\ h = 1, 2, ..., c \end{cases} \tag{4.85}$$

where $\hat{\varepsilon}_{kh}$ may be either $\hat{\varepsilon}_{kh,\text{woi}}$ or $\hat{\varepsilon}_{kh,\text{wi}}$, $b$ is the number of experiment sites, and $c$ the number of observed hours within site. In practice, $b = 5$ and $c = 24$. The parameter $\alpha$ is an overall mean, $\beta_k$ is the random effect of the $k$th site, and $e_{kh}$ is a random error. We assume that the $\beta_k$'s are normally and independently distributed (NID) with mean zero and variance $\sigma_\beta^2$, the $e_{kh}$'s $NID\,(0, \sigma_e^2)$, and that $\beta_k$ and $e_{kh}$ are independent. This *random effects model* (see, for instance, [25, Section 3-7], [26, Chapter 24]) actually presupposes that our experiment sites were selected randomly from all possible sites (all urban road meters in Sweden). Then, inference could be made about all sites. In our case, since the sites were chosen purposively, we must interpret our results with caution.

We start by testing the hypothesis $H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$. The ANOVA's for our data are shown in Appendix E.2.1. We see that our conclusions differ for different treatments of the imputations in the valve measurements. If the imputations are removed, the null hypothesis is not rejected at the 0.05 level of significance. If, on the other hand, the imputations are retained, the null hypothesis is rejected. Hence, we do not get a clear indication as to whether there is a variability between sites or not.

We are further interested in estimating the mean $\mu_{\hat{\varepsilon}} = \alpha$ of $\hat{\varepsilon}_{kh}$. From [26, Eq. (24.15)], a $100(1-\alpha)$ percent confidence interval on $\mu_{\hat{\varepsilon}}$ is given by

$$\bar{\bar{\hat{\varepsilon}}} \pm t_{1-\alpha/2,b-1} \sqrt{\frac{MS_{\text{site}}}{bc}} \tag{4.86}$$

where $\bar{\bar{\hat{\varepsilon}}} = \sum_{k=1}^{b} \sum_{h=1}^{c} \hat{\varepsilon}_{kh}$ and $MS_{\text{site}}$ is the mean square due to sites. By use of Equation (4.86) and the ANOVA's in Appendix E.2.1, the interval estimates of $\mu_{\hat{\varepsilon}}$ in Table 4.4 are obtained. Again, our conclusions differ for different treatments of the imputations in the valve measurements. If the imputations are removed, the hypothesis of $\mu_{\hat{\varepsilon}} = 1$ is not rejected at the 0.05 level of significance. If, on the other hand, the imputations are retained,

| Imputation error | 95% confidence interval for $\mu_{\hat{\varepsilon}}$ |
|---|---|
| $\hat{\varepsilon}_{kh,\mathrm{woi}}$ | $1.10398 \pm 0.22136$ |
| $\hat{\varepsilon}_{kh,\mathrm{wi}}$ | $0.80832 \pm 0.16808$ |

Table 4.4: Confidence intervals for $\mu_{\hat{\varepsilon}}$, calculated with the imputations in the valve measurements removed and retained, respectively.

the hypothesis is rejected. Thus, it remains an open question whether the number of imputed vehicles is conditionally unbiased for the true number of missing vehicles or not.

**Evaluation of the error model for $\hat{\theta}^{(2)}$**

No matter if the additive or the multiplicative error model for $\hat{\theta}_{kh}^{(2)}$ is considered; if the model is correct, the observed errors in $\hat{\theta}_{kh}^{(2)}$ should not reveal any obvious patterns if plotted against $\hat{\theta}_{kh}$ or the registered flows $n_{r_{kh(0)}}$. For the estimator $\hat{\theta}_{kh}^{(2)}$ to be unbiased for $\hat{\theta}_{kh}$ (and thus, hopefully, for the true registration probability $\theta_{kh}$) the errors, when plotted against $\hat{\theta}_{kh}$, ought to scatter around the relevant reference line (placed at level zero for the additive errors; level one for the multiplicative errors).

We start by the observed errors under the *additive* model (Equations (4.74) and (4.75)). In Figure 4.4, we see a tendency for the plus signs to scatter above the reference line, and for the dots to scatter below the line. These point swarms represent the two extremes in terms of treatment of imputations in the valve measurements – the location of the 'true' swarm ought to be somewhere in between. We do not see a strong tendency of the error variance to change with the size of $\hat{\theta}_{kh}$. The scarcity of observations for small values of $\hat{\theta}_{kh}$ makes it hard though to draw any certain conclusions. When the errors are plotted against the number of registered vehicles – see Figure 4.5 – we see signs of dependency between the errors and the registered flows. It seems that the true probability is overestimated for low flows, but underestimated for high flows.

Now consider the observed errors under the *multiplicative* model (Equations (4.76) and (4.77)). In Figure 4.6, we see again the tendency of the two

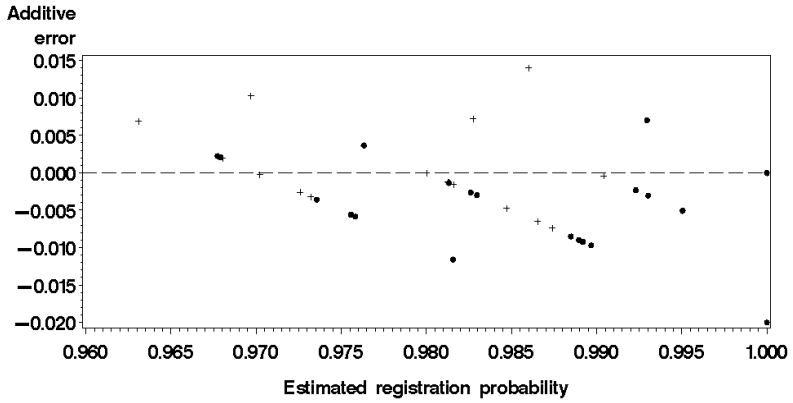


Figure 4.4: Estimated errors in $\hat{\theta}_{kh}^{(2)}$ under the *additive* error model, versus estimated registration probabilities, for site 4. One data point corresponds to one hour. Dots indicate $\hat{\epsilon}_{kh,\mathrm{woi}}$ plotted against $\hat{\theta}_{kh,\mathrm{woi}}$; plus signs indicate $\hat{\epsilon}_{kh,\mathrm{wi}}$ plotted against $\hat{\theta}_{kh,\mathrm{wi}}$. A horizontal reference line indicates the desired expected value of the errors. The diagonal pattern in the observations is a result of the MEs (used in the calculations of the errors) only being available as integers.

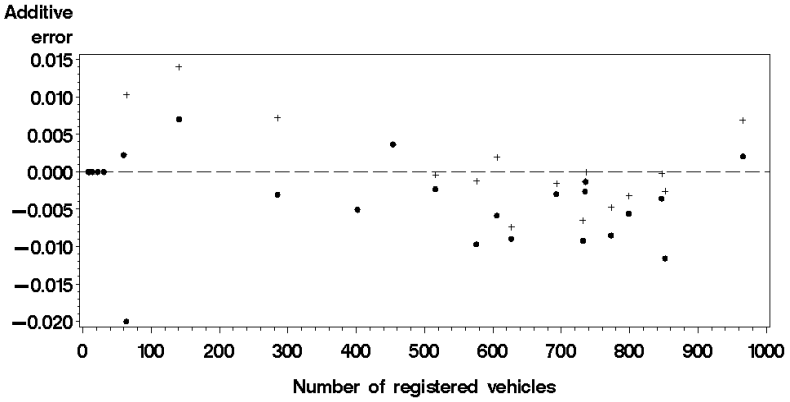


Figure 4.5: Estimated errors in $\hat{\theta}_{kh}^{(2)}$ under the *additive* error model, versus registered flow, for site 4. One data point corresponds to one hour. Dots indicate $\hat{\epsilon}_{kh,\text{woi}}$ plotted against $n_{r_{kh(0)}}$; plus signs indicate $\hat{\epsilon}_{kh,\text{wi}}$ plotted against $n_{r_{kh(0)}}$. A horizontal reference line indicates the desired expected value of the errors.

point swarms to lie above and below the reference line. And again, as far as we can tell, the error variance seems to be independent of $\hat{\theta}_{kh}$. When the errors are plotted against the number of registered vehicles – see Figure 4.7 – we see, however, signs of dependency between the variables. The pattern is the same as in Figure 4.5.

Both the additive and the multiplicative error model states that the variance of the errors is independent of the site. To investigate this, we use the same ANOVA model as in Equation (4.85) – only with $\hat{\varepsilon}_{kh}$ replaced by $\hat{\epsilon}_{kh}$ (which may represent either $\hat{\epsilon}_{kh,\text{woi}}$ in Equation (4.74) or (4.76), or $\hat{\epsilon}_{kh,\text{wi}}$ in Equation (4.75) or (4.77)). Again, the aim is to test the hypothesis $H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$. The corresponding ANOVA tables are given in Appendices E.2.2 and E.2.3. We see that throughout, the null hypothesis is rejected at 0.05 level of significance. In other words, contrary to what our models state, there seems to be a variability due to site in the error in $\hat{\theta}_{kh}^{(2)}$.

We proceed by estimating the mean $\mu_{\hat{\epsilon}} = \alpha$ of $\hat{\epsilon}_{kh}$. By use of Equation (4.86) with $\hat{\varepsilon}_{kh}$ replaced by $\hat{\epsilon}_{kh}$, and the ANOVA's in Appendices E.2.2

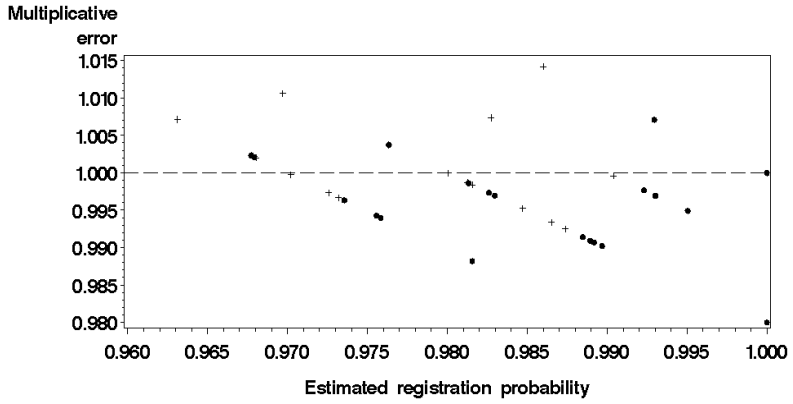


Figure 4.6: Estimated errors in $\hat{\theta}_{kh}^{(2)}$ under the *multiplicative* error model, versus estimated registration probabilities, for site 4. One data point corresponds to one hour. Dots indicate $\hat{\epsilon}_{kh,\mathrm{woi}}$ plotted against $\hat{\theta}_{kh,\mathrm{woi}}$; plus signs indicate $\hat{\epsilon}_{kh,\mathrm{wi}}$ plotted against $\hat{\theta}_{kh,\mathrm{wi}}$. A horizontal reference line indicates the desired expected value of the errors. The diagonal pattern in the observations is a result of the MEs (used in the calculations of the errors) only being available as integers.

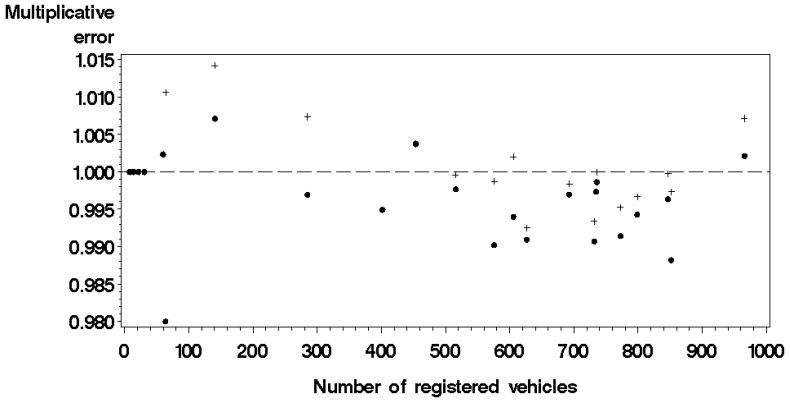


Figure 4.7: Estimated errors in $\hat{\theta}_{kh}^{(2)}$ under the *multiplicative* error model, versus registered flow, for site 4. One data point corresponds to one hour. Dots indicate $\hat{\epsilon}_{kh,\mathrm{woi}}$ plotted against $n_{r_{kh(0)}}$; plus signs indicate $\hat{\epsilon}_{kh,\mathrm{wi}}$ plotted against $n_{r_{kh(0)}}$. A horizontal reference line indicates the desired expected value of the errors.

and E.2.3, the interval estimates of $\mu_{\hat{\epsilon}}$ in Table 4.5 are obtained. At the 0.05 level of significance, for the additive error model, the hypothesis of $\mu_{\hat{\epsilon}} = 0$ is not rejected. Also, for the multiplicative model, the hypothesis of $\mu_{\hat{\epsilon}} = 1$ is not rejected. These results stand, no matter how the imputations in the valve measurements are treated.

### Empirical behavior of proposed estimators

Obviously, our limited data material does not allow us to study the long-run performances of the estimators of flow and travel time, but can only give some indication of the same. In Tables 4.6 and 4.7, as expected, the Strategy 0 estimates all fall below one. The missing data adjusted estimates under Strategy 1 and 2, on the other hand, look quite well. Depending on what entity is used to standardize the flow estimates, for both strategies, their averages land slightly below or above one (with the true average expected to be somewhere in between). The averages of the standardized travel estimates under Strategy 1 and 2 land slightly above one. However, most likely, the

| Error model for $\hat{\theta}^{(2)}$ | Imputation error | 95% confidence interval for $\mu_{\hat{\epsilon}}$ |
|---|---|---|
| Additive | $\hat{\epsilon}_{kh,\mathrm{woi}}$ | $-0.00743 \pm 0.00911$ |
| Additive | $\hat{\epsilon}_{kh,\mathrm{wi}}$ | $0.00796 \pm 0.01706$ |
| Multiplicative | $\hat{\epsilon}_{kh,\mathrm{woi}}$ | $0.99103 \pm 0.01301$ |
| Multiplicative | $\hat{\epsilon}_{kh,\mathrm{wi}}$ | $1.00963 \pm 0.02105$ |

Table 4.5: Confidence intervals for $\mu_{\hat{\epsilon}}$ by error model, calculated with the imputations in the valve measurements removed and retained, respectively.

| Site | $\tilde{y}^{(0)}_{\mathrm{woi}}$ | $\tilde{y}^{(1)}_{\mathrm{woi}}$ | $\tilde{y}^{(2)}_{\mathrm{woi}}$ | $\tilde{y}^{(0)}_{\mathrm{wi}}$ | $\tilde{y}^{(1)}_{\mathrm{wi}}$ | $\tilde{y}^{(2)}_{\mathrm{wi}}$ |
|---|---|---|---|---|---|---|
| 1 | 0.99077 | 1.00279 | 1.00783 | 0.98733 | 0.99931 | 1.00429 |
| 2 | 0.95656 | 1.00648 | 1.01129 | 0.94165 | 0.99079 | 0.99550 |
| 3 (one dir.) | 0.82990 | 1.05230 | 1.04836 | 0.77524 | 0.98301 | 0.97928 |
| 4 | 0.98232 | 0.99856 | 1.00476 | 0.97846 | 0.99464 | 1.00079 |
| 5 | 0.97052 | 0.99966 | 1.00440 | 0.96793 | 0.99699 | 1.00171 |
| Mean | 0.94601 | 1.01196 | 1.01530 | 0.93012 | 0.99295 | 0.99631 |

Table 4.6: Standardized estimates of flow, by site. (For site 3, only data from $T_2$ are used.)

| Site | $\tilde{z}^{(0)}_{\mathrm{woi}}$ | $\tilde{z}^{(1)}_{\mathrm{woi}}$ | $\tilde{z}^{(2)}_{\mathrm{woi}}$ |
|---|---|---|---|
| 1 | 0.98854 | 1.00051 | 1.00550 |
| 2 | 0.94958 | 1.00097 | 1.00570 |
| 3 (one dir.) | 0.81594 | 1.04619 | 1.04157 |
| 4 | 0.98234 | 0.99864 | 1.00483 |
| 5 | 0.96555 | 0.99475 | 0.99947 |
| Mean | 0.94039 | 1.00821 | 1.01142 |

Table 4.7: Standardized estimates of travel time, by site. (For site 3, only data from $T_2$ are used.)

| Site | $\tilde{u}^{(0)}_{k,\text{woi}}$ | $\tilde{u}^{(1)}_{k,\text{woi}}$ | $\tilde{u}^{(2)}_{k,\text{woi}}$ | $\tilde{u}^{(0)}_{k,\text{wi}}$ | $\tilde{u}^{(1)}_{k,\text{wi}}$ | $\tilde{u}^{(1)}_{k,\text{wi}}$ |
|---|---|---|---|---|---|---|
| 1 | 1.00226 | 1.00227 | 1.00227 | 0.99878 | 0.99879 | 0.99879 |
| 2 | 1.00735 | 1.00551 | 1.00554 | 0.99165 | 0.98983 | 0.98986 |
| 3 (one dir.) | 1.01711 | 1.00584 | 1.00647 | 0.95013 | 0.93960 | 0.94019 |
| 4 | 0.99999 | 0.99992 | 0.99991 | 0.99605 | 0.99599 | 0.99598 |
| 5 | 1.00515 | 1.00493 | 1.00492 | 1.00247 | 1.00225 | 1.00224 |
| Mean | 1.01308 | 1.00146 | 1.00199 | 0.99720 | 0.98576 | 0.98629 |

Table 4.8: Standardized estimates of average speed, by site. (For site 3, only data from $T_2$ are used.)

travel time estimates are standardized with too small a value (since the imputations are ignored). In all, from Tables 4.6 and 4.7, it is far from obvious which adjustment strategy (1 or 2) ought to be recommended.

Now consider the standardized estimates of average speed in Table 4.8. Formally, we can not use these estimates to evaluate the performances of present or proposed estimators of $R$. Still, the average speed $u_k$ is the counterpart on 'element-level' to the average speed $R$ for all roads. The estimates in Table 4.8, including the Strategy 0 estimates, are very close to one. We take this as a small hint that missing data adjustments are not a necessity when estimating $R$.

### 4.5.6 Summary of empirical findings

Since we could not calculate registration probabilities for individual vehicles, but only by hour, we were not able to check the assumptions of the registration model. However, we seized the opportunity to see whether the data indicated in favor of merging hours into larger homogeneity groups. This seemed to be the case, but it did not appear immediately clear where the borders should be drawn.

Under the multiplicative imputation error model, the conditional expectation and variance of the errors are independent of the number of registered vehicles and of the site. Our data gave us no obvious reason to reject independency between the errors and the number of registered vehicles. We were not able to establish whether the errors are site-independent or not, since

the result of our (approximative) ANOVA test proved to be sensitive to how imputations in the valve measurements were treated. For the same reason, we did not get a clear-cut answer on whether the error expectation is equal to one (and hence are not able to say whether the Strategy 1 estimators are unbiased or not).

Under both the additive and the multiplicative error models for the estimator of the registration probability, the errors seemed independent of the 'true' probability. We noted, however, some alarming signs of dependencies between errors and registered flow. It seemed as if the true probability might be overestimated for low flows, while underestimated for high flows. Both error models state that the errors are site independent. Throughout our ANOVA tests, however, the null hypothesis of zero variance due to site was rejected. This objection to the models requires further investigation. The Strategy 2 estimators are unbiased if the estimated registration probabilities are unbiased for their true counterparts. We tested for this too, and obtained results that suggest that unbiasedness is in fact attained, no matter if the errors are additive or multiplicative.

For our five experimental sites, we estimated the flow, travel time, and average speed, then compared the estimates with the 'true' values. Under Strategy 0, as expected, the flow and travel time were clearly underestimated. Under both Strategy 1 and 2, on the other hand, the estimates of flow and travel time ended up reasonably close to the 'truth.' Under all strategies, the estimates of average speed came quite close to the 'truth.' The last result is far from concluding evidence. Still, we take it as a small hint that the present estimator of average speed is not overly sensitive to missing data.

## 4.6 Summary

Our suggested strategies for missing data adjustments are easy to implement. Still, the implementation is only of value if the adjustment estimators are likely to remove bias due to missing data. Whether they really get the job done is not that easy to establish. In fact, it is not even a matter of course that adjustments are at all necessary. Some of our empirical findings hint that the present unadjusted estimator of average speed may be surprisingly

resistant to bias due to missing data.

In our investigation of the estimators' theoretical properties, we made use of several models. We did not build complicated models, trying to get as close to reality as possible, but strived instead for simplicity. Despite this, the expressions for the estimators' expectations and variances turned out a bit messy. We were privileged to be able to supplement the theoretical analysis by use of some empirical data. Most of our model assumptions seemed to agree reasonably well with these data. In addition, the adjustment estimators seemed to produce better (less biased) estimates of the totals $t_y$ and $t_z$ than the current unadjusted estimators. We were not able to tell how their variances stood in comparison. None of the adjustment strategies showed its clear superiority to the other. Also, as already mentioned, it remains an open question whether the estimator of average speed really needs any missing data adjustments.

# Chapter 5

# Allocation problems

## 5.1 Introduction

In this chapter, we turn our attention to the sampling error of the survey estimates. More precisely, we are interested in evaluating the current allocation of the total sample over sampling stages. Our method of doing this is to estimate the components of the total variance of the estimator of $R$ arising from each sampling stage and then analyze their relative sizes. As means for a possible reallocation of the sample, we also present formulae for optimum sampling sizes. Possible nonsampling errors, which may bias and increase the variance of the survey estimators, are here ignored.

At present, in all but the first sampling stage, only one sampling unit per stratum is selected. Since units are drawn with replacement in stage one, the total variances of the estimators can still be estimated. The variance contributions from each sampling stage are, however, inseparable. We circumvent this problem by making use of a fictitious sampling design and some experimental data. In this way, the required variance component estimates are calculated for a domain of study.

## 5.2 Estimation of variance components

This section addresses the problem of estimating sampling stage variance components for the speed survey. By way of introduction, we present the

estimators which would have been applicable if the sampling sizes had exceeded one in each sampling stage. We continue by considering a situation where the sample sizes exceed one in the first and third sampling stages, but are equal to one in stage two. Under these circumstances, not all components can be estimated (more precisely, the variance contributions from the first and second sampling stages cannot be separated). We show how to make use of a fictitious design to enable estimation of all components. Finally, these formulae are used to calculate variance component estimates from a set of experimental data.

### 5.2.1 At least two observations in each sampling stage

Assume that two or more units have been selected from each stratum in each sampling stage in the speed survey. For this situation, estimators of the components of $V_p\left(\hat{t}_a\right)$ and $AV_p\left(\hat{R}\right)$ are available. Although in reality we do not face this favorable situation, an investigation of the estimators that ideally could have been used still serves as the natural starting point for our work.

**Estimation of the components of $V_p\left(\hat{t}_a\right)$**

In order to estimate the variance $V_p\left(\hat{t}_a\right)$ of $\hat{t}_a$, it is not necessary to estimate each of its components separately. From [29, Result 4.5.1], $V_p\left(\hat{t}_a\right)$ is unbiasedly estimated by

$$\hat{V}_{3\mathrm{st}}\left(\hat{t}_a\right) = \frac{1}{m_I\left(m_I-1\right)} \sum_{\nu=1}^{m_I} \left( \frac{\hat{t}_{\pi a i_\nu}}{p_{i_\nu}} - \hat{t}_a \right)^2 . \tag{5.1}$$

The computationally simple formula is due to the fact that sampling with replacement is used at the first sampling stage. We are, however, interested in estimating each variance component separately. By slight modification of [29, Result 4.4.3], unbiased estimators of the variance components $V_{\mathrm{TSU}}\left(\hat{t}_a\right)$, $V_{\mathrm{SSU}}\left(\hat{t}_a\right)$ and $V_{\mathrm{PSU}}\left(\hat{t}_a\right)$ are given, respectively, by

$$\hat{V}_{\mathrm{TSU}}\left(\hat{t}_a\right) = \frac{1}{m_I^2} \sum_{\nu=1}^{m_I} \frac{1}{p_{i_\nu}^2} \left( \frac{N_{I I i_\nu}}{n_{I I i_\nu}} \right)^2 \sum\nolimits_{s_{I I i_\nu}} \hat{V}_{a i_\nu q} \tag{5.2}$$

where

$$\hat{V}_{ai_\nu q} = N_{i_\nu q}^2 \frac{1-f_{i_\nu q}}{n_{i_\nu q}} S^2_{as_{i_\nu q}}; \quad f_{i_\nu q} = n_{i_\nu q}/N_{i_\nu q};$$
$$S^2_{as_{i_\nu q}} = \frac{1}{n_{i_\nu q}-1} \sum\nolimits_{s_{i_\nu q}} \left(a_k - \frac{t_{ai_\nu q}}{n_{i_\nu q}}\right)^2$$

for $q \in i_\nu$ and every $i_\nu$ that is a component of $os_I$,

$$\hat{V}_{\mathrm{SSU}}\left(\hat{t}_a\right) = \frac{1}{m_I^2} \sum_{\nu=1}^{m_I} \frac{\hat{V}_{ai_\nu}}{p_{i_\nu}^2} - \hat{V}_{\mathrm{TSU}}\left(\hat{t}_a\right) \tag{5.3}$$

where

$$\hat{V}_{ai_\nu} = N_{II i_\nu}^2 \frac{1-f_{II i_\nu}}{n_{II i_\nu}} S^2_{\hat{t}_a s_{II i_\nu}} + \frac{N_{II i_\nu}}{n_{II i_\nu}} \sum\nolimits_{s_{II i_\nu}} \hat{V}_{ai_\nu q};$$
$$f_{II i_\nu} = n_{II i_\nu}/N_{II i_\nu};$$
$$S^2_{\hat{t}_a s_{II i_\nu}} = \frac{1}{n_{II i_\nu}-1} \sum\nolimits_{s_{II i_\nu}} \left(\hat{t}_{\pi a i_\nu q} - \frac{\hat{t}_{\pi a i_\nu}}{n_{II i_\nu}}\right)^2$$

for every $i_\nu$ that is a component of $os_I$, and

$$\hat{V}_{\mathrm{PSU}}\left(\hat{t}_a\right) = \hat{V}_{3\mathrm{st}}\left(\hat{t}_a\right) - \hat{V}_{\mathrm{SSU}}\left(\hat{t}_a\right) - \hat{V}_{\mathrm{TSU}}\left(\hat{t}_a\right). \tag{5.4}$$

**Estimation of the components of $AV_p\left(\hat{R}\right)$**

Define a new variable $e = y - \hat{R}z$. From [27, Section 6.8.2], an estimator of the variance $AV_p\left(\hat{R}\right)$ of $\hat{R}$ is given by

$$\begin{aligned}
\hat{V}_{3\mathrm{st}}\left(\hat{R}\right) &= \frac{1}{\hat{t}_z^2} \hat{V}_{3\mathrm{st}}\left(\hat{t}_e\right) \\
&= \frac{1}{\hat{t}_z^2} \frac{1}{m_I\left(m_I-1\right)} \sum_{\nu=1}^{m_I} \left(\frac{\hat{t}_{\pi e i_\nu}}{p_{i_\nu}} - \hat{t}_e\right)^2 \\
&= \frac{1}{\hat{t}_z^2} \frac{1}{m_I\left(m_I-1\right)} \sum_{\nu=1}^{m_I} \left(\frac{\hat{t}_{\pi y i_\nu}}{p_{i_\nu}} - \hat{R}\frac{\hat{t}_{\pi z i_\nu}}{p_{i_\nu}}\right)^2
\end{aligned} \tag{5.5}$$

where the last equality holds since $\hat{t}_e = 0$.

Estimators of the variance components $AV_{\text{TSU}}\left(\hat{R}\right)$, $AV_{\text{SSU}}\left(\hat{R}\right)$ and $AV_{\text{PSU}}\left(\hat{R}\right)$ are given, respectively, by

$$\hat{V}_{\text{TSU}}\left(\hat{R}\right) = \frac{\hat{V}_{\text{TSU}}\left(\hat{t}_e\right)}{\hat{t}_z^2}; \quad \hat{V}_{\text{SSU}}\left(\hat{R}\right) = \frac{\hat{V}_{\text{SSU}}\left(\hat{t}_e\right)}{\hat{t}_z^2};$$
$$\hat{V}_{\text{PSU}}\left(\hat{R}\right) = \frac{\hat{V}_{\text{PSU}}\left(\hat{t}_e\right)}{\hat{t}_z^2} \tag{5.6}$$

where $\hat{V}_{\text{TSU}}\left(\hat{t}_e\right)$, $\hat{V}_{\text{SSU}}\left(\hat{t}_e\right)$ and $\hat{V}_{\text{PSU}}\left(\hat{t}_e\right)$ are obtained from Equations (5.2) to (5.4) by letting $a = e$.

### 5.2.2 One observation in the second stage

We now turn to a design with greater likeness to the real speed survey design than the one described in Section 5.2.1. It is still assumed that two or more units are selected from each stratum in the first and third sampling stages; the sample size is, however, now equal to one within stratum in the second stage.

A sample of size one in stage two does not prevent us from estimating the total variance of $\hat{t}_a$, or the last-stage component $V_{\text{TSU}}\left(\hat{t}_a\right)$, as in Section 5.2.1. Hence, we are also still able to estimate $V_{\text{PSU}}\left(\hat{t}_a\right) + V_{\text{SSU}}\left(\hat{t}_a\right)$. The small sample size does, however, preclude us from estimating $V_{\text{PSU}}\left(\hat{t}_a\right)$ and $V_{\text{SSU}}\left(\hat{t}_a\right)$ separately. The corresponding estimation problem holds for $\hat{R}$. We choose to tackle the problem as follows. First we note that for $V_{\text{SSU}}\left(\hat{t}_a\right) \neq 0$,

$$V_{\text{PSU}}\left(\hat{t}_a\right) + V_{\text{SSU}}\left(\hat{t}_a\right) = V_{\text{SSU}}\left(\hat{t}_a\right)\left(C\left(\hat{t}_a\right) + 1\right) \tag{5.7}$$

where $C\left(\hat{t}_a\right) = V_{\text{PSU}}\left(\hat{t}_a\right)/V_{\text{SSU}}\left(\hat{t}_a\right)$. In the same manner, for $AV_{\text{SSU}}\left(\hat{R}\right) \neq 0$,

$$AV_{\text{PSU}}\left(\hat{R}\right) + AV_{\text{SSU}}\left(\hat{R}\right) = AV_{\text{SSU}}\left(\hat{R}\right)\left(C\left(\hat{R}\right) + 1\right) \tag{5.8}$$

where $C\left(\hat{R}\right) = AV_{\text{PSU}}\left(\hat{R}\right)/AV_{\text{SSU}}\left(\hat{R}\right)$. Next, we formulate a fictitious sampling design, formulated so as to fulfil the criteria

- closely related to the one actually in use, and
- admitting separate estimation of each sampling stage component.

Finally, we derive estimators of (the closest equivalents to) the ratios $C\left(\hat{t}_a\right)$ and $C\left(\hat{R}\right)$ under the fictitious design and use those to estimate $V_{\mathrm{PSU}}\left(\hat{t}_a\right)$ and $V_{\mathrm{SSU}}\left(\hat{t}_a\right)$, $AV_{\mathrm{PSU}}\left(\hat{R}\right)$ and $AV_{\mathrm{SSU}}\left(\hat{R}\right)$, separately.

**Formulation of a fictitious sampling design**

Under our fictitious design, PSUs and TSUs are selected *without* replacement; SSUs *with* replacement, as follows.

**Stage I'** First, the ordered sample $os_I$ is drawn as in stage I in Section 2.4. The set of distinct PSUs which occur at least twice in $os_I$ then make up the stage I' set-sample $s_{I'}$ of PSUs of size $n_{I'}$.

**Stage II'** For every $i \in s_{I'}$, a sample of SSUs is drawn with simple random sampling with replacement (SIR). At every draw, $p_q = 1/N_{IIi}$ is the probability of selecting the $q$th SSU. Let $q_{\nu'}$ denote the SSU selected in the $\nu'$th draw, $\nu' = 1, ..., m_{II'i}$, where $m_{II'i}$ is the number of draws. The probability of selecting $q_{\nu'}$ is denoted $p_{q_{\nu'}}$. If the $q$th SSU is selected in the $\nu'$th draw, then $p_{q_{\nu'}} = p_q$. The vector of selected SSUs, $\left(q_{1'}, ..., q_{\nu'}, ..., q_{m_{II'i}}\right)$, is the resulting ordered sample $os_{II'i}$.

**Stage III'** For every $q_{\nu'}$ that is a component of $os_{II'i}$, an SI sample $s_{iq_{\nu'}}$ of TSUs of size $n_{iq_{\nu'}}$ is selected.

Note the resemblance to the actual design in Section 2.4. What we have done here is to transform the ordered sample in stage I into a set sample, and 'move' the with-replacement sampling one step down the stage hierarchy from the first to the second sampling stage. The main advantage of this procedure is that we gain access to more than one SSU drawing.

The sampling method specified for the first stage is not of standard type. Hence, for estimation purposes, the relevant first and second order inclusion probabilities $\pi_{I'i}$ and $\pi_{I'ij}$ need to be derived. As shown in Appendix D, the probability $\pi_{I'i}$ that PSU $i$ will be included in $s_{I'}; i \in U_I$, is given by

$$\pi_{I'i} = 1 - (1 - p_i)^{m_I} \left(1 + m_I \frac{p_i}{1 - p_i}\right) \tag{5.9}$$

and the probability $\pi_{I'ij}$ that both PSU $i$ and $j$ will be included in $s_{I'}; i, j \in U_I$, by

$$\pi_{I'ij} = 1 - (1-p_i)^{m_I}\left(1 + m_I\frac{p_i}{1-p_i}\right) - (1-p_j)^{m_I}\left(1 + m_I\frac{p_j}{1-p_j}\right) - (1-p_i-p_j)^{m_I-1} m_I\left[p_i + p_j + (m_I - 1)\frac{p_i p_j}{1-p_i-p_j}\right]. \quad (5.10)$$

**Use of the fictitious design**

For the sake of completeness, we start by presenting the estimators of $t_a$ and $R$ and their variances under the fictitious design, and continue by giving the estimators of the variances and their sampling stage components. The impatient reader is encouraged to proceed directly to subsection 'Estimation of the components of $V_p(\hat{t}_a)$ and $AV_p\left(\hat{R}\right)$', where our proposal for estimation of the components of $V_p(\hat{t}_a)$ and $AV_p\left(\hat{R}\right)$ by help of the fictitious design is summarized.

**The estimator $\hat{t}'_a$ of $t_a$** Under the fictitious design, from [29, Result 4.4.1], an unbiased estimator of $t_a$ is given by

$$\hat{t}'_a = \sum\nolimits_{s_{I'}} \frac{\hat{t}'_{\mathrm{pwr}ai}}{\pi_{I'i}} \quad (5.11)$$

where $\hat{t}'_{\mathrm{pwr}ai} = (N_{III}/m_{II'i})\sum_{\nu'=1}^{m_{II'i}} \hat{t}'_{\pi aiq_{\nu'}}$, $\hat{t}'_{\pi aiq_{\nu'}} = \left(N_{iq_{\nu'}}/n_{iq_{\nu'}}\right)\sum_{s_{iq_{\nu'}}} a_k$, and $N_{iq_{\nu'}}$ is the number of one-meter road sites in small area $q_{\nu'}$. (If the $q$th SSU was selected in the $\nu'$th draw, then $\hat{t}'_{\pi aiq_{\nu'}} = \hat{t}_{\pi aiq} = (N_{iq}/n_{iq})\sum_{s_{iq}} a_k$.)

The variance of $\hat{t}'_a$ is given by

$$V_{p'}(\hat{t}'_a) = \sum\sum\nolimits_{U_I} \Delta_{I'ij}\frac{t_{ai}}{\pi_{I'i}}\frac{t_{aj}}{\pi_{I'j}} + \sum\nolimits_{U_I}\frac{V'_{ai}}{\pi_{I'i}} \quad (5.12)$$

where $\Delta_{I'ij} = \pi_{I'ij} - \pi_{I'i}\pi_{I'j}$, $V'_{ai}$ is the variance of $\hat{t}'_{\mathrm{pwr}ai}$ with respect to the last two sampling stages:

$$V'_{ai} = V'_{aIIi} + \frac{N_{III}}{m_{II'i}}\sum_{q=1}^{N_{III}} V_{aiq} \quad (5.13)$$

where

$$V'_{aIIi} = \frac{N_{IIi}\,(N_{IIi}-1)}{m_{II'i}} S^2_{t_a U_i}.$$

Equivalently, the variance $V_{p'}\big(\hat{t}'_a\big)$ can be written as

$$V_{p'}\big(\hat{t}'_a\big) = V'_{\mathrm{PSU}}\big(\hat{t}'_a\big) + V'_{\mathrm{SSU}}\big(\hat{t}'_a\big) + V'_{\mathrm{TSU}}\big(\hat{t}'_a\big) \tag{5.14}$$

where

$$V'_{\mathrm{TSU}}\big(\hat{t}'_a\big) = \sum\nolimits_{U_I} \frac{1}{\pi_{I'i}} \frac{N_{IIi}}{m_{II'i}} \sum_{q=1}^{N_{IIi}} V_{aiq}, \tag{5.15}$$

$$V'_{\mathrm{SSU}}\big(\hat{t}'_a\big) = \sum\nolimits_{U_I} \frac{V'_{aIIi}}{\pi_{I'i}}, \tag{5.16}$$

and

$$V'_{\mathrm{PSU}}\big(\hat{t}'_a\big) = \sum\sum\nolimits_{U_I} \Delta_{I'ij} \frac{t_{ai}}{\pi_{I'i}} \frac{t_{aj}}{\pi_{I'j}}. \tag{5.17}$$

**The estimator $\hat{R}'$ of $R$** Under the fictitious design, from [29, Result 5.6.2], an approximately unbiased estimator of $R$ is given by

$$\hat{R}' = \frac{\hat{t}'_y}{\hat{t}'_z} = \frac{\sum_{s_{I'}} \frac{\hat{t}'_{\mathrm{pwr}yi}}{\pi_{I'i}}}{\sum_{s_{I'}} \frac{\hat{t}'_{\mathrm{pwr}zi}}{\pi_{I'i}}}. \tag{5.18}$$

The estimator $\hat{R}'$ has the approximate (Taylor) variance

$$\begin{aligned} AV_{p'}\Big(\hat{R}'\Big) &= \frac{1}{t_z^2} V_{p'}\big(\hat{t}'_E\big) \\ &= \frac{1}{t_z^2}\left(\sum\sum\nolimits_{U_I} \Delta_{I'ij} \frac{t_{Ei}}{\pi_{I'i}} \frac{t_{Ej}}{\pi_{I'j}} + \sum\nolimits_{U_I} \frac{V'_{Ei}}{\pi_{I'i}}\right) \\ &= \frac{1}{t_z^2}\left(\sum\sum\nolimits_{U_I} \Delta_{I'ij} \frac{t_{yi} - Rt_{zi}}{\pi_{I'i}} \frac{t_{yj} - Rt_{zj}}{\pi_{I'j}} + \sum\nolimits_{U_I} \frac{V'_{Ei}}{\pi_{I'i}}\right) \end{aligned} \tag{5.19}$$

where $V'_{Ei}$ is obtained from Equation (5.13) by letting the variable $a$ equal $E$.

The approximate variance $AV_{p'}\left(\hat{R}'\right)$ can equivalently be written as

$$AV_{p'}\left(\hat{R}'\right) = AV'_{\mathrm{PSU}}\left(\hat{R}'\right) + AV'_{\mathrm{SSU}}\left(\hat{R}'\right) + AV'_{\mathrm{TSU}}\left(\hat{R}'\right) \tag{5.20}$$

$$= \frac{V'_{\mathrm{PSU}}\left(\hat{t}'_E\right)}{t_z^2} + \frac{V'_{\mathrm{SSU}}\left(\hat{t}'_E\right)}{t_z^2} + \frac{V'_{\mathrm{TSU}}\left(\hat{t}'_E\right)}{t_z^2} \tag{5.21}$$

where $V'_{\mathrm{TSU}}\left(\hat{t}'_E\right)$, $V'_{\mathrm{SSU}}\left(\hat{t}'_E\right)$ and $V'_{\mathrm{PSU}}\left(\hat{t}'_E\right)$ are obtained from Equations (5.15) to (5.17) by letting $a = E$.

**Estimation of the components of** $V_{p'}\left(\hat{t}'_a\right)$ From [29, Result 4.4.1], under the fictitious design, an unbiased estimator of $V_{p'}\left(\hat{t}'_a\right)$ is given by

$$\hat{V}'_{3\mathrm{st}}\left(\hat{t}'_a\right) = \sum\sum\nolimits_{s_{I'}} \frac{\Delta_{I'ij}}{\pi_{I'ij}} \frac{\hat{t}'_{\mathrm{pwr}ai}}{\pi_{I'i}} \frac{\hat{t}'_{\mathrm{pwr}aj}}{\pi_{I'j}}. \tag{5.22}$$

By slight modification of [29, Result 4.4.3], unbiased estimators of $V'_{\mathrm{TSU}}\left(\hat{t}'_a\right)$, $V'_{\mathrm{SSU}}\left(\hat{t}'_a\right)$ and $V'_{\mathrm{PSU}}\left(\hat{t}'_a\right)$ are given, respectively, by

$$\hat{V}'_{\mathrm{TSU}}\left(\hat{t}'_a\right) = \sum\nolimits_{s'_I} \frac{1}{\pi^2_{I'i}} \left(\frac{N_{IIi}}{m_{II'i}}\right)^2 \sum\nolimits_{\nu'=1}^{m_{II'i}} \hat{V}'_{aiq_{\nu'}} \tag{5.23}$$

where

$$\hat{V}'_{aiq_{\nu'}} = N^2_{iq_{\nu'}} \frac{1 - f_{iq_{\nu'}}}{n_{iq_{\nu'}}} S^2_{as_{iq_{\nu'}}}; \quad f_{iq_{\nu'}} = n_{iq_{\nu'}}/N_{iq_{\nu'}};$$

$$S^2_{as_{iq_{\nu'}}} = \frac{1}{n_{iq_{\nu'}} - 1} \sum\nolimits_{s_{iq_{\nu'}}} \left(a_k - \frac{t_{a_{iq_{\nu'}}}}{n_{iq_{\nu'}}}\right)^2$$

for every $q_{\nu'}$ that is a component of $os_{II'i}$ and $i \in s'_I$,

$$\hat{V}'_{\mathrm{SSU}}\left(\hat{t}'_a\right) = \sum\nolimits_{s'_I} \frac{\hat{V}'_{ai}}{\pi^2_{I'i}} - \hat{V}'_{\mathrm{TSU}}\left(\hat{t}'_a\right) \tag{5.24}$$

where

$$\hat{V}'_{ai} = \frac{N^2_{IIi}}{m_{II'i}} S^2_{\hat{t}'_a os_{II'i}};$$

$$S^2_{\hat{t}'_a os_{II'i}} = \frac{1}{m_{II'i} - 1} \sum\nolimits_{\nu'=1}^{m_{II'i}} \left(\hat{t}_{\pi aiq_{\nu'}} - \frac{\sum_{\nu'=1}^{m_{II'i}} \hat{t}_{\pi aiq_{\nu'}}}{m_{II'i}}\right)^2$$

for $i \in s'_I$, and

$$\hat{V}'_{\mathrm{PSU}}\left(\hat{t}'_a\right) = \hat{V}'_{3\mathrm{st}}\left(\hat{t}'_a\right) - \hat{V}'_{\mathrm{SSU}}\left(\hat{t}'_a\right) - \hat{V}'_{\mathrm{TSU}}\left(\hat{t}'_a\right). \tag{5.25}$$

**Estimation of the components of $AV_{p'}\left(\hat{R}'\right)$** Define the variable $e' = y - \hat{R}'z$. Under the fictitious design, from [29, Result 5.6.2], an estimator of $AV_{p'}\left(\hat{R}'\right)$ is given by

$$\begin{aligned}\hat{V}'_{3\text{st}}\left(\hat{R}'\right) &= \frac{1}{\left(\hat{t}'_z\right)^2}\hat{V}'_{3\text{st}}\left(\hat{t}'_{e'}\right) \\ &= \frac{1}{\left(\hat{t}'_z\right)^2}\sum\sum_{s_{I'}}\frac{\Delta_{I'ij}}{\pi_{I'ij}}\frac{\hat{t}'_{\text{pwr}e'i}}{\pi_{I'i}}\frac{\hat{t}'_{\text{pwr}e'j}}{\pi_{I'j}}.\end{aligned} \tag{5.26}$$

Estimators of the variance components $AV'_{\text{TSU}}\left(\hat{R}'\right)$, $AV'_{\text{SSU}}\left(\hat{R}'\right)$ and $AV'_{\text{PSU}}\left(\hat{R}'\right)$ are given, respectively, by

$$\begin{aligned}\hat{V}'_{\text{TSU}}\left(\hat{R}'\right) = \frac{\hat{V}'_{\text{TSU}}\left(\hat{t}'_{e'}\right)}{\left(\hat{t}'_z\right)^2}; \quad \hat{V}'_{\text{SSU}}\left(\hat{R}'\right) = \frac{\hat{V}'_{\text{SSU}}\left(\hat{t}'_{e'}\right)}{\left(\hat{t}'_z\right)^2}; \\ \hat{V}'_{\text{PSU}}\left(\hat{R}'\right) = \frac{\hat{V}'_{\text{PSU}}\left(\hat{t}'_{e'}\right)}{\left(\hat{t}'_z\right)^2}\end{aligned} \tag{5.27}$$

where $\hat{V}'_{\text{TSU}}\left(\hat{t}'_{e'}\right)$, $\hat{V}'_{\text{SSU}}\left(\hat{t}'_{e'}\right)$ and $\hat{V}'_{\text{PSU}}\left(\hat{t}'_{e'}\right)$ are obtained from Equations (5.23) to (5.25) by letting $a = e'$.

**Estimation of the components of $V_p\left(\hat{t}_a\right)$ and $AV_p\left(\hat{R}\right)$** As estimators of $C\left(\hat{t}_a\right)$ and $C\left(\hat{R}\right)$, we suggest using the estimators of the corresponding population entities under the fictitious design. That is, estimate $C\left(\hat{t}_a\right)$ by

$$\hat{C}'\left(\hat{t}_a\right) = \frac{\hat{V}'_{\text{PSU}}\left(\hat{t}'_a\right)}{\hat{V}'_{\text{SSU}}\left(\hat{t}'_a\right)} \tag{5.28}$$

and $C\left(\hat{R}\right)$ by

$$\hat{C}'\left(\hat{R}\right) = \frac{\hat{V}'_{\text{PSU}}\left(\hat{R}\right)}{\hat{V}'_{\text{SSU}}\left(\hat{R}\right)} = \frac{\hat{V}'_{\text{PSU}}\left(\hat{t}'_{e'}\right)}{\hat{V}'_{\text{SSU}}\left(\hat{t}'_{e'}\right)}. \tag{5.29}$$

The resulting estimators of $V_{\text{SSU}}\left(\hat{t}_a\right)$ and $V_{\text{PSU}}\left(\hat{t}_a\right)$ are

$$\hat{V}_{\text{SSU}}\left(\hat{t}_a\right) = \frac{\hat{V}_{3\text{st}}\left(\hat{t}_a\right) - \hat{V}_{\text{TSU}}\left(\hat{t}_a\right)}{\hat{C}'\left(\hat{t}_a\right) + 1} \tag{5.30}$$

and

$$\hat{V}_{\mathrm{PSU}}\left(\hat{t}_a\right) = \hat{V}_{3\mathrm{st}}\left(\hat{t}_a\right) - \hat{V}_{\mathrm{SSU}}\left(\hat{t}_a\right) - \hat{V}_{\mathrm{TSU}}\left(\hat{t}_a\right) \tag{5.31}$$

respectively, where $\hat{V}_{\mathrm{TSU}}\left(\hat{t}_a\right)$ is given by Equation (5.2) and $\hat{V}_{3\mathrm{st}}\left(\hat{t}_a\right)$ by Equation (5.1). In the same manner, our suggested estimators of $AV_{\mathrm{SSU}}\left(\hat{R}\right)$ and $AV_{\mathrm{PSU}}\left(\hat{R}\right)$ are given by

$$\hat{V}_{\mathrm{SSU}}\left(\hat{R}\right) = \frac{\hat{V}_{3\mathrm{st}}\left(\hat{R}\right) - \hat{V}_{\mathrm{TSU}}\left(\hat{R}\right)}{\hat{C}'\left(\hat{R}\right) + 1} \tag{5.32}$$

and

$$\hat{V}_{\mathrm{PSU}}\left(\hat{t}_a\right) = \hat{V}_{3\mathrm{st}}\left(\hat{R}\right) - \hat{V}_{\mathrm{SSU}}\left(\hat{R}\right) - \hat{V}_{\mathrm{TSU}}\left(\hat{R}\right) \tag{5.33}$$

respectively (with $\hat{V}_{\mathrm{TSU}}\left(\hat{R}\right)$ as in Equation (5.6) and $\hat{V}_{3\mathrm{st}}\left(\hat{R}\right)$ as in Equation (5.5)).

There is no guarantee for $\hat{V}'_{\mathrm{PSU}}\left(\hat{t}'_a\right)$ or $\hat{V}'_{\mathrm{SSU}}\left(\hat{t}'_a\right)$ to take on positive values. In case any of them is negative, it does not make sense to calculate $\hat{C}'\left(\hat{t}_a\right)$, and the variance component estimators in Equations (5.30) and (5.31) must be abandoned. Correspondingly, the estimators in Equations (5.32) and (5.33) should not be used if $\hat{V}'_{\mathrm{PSU}}\left(\hat{R}\right)$ or $\hat{V}'_{\mathrm{SSU}}\left(\hat{R}\right)$ is negative.

### 5.2.3 Calculation of variance component estimates from real data

In the speed survey, the sample size within stratum is one in both the second and the third sampling stage. To render variance component estimation possible, in accordance with Section 5.2.2, within the frame of the main 2001 survey, some experimental data were collected. Here, we present the design and outcome of this experiment.

#### Data collection and processing

The collection of experimental data was restricted to one PSU stratum: the South-Eastern SNRA region and the size class 'Large major population center of a municipality.' From this set of population centers, as part of the

| Development type | Road type | Number of pairs |
|---|---|---|
| City | M70 | 10 |
| City | M50 | 0 |
| City | Other | 6 |
| Industrial | M70 | 8 |
| Industrial | M50 | 3 |
| Industrial | Other | 10 |
| Residential | M70 | 5 |
| Residential | M50 | 2 |
| Residential | Other | 8 |
| Other | M70 | 7 |
| Other | M50 | 1 |
| Other | Other | 9 |

Table 5.1: The number of observation pairs, for each combination of SSU and TSU stratum.

main survey, a sample of 75 road sites was selected. Our plan was to double this number by selecting an additional road site from each chosen small area (within PSU drawing). For various reasons (such as missing data problems), for six chosen small areas, data were obtained from only one site. We decided to exclude these small areas from the experiment, which left us with $69 \times 2$ observations on traffic flow and travel time. These 69 observation pairs are distributed among the four SSU strata and three TSU strata (see Section 2.4.3) as shown in Table 5.1. We see that, throughout, we have very few observations on M50 roads. Therefore, this stratum is left out of further consideration.

For the strata thus comprised by the experiment, the sample sizes according to the fictitious design are: $n_{I'} = 3$ in the first stage; $m_{II'i} = 2$, 4 and 4, respectively, in the second stage; and $n_{iq_{\nu'}} = 2$ in the final sampling stage.

| Development type | Road type | $\hat{V}_{3\text{st}}\left(\hat{R}\right)$ | $\hat{V}_{\text{TSU}}\left(\hat{R}\right)$ | $\hat{V}_{\text{SSU}}\left(\hat{R}\right)$ | $\hat{V}_{\text{PSU}}\left(\hat{R}\right)$ |
|---|---|---|---|---|---|
| City | M70 | 5.6367 | 7.5538 | – | – |
| City | Other | 10.8389 | 11.2937 | – | – |
| Industrial | M70 | 1.2929 | 1.1497 | * | * |
| Industrial | Other | 2.1851 | 5.9771 | – | – |
| Residential | M70 | 9.4960 | 3.0601 | 6.4359 | * |
| Residential | Other | 9.4112 | 13.1166 | – | – |
| Other | M70 | 5.9345 | 1.3650 | 4.5695 | * |
| Other | Other | 6.6365 | 5.6771 | 0.9594 | * |

Table 5.2: Estimates (in km/h) of the approximate variance of $\hat{R}$ and its components, for various combinations of SSU and TSU strata.

**Results**

For each combination of SSU and TSU stratum, variances and variance components are estimated in accordance with Section 5.2.2. The estimates for $\hat{R}$ are presented in Table 5.2; the corresponding estimates for $\hat{t}_y$ and $\hat{t}_z$ in Appendix F. To simplify the estimation task, a few shortcuts are taken. As mentioned in Section 2.4.3, SSUs are really selected with pps rather than SI sampling within stratum Residential areas. This exception is disregarded, and the estimates for residential areas calculated as if SI sampling was used. Also, our estimates refer to a single (arbitrary) 24-hour period within the time period of study, rather than the whole period.

In Table 5.2 and Appendix F, there are many hyphens and asterisks replacing numbers. The *hyphens* are used for cases where the difference $\hat{V}_{3\text{st}} - \hat{V}_{\text{TSU}}$ is negative; then, we do not attempt to estimate $\hat{V}_{\text{SSU}}$ or $\hat{V}_{\text{PSU}}$. The *asterisks* are used when $\hat{V}_{3\text{st}} - \hat{V}_{\text{TSU}}$ is positive but $\hat{V}'_{\text{PSU}}$ or $\hat{V}'_{\text{SSU}}$ is negative. If $\hat{V}'_{\text{PSU}}$ is negative, $\hat{V}_{\text{SSU}}$ is calculated as $\hat{V}_{3\text{st}} - \hat{V}_{\text{TSU}}$ whereas $\hat{V}_{\text{PSU}}$ is marked with an asterisk. Correspondingly, if $\hat{V}'_{\text{SSU}}$ is negative, $\hat{V}_{\text{PSU}}$ is calculated as $\hat{V}_{3\text{st}} - \hat{V}_{\text{TSU}}$ whereas $\hat{V}_{\text{SSU}}$ is marked with an asterisk. If $\hat{V}'_{\text{PSU}}$ *and* $\hat{V}'_{\text{SSU}}$ are negative (occurs only once), $\hat{V}_{\text{PSU}}$ and $\hat{V}_{\text{SSU}}$ are both marked with asterisks. We take the many negative variance estimates as a manifestation of the uncertainty in the estimates due to small sample sizes. A much larger

experiment is needed to reduce this uncertainty.

Since $R$ is the most important parameter, we focus on Table 5.2. Typically, $\hat{V}_{\mathrm{TSU}}\left(\hat{R}\right)$ is nearly as large as (or even larger than) $\hat{V}_{\mathrm{3st}}\left(\hat{R}\right)$. Thus, our main conclusion is that $AV_{\mathrm{TSU}}\left(\hat{R}\right)$ seems to predominate among the components of the variance of $\hat{R}$.

## 5.3 Optimum allocation over sampling stages

In this section, formulae for determination of optimum sampling sizes in each sampling stage are presented. As in most parts of this thesis, the stratification in each sampling stage is ignored. Extension of the theory presented here to the stratified case is, nevertheless, straightforward.

### 5.3.1 Conditions and general solution

The conditions for allocation are the following. The variance of $\hat{t}_a$ and the approximate variance of $\hat{R}$ both fit into the general variance expression

$$V = \frac{A_1}{x_1} + \sum_{i=1}^{N_I} \frac{A_{IIi}}{x_{IIi}} + \sum_{i=1}^{N_I} \sum\nolimits_{U_{IIi}} \frac{A_{iq}}{x_{iq}} \tag{5.34}$$

where the $x$'s are given by

$$x_1 = m_I \tag{5.35}$$

$$x_{IIi} = m_I n_{IIi}; \quad i \in U_I \tag{5.36}$$

$$x_{iq} = m_I n_{IIi} n_{iq}; \quad q \in U_{IIi}; i \in U_I, \tag{5.37}$$

and the $A$'s are constants with respect to the $x$'s.

Let $C_I$ denote the cost of conducting one PSU drawing. Within PSU $i \in U_I$, the listing cost per SSU, and the cost of selecting one SSU, are denoted $C^l_{IIi}$ and $C^s_{IIi}$, respectively. In the same manner, within selected SSU $q \in U_{IIi}$ and PSU $i \in U_I$, the listing cost per TSU and the cost of observing a selected TSU are denoted $C^l_{iq}$ and $C^s_{iq}$, respectively. The variable

costs of the survey can now be described by the linear function

$$VC = m_I C_I + \sum_{\nu=1}^{m_I} N_{III_\nu} C^l_{III_\nu} + \sum_{\nu=1}^{m_I} n_{III_\nu} C^s_{III_\nu} + \sum_{\nu=1}^{m_I} \sum\nolimits_{s_{III_\nu}} N_{i_\nu q} C^l_{i_\nu q} + \sum_{\nu=1}^{m_I} \sum\nolimits_{s_{III_\nu}} n_{i_\nu q} C^s_{i_\nu q}. \tag{5.38}$$

If PSU $i \in U_I$ was selected in the $\nu$th draw, then $C^l_{III_\nu} = C^l_{III}$, $C^s_{III_\nu} = C^s_{III}$, $C^l_{i_\nu q} = C^l_{iq}$ and $C^s_{i_\nu q} = C^s_{iq}$.

In practice, the listing and selection costs $C^l_{III}$ and $C^s_{III}$ may be approximately constant over PSUs. The listing costs $C^l_{iq}$ are, however, certain to vary substantially between SSUs. For each chosen small area, as described in Section 2.4.2, the list of PSUs is prepared from a city map. This procedure can be very time-consuming in areas with complex road networks, whilst quite fast in areas containing only a few roads. (The differences in listing costs are mitigated, but hardly removed, by the stratification of SSUs.)

Since $VC$ is a random variable (it depends on the random samples $os_I$ and $s_{III_\nu}$), we do not want to base an optimization problem directly upon it. Instead, we follow established practice and use its expectation. Under the sampling design described in Section 2.4, the expected value of $VC$ is given by

$$EVC = m_I C_I + m_I \sum_{i=1}^{N_I} p_i N_{III} C^l_{III} + m_I \sum_{i=1}^{N_I} p_i n_{III} C^s_{III} + m_I \sum_{i=1}^{N_I} p_i \sum\nolimits_{U_{III}} \frac{n_{III}}{N_{III}} N_{iq} C^l_{iq} + m_I \sum_{i=1}^{N_I} p_i \sum\nolimits_{U_{III}} \frac{n_{III}}{N_{III}} n_{iq} C^s_{iq}. \tag{5.39}$$

Equivalently, the expected variable cost can be expressed as

$$EVC = x_1 a_1 + \sum_{i=1}^{N_I} x_{III} a_{III} + \sum_{i=1}^{N_I} \sum\nolimits_{U_{III}} x_{iq} a_{iq} \tag{5.40}$$

with the $x$'s as in Equations (5.35) to (5.40), and the $a$'s given by

$$a_1 = C_I + \sum_{i=1}^{N_I} p_i N_{IIi} C^l_{IIi} \tag{5.41}$$

$$a_{IIi} = p_i \left( C^s_{IIi} + \sum\nolimits_{U_{IIi}} \frac{N_{iq}}{N_{IIi}} C^l_{iq} \right); \quad i \in U_I \tag{5.42}$$

$$a_{iq} = p_i \sum\nolimits_{U_{IIi}} \frac{C^s_{iq}}{N_{IIi}}; \quad q \in U_{IIi}; i \in U_I. \tag{5.43}$$

The allocation problem has two possible formulations:

- Minimize the variance $V$ in Equation (5.34) with respect to $x_1$, $x_{IIi}$ and $x_{iq}$ under the expected cost constraint

  $$EVC = C_0$$

  where $EVC$ is given by Equation (5.40), or

- minimize the expected cost $EVC$ in Equation (5.40) with respect to $x_1$, $x_{IIi}$ and $x_{iq}$ under the variance constraint

  $$V = V_0$$

  where $V$ is given by Equation (5.34).

We restrict our attention here to the second case. If the $A$'s are all greater than zero, from [7, p. 15], this minimization problem has the analytical solution

$$x_1 = K \sqrt{\frac{A_1}{a_1}} \tag{5.44}$$

$$x_{IIi} = K \sqrt{\frac{A_{IIi}}{a_{IIi}}}; \quad i \in U_I \tag{5.45}$$

$$x_{iq} = K \sqrt{\frac{A_{iq}}{a_{iq}}}; \quad q \in U_{IIi}; i \in U_I \tag{5.46}$$

where

$$K = \frac{1}{V_0} \left( \sqrt{a_1 A_1} + \sum_{i=1}^{N_I} \sqrt{a_{IIi} A_{IIi}} + \sum_{i=1}^{N_I} \sum\nolimits_{U_{IIi}} \sqrt{a_{iq} A_{iq}} \right).$$

The resulting optimum sampling sizes are given by

$$m_I = K\sqrt{\frac{A_1}{a_1}} \tag{5.47}$$

$$n_{IIi} = \sqrt{\frac{A_{IIi}a_1}{a_{IIi}A_1}}; \quad i \in U_I \tag{5.48}$$

$$n_{iq} = \sqrt{\frac{A_{iq}a_{IIi}}{a_{iq}A_{IIi}}}; \quad q \in U_{IIi}; i \in U_I. \tag{5.49}$$

(For a solution of the minimization problem if the $A$'s are *not* all greater than zero, see [7, p. 15].)

### 5.3.2 Solutions for $t_a$ and $R$

The general variance expression in Equation (5.34) turns into the variance of $\hat{t}_a$ if the $A$'s are defined as

$$A_1 = \sum_{i=1}^{N_I} p_i \left(\frac{t_{ai}}{p_i} - t_a\right)^2 - \sum_{i=1}^{N_I} \frac{N_{IIi}}{p_i} S^2_{t_a U_i} \tag{5.50}$$

$$A_{IIi} = \frac{N_{IIi}}{p_i}\left(N_{IIi}S^2_{t_a U_i} - \sum\nolimits_{U_{IIi}} N_{iq}S^2_{aU_{iq}}\right); \quad i \in U_I \tag{5.51}$$

$$A_{iq} = \frac{N_{IIi}}{p_i} N^2_{iq} S^2_{aU_{iq}}; \quad q \in U_{IIi}; i \in U_I. \tag{5.52}$$

Insertion in Equations (5.47) to (5.49) gives the solution for $t_a$ (if the $A$'s are all greater than zero). Similarly, in order to transform Equation (5.34) into the approximate variance of $\hat{R}$, define the $A$'s as

$$A_1 = \frac{1}{t_z^2}\left[\sum_{i=1}^{N_I} p_i \left(\frac{t_{Ei}}{p_i} - t_E\right)^2 - \sum_{i=1}^{N_I} \frac{N_{IIi}}{p_i} S^2_{t_E U_i}\right] \tag{5.53}$$

$$A_{IIi} = \frac{1}{t_z^2}\frac{N_{IIi}}{p_i}\left(N_{IIi}S^2_{t_E U_i} - \sum\nolimits_{U_{IIi}} N_{iq}S^2_{EU_{iq}}\right); \quad i \in U_I \tag{5.54}$$

$$A_{iq} = \frac{1}{t_z^2}\frac{1}{p_i} N_{IIi} N^2_{iq} S^2_{EU_{iq}}; \quad q \in U_{IIi}; i \in U_I. \tag{5.55}$$

If the $A$'s are all greater than zero, the solution for $R$ is obtained from Equations (5.47) to (5.49).

### 5.3.3 On use of the solutions

The formulae for optimum sample sizes in Equations (5.47) to (5.49) are not very complicated. The real problems start when they are to be used. Since $R$ is the parameter of main interest, it ought to govern the allocation. But let us take a second glance at Equations (5.53) to (5.55). Among several unknown population entities, the $A$'s include the population variances of $t_E$, $S^2_{t_{EU_i}}$, for all PSUs, as well as the population variances of $E$, $S^2_{EU_{iq}}$, for all SSUs. Use of Equations (5.47) to (5.49) thus require estimates of all these variances. The survey data will not suffice for calculating the demanded estimates, but large amounts of additional data need to be collected.

## 5.4 Summary

Is the current allocation of $s$ over sampling stages the most efficient, or is there room for improvement? In order to answer this question, we pursued the theoretical work of Sections 5.2.1 and 5.2.2, and conducted the experiment reported in Section 5.2.3. Our efforts resulted in the variance component estimates for $\hat{R}$ presented in Table 5.2. From this table, it looks as if the final sampling stage contributes the most to the total variance of $\hat{R}$. We conclude that, for unchanged size of $s$, the precision of $\hat{R}$ would probably improve if the sample sizes in stage three were increased, and the number of drawings in stage one decreased correspondingly (there is no room for decreasing the sample sizes in stage two, since they are already at minimum).

Our advice on reallocation of the total sample is, by necessity, quite vague. The theoretical tools for choosing the sampling sizes in an optimum manner are provided in Section 5.3. However, the formulae presented there involve many unknown population quantities and hence may be hard to use in practice.

# Chapter 6

# Survey models

## 6.1 Introduction

In previous chapters, the isolated impact of some different sources of error on the survey estimators has been investigated. Here, we adopt a comprehensive view towards those errors by formulating survey models for the estimators of $t_a$ and $R$.

Generally speaking, a survey model (mixed error model, total error model) is a model that accommodates several sources of error and possible interrelationships among them. Knowledge of the relative importance of different sources of errors can be used as an aid in making decisions on how available survey resources should be allocated. Since attempts to reduce or control errors of one type may have adverse effects on some other component of the total error, knowledge of interrelationships among different sources of error is important. Research on survey models date from the 1940s and was initially dominated by work performed at the U.S. Census Bureau. A review of model development before 1970 is given in [11]; for later development, see [12]. For examples of some general models, see [24, Chapter 12] and [29, Chapter 16].

## 6.2 The estimators $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$

Due to the frame errors discussed in Chapter 3 and the missing data problem treated in Chapter 4, instead of the prototype estimator $\hat{t}_a$, we can only

observe

$$\hat{t}_{F\hat{a}^{(c)}} = \frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{II i_\nu}}{n_{II i_\nu}}\sum\nolimits_{s_{II i_\nu}} \hat{t}_{F\pi\hat{a}^{(c)} i_\nu q} \tag{6.1}$$

where

$$\hat{t}_{F\pi\hat{a}^{(c)} i_\nu q} = \frac{N_{Fi_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{Fi_\nu q}} \hat{a}_k^{(c)}.$$

Also, we do not have access to $\hat{R}$ but only to

$$\hat{R}_F^{(c)} = \frac{\hat{t}_{F\hat{y}^{(c)}}}{\hat{t}_{F\hat{z}^{(c)}}}. \tag{6.2}$$

The statistical properties of $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$ are investigated in Section 6.3.

## 6.3 Survey models for $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$

In order to derive the expectations and variances of $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$, we need to take several sources of randomness into account: three stages of sampling as well as the procedures that generate $N_{Fiq}$ and $\hat{a}_k^{(c)}$. An error model $\mathrm{m}_1$ for $N_{Fiq}$ was formulated in Section 3.3; an error model $\mathrm{m}_2$ for $\hat{a}_k^{(c)}$ in Section 4.4.2. We have no reason to believe that the two types of error are somehow related, but assume that the mechanism that generates $\hat{a}_k^{(c)}$ is unconfounded with the one that generates $N_{Fiq}$. That is, the probability of a certain outcome of $\hat{a}_k^{(c)}$ is assumed to be unaffected by $N_{Fiq}$.

In the following, expectations and variances are indicated by subscript $p\mathrm{m}_1\mathrm{m}_2$ if taken with respect jointly to the sampling design $p$ and model $\mathrm{m}_1$ and $\mathrm{m}_2$. Conditional expectations and variances are indicated by '|'; for instance, $E_{\mathrm{m}_2|\mathrm{m}_1}$ denotes expectation with respect to model $\mathrm{m}_2$, conditional on model $\mathrm{m}_1$.

### 6.3.1 Properties of $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$

Consider first the estimator $\hat{t}_{F\hat{a}^{(c)}}$ of $t_a$. By use of conditioning, the expected value of $\hat{t}_{F\hat{a}^{(c)}}$ can be written as

$$E_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right) = E_p E_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\hat{a}^{(c)}}\right) \tag{6.3}$$

and its total variance as

$$\begin{aligned} V_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= V_p E_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + E_p V_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &\quad + E_p E_{\mathrm{m}_1|p} V_{\mathrm{m}_2|p\mathrm{m}_1}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &= V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big)\,. \end{aligned} \tag{6.4}$$

The $V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ term is due to the sample selection: in a total enumeration of all road sites, $V_1 = 0$. The $V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ term arises from variability in $\hat{t}_{F\pi\gamma\left(\hat{a}^{(c)}\right)i_\nu q}$ due to different realisations of $N_{Fi_\nu q}$: if all $\sigma^2_{iq} = 0$, then $V_2 = 0$. The $V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ term, finally, arises from variability in $\hat{a}_k$ for individual road sites: if all $\delta\big(\hat{a}^{(c)}\big)_k = 0$, then $V_3 = 0$. If the speed survey did not suffer from any frame errors, $V_3$ would correspond to the 'measurement variance' in [24, Equation (12.9)] or the 'simple measurement variance' in [29, Equation (16.4.5)].

By additional use of conditioning, the $V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ term can be written as the sum of three components, representing the variation contribution due to each sampling stage:

$$\begin{aligned} V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= V_I E_{II} E_{III} E_{\mathrm{m}_1\mathrm{m}_2|p}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + E_I V_{II} E_{III} E_{\mathrm{m}_1\mathrm{m}_2|p}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &\quad + E_I E_{II} V_{III} E_{\mathrm{m}_1\mathrm{m}_2|p}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &= V_{1,\mathrm{PSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,\mathrm{SSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,\mathrm{TSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \end{aligned} \tag{6.5}$$

where $V_{1,\mathrm{PSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ is due to the initial sampling of PSUs, $V_{1,\mathrm{SSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ to the second-stage sampling of SSUs, and $V_{1,\mathrm{TSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big)$ to the final-stage sampling of TSUs.

We are now ready for the following theorem:

**Theorem 6.3.1** *Jointly under the sampling design p in Section 2.4, the error model $m_1$ in Section 3.3, and error model $m_2$ in Section 4.4.2, the expected value of $\hat{t}_{\hat{a}^{(c)}}$ is given by*

$$E_{pm_1m_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \sum_{i=1}^{N_I} E_{pm_1m_2}\big(\hat{t}_{F\pi\hat{a}^{(c)}i}\big) \tag{6.6}$$

*where $E_{pm_1m_2}\big(\hat{t}_{F\pi\hat{a}^{(c)}i}\big) = \sum_{U_{II}i}\big(\mu_{iq}/N_{iq}\big)\, t_{\gamma\left(\hat{a}^{(c)}\right)iq}$. The variance of $\hat{t}_{F\hat{a}^{(c)}}$ is*

*given by*

$$\begin{aligned} V_{pm_1m_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &= V_{1,PSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,SSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,TSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &\quad + V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) \end{aligned} \tag{6.7}$$

*where*

$$V_{1,PSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left( \frac{E_{pm_1m_2}\big(\hat{t}_{F\pi\hat{a}^{(c)}i}\big)}{p_i} - E_{pm_1m_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \right)^2 \tag{6.8}$$

$$\begin{aligned} V_{1,SSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{1}{p_i} N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \sum\nolimits_{U_{IIi}} \Bigg( & \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq} \\ & - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq} \Bigg)^2 \end{aligned} \tag{6.9}$$

$$V_{1,TSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left( \frac{\mu_{iq}}{N_{iq}} \right)^2 V_{\gamma\left(\hat{a}^{(c)}\right)iq} \tag{6.10}$$

$$V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\sigma_{iq}^2}{N_{iq}^2} \left( V_{\gamma\left(\hat{a}^{(c)}\right)iq} + t^2_{\gamma\left(\hat{a}^{(c)}\right)_{iq}} \right) \tag{6.11}$$

*and*

$$V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}^2 + \sigma_{iq}^2}{N_{iq} n_{iq}} \sum\nolimits_{U_{iq}} \delta\big(\hat{a}^{(c)}\big)_k \,. \tag{6.12}$$

The proof of Theorem 6.3.1 is given in Appendix B.1.4.

Having come this far, it is an easy matter to derive the statistical properties of $\hat{R}_F^{(c)}$. The following theorem is proven by a slight generalization of the results in [27, Section 6.8.2.]:

**Theorem 6.3.2** *Jointly under the sampling design p in Section 2.4, the error model $m_1$ in Section 3.3, and error model $m_2$ in Section 4.4.2, the estimator $\hat{R}_F^{(c)}$ is approximately unbiased for*

$$R_F^{(c)} = \frac{E_{pm_1m_2}\big(\hat{t}_{F\hat{y}^{(c)}}\big)}{E_{pm_1m_2}\big(\hat{t}_{F\hat{z}^{(c)}}\big)}. \tag{6.13}$$

*The approximate variance of* $\hat{R}_F^{(c)}$ *is given by*

$$\begin{aligned} AV_{pm_1m_2}\left(\hat{R}_F^{(c)}\right) &= V_1\left(\hat{R}_F^{(c)}\right) + V_2\left(\hat{R}_F^{(c)}\right) + V_3\left(\hat{R}_F^{(c)}\right) \\ &= \frac{V_1\left(\hat{t}_{F\pi\hat{E}^{(c)}}\right)}{t_z^2} + \frac{V_2\left(\hat{t}_{F\pi\hat{E}^{(c)}}\right)}{t_z^2} + \frac{V_3\left(\hat{t}_{F\pi\hat{E}^{(c)}}\right)}{t_z^2} \end{aligned} \tag{6.14}$$

*where the variances of* $\hat{t}_{F\pi\hat{E}^{(c)}}$ *are obtained from the corresponding variances of* $\hat{t}_{F\hat{a}^{(c)}}$ *in Theorem 6.3.1 by replacing* $\hat{a}^{(c)}$ *with* $\hat{E}_F^{(c)} = \hat{y}^{(c)} - R_F^{(c)}\hat{z}^{(c)}$.

For both $\hat{t}_{F\hat{a}^{(c)}}$ and $R_F^{(c)}$, it is tempting to call $V_1$ the sampling variance, $V_2$ the frame errors variance, and $V_3$ the variance due to missing data. These interpretations are, however, somewhat misleading. The errors due to missing data and frame imperfections are entwined closely together, and both have the potential to influence all components of the model. This follows since $V_1$ includes expected values of both $N_{Fiq}$ and $\hat{a}_k^{(c)}$; $V_2$ expected values of $\hat{a}_k^{(c)}$; and $V_3$ both expected values and variances of $N_{Fiq}$.

### 6.3.2 Simplifications and connections with earlier work

The expectations and variances of $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$, as presented in Theorems 6.3.1 and 6.3.2, are quite complicated and thus hard to evaluate. We now try to simplify the expressions by making two assumptions:

- The frame road lengths are unbiased for the true road lengths.
- The missing data adjusted estimators $\hat{a}_k^{(1)}$ and $\hat{a}_k^{(2)}$ are unbiased for $a_k$.

The first assumption, which receives some confirmation in our frame errors investigation in Section 3.5, implies that we can replace $\mu_{iq}$ by $N_{iq}$. Our results in Section 4.5 give some support for the second assumption, which means that we can substitute $\gamma\left(\hat{a}^{(c)}\right)_k$ by $a_k$ if $c = 1$ or 2.

Under the two assumptions, if estimation strategy $c = 1$ or 2 is employed to adjust for missing data, the following holds.

The estimator $\hat{t}_{F\hat{a}^{(c)}}$ is unbiased for $t_a$, $\hat{R}_F^{(c)}$ is approximately unbiased for $R$, and $V_1$ equals the sampling variance of the estimator in question (that is, the variance of the corresponding prototype estimator). Then, since

the components $V_{1,\mathrm{PSU}}$, $V_{1,\mathrm{SSU}}$ and $V_{1,\mathrm{TSU}}$ now solely represent the variation contribution due to each sampling stage, our investigation in Chapter 5 of their relative sizes applies. The investigation suggests that, for $\hat{R}_F^{(c)}$, the $V_{1,\mathrm{TSU}}$ component prevails among the three.

Furthermore, for $\hat{t}_{F\hat{a}^{(c)}}$,

$$V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\sigma_{iq}^2}{N_{iq}^2}\left(V_{aiq}+t_{aiq}^2\right) \tag{6.15}$$

$$V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\left(\frac{N_{iq}}{n_{iq}}+\frac{\sigma_{iq}^2}{N_{iq}n_{iq}}\right) \times\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\,. \tag{6.16}$$

(The terms $V_2$ and $V_3$ for $\hat{R}_F^{(c)}$ are simplified correspondingly.) We see that all errors no longer affect all components. Besides $V_1$ being equal to the sampling variance, $V_2$ does not include expected values of $\hat{a}_k^{(c)}$ (which may differ from the true values) but the $a_k$'s themselves. Hence, the $V_2$ term now truly deserves to be called the frame errors variance. The $V_3$ term, however, is still not a pure missing data variance.

Assume further, as in the multiplicative error model for $N_{Fiq}$ in Chapter 3, that $\sigma_{iq}^2 = \tau^2 N_{iq}^2$ (where $\tau^2$ is constant as function of $N_{iq}$). Then,

- from Corollary 3.4.3, the $V_2$ term for $\hat{R}_F^{(c)}$ can be written as

$$V_2\Big(\hat{R}_F^{(c)}\Big) = \tau^2 A V_p\Big(\hat{R}\Big)\,. \tag{6.17}$$

  According to Table 3.3, a 95 percent upper-bounded confidence interval for $\tau^2$ is given by $[0, 0.00848]$.

- the $V_3$ term for $\hat{R}_F^{(c)}$ is given by

$$V_3\Big(\hat{R}_F^{(c)}\Big) = \big(1+\tau^2\big)\frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}} \times\sum\nolimits_{U_{iq}}\delta\Big(\hat{E}^{(c)}\Big)_k \tag{6.18}$$

  where $\hat{E}^{(c)} = \hat{y}^{(c)} - R\hat{z}^{(c)}$. In Section 4.4.2, the variance $\delta\Big(\hat{E}_F^{(c)}\Big)_k$ is derived for the adjustment strategies $c = 1$ and 2 and various special cases of those. (Unfortunately, the resulting expressions are quite complicated and involve several unknown population entities.)

### 6.3.3 Decompositions of MSE

In the literature, a survey model for an estimator is often formulated as a decomposition of its mean square error (MSE) – see for instance [24, Section 12.2] or [29, Chapter 16]. Consider again the estimator $\hat{t}_{F\hat{a}^{(c)}}$ of $t_a$. By definition, the MSE of $\hat{t}_{F\hat{a}^{(c)}}$, with respect jointly to the sampling design $p$ in Section 2.4 and the error models $\mathrm{m}_1$ in Section 3.3 and $\mathrm{m}_2$ in Section 4.4.2, is given by

$$\begin{aligned} MSE_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right) &= E_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}} - t_a\right)^2 \\ &= V_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right) + \left(B_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right)\right)^2 \end{aligned} \tag{6.19}$$

where $B_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right) = E_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right) - t_a$ is the bias of $\hat{t}_{F\hat{a}^{(c)}}$ as estimator of $t_a$. The total variance is, in turn, composed of variances arising from different sources. The relevant components are derived in Theorem 6.3.1, as is the expectation of $\hat{t}_{F\hat{a}^{(c)}}$. Hence, in this theorem, all the tools for making an MSE decomposition for $\hat{t}_{F\hat{a}^{(c)}}$ are provided. Likewise, an MSE decomposition for $\hat{R}_F^{(c)}$ can be made by use of Theorem 6.3.2.

## 6.4 Summary

In this chapter, we tried to take a comprehensive view of the impact of some error sources on the speed survey estimators. We derived the expectations and variances of the estimators, taking both the sampling design and models for the procedures that generate $N_{Fiq}$ and $\hat{a}_k^{(c)}$ into account. In order to evaluate the resulting expressions, it would be necessary to design (and conduct) experiments in such a way that the various errors were simultaneously controlled for. Neither the time nor the resources were available to perform such experiments, but we noted some special cases under which results from earlier chapters applied.

# Chapter 7

# Summary and final remarks

Within the framework of this thesis, it would have been impracticable for us to inspect all possible quality problems associated with the speed survey. Instead, we restricted our attention to two selected nonsampling errors and an allocation aspect of the sampling error.

A problem with erroneous road lengths in the frames used in the final sampling stage was analyzed in Chapter 3. The analysis rested on a frame error model, in which the frame road length for a small area was described as a function of the true length and a random error. Two special cases of the model were the subjects of particular attention: an additive and a multiplicative relationship, respectively, between the true length and the random error. Our investigation revealed that, theoretically, if the errors in the frame road lengths were multiplicative with an expectation of one and constant variance, the length error had no bias effect on the survey estimators. Also, for the estimator of average speed, its relative (approximate) variance increase due to the length error was simply equal to the error variance. According to our experiment, the additive model described the errors in the frame road lengths slightly better than the multiplicative model did. The choice of model may, however, not be crucial since the errors, at any rate, appeared to be quite small. Our observed multiplicative errors showed a mean close to one, and a variance of less than one percent. Thus, it seems as if the speed survey estimators are not biased by the length error, and that the variance increase due to the length error, for the average speed estimator, was small. The

conclusion was based on limited data and, ideally, should be verified by a larger experiment.

Chapter 4 dealt with the problem that, currently, a number of the vehicles passing a selected site often remain unobserved. For site $k$, the plan is to observe the number of passing vehicles, $y_k$, and the sum of their travel times (the times they take to pass the site), $z_k$. If some vehicles are not noted, $y_k$ and $z_k$ must be estimated. The current estimates are obtained by simply ignoring the missing data. We suggested dividing the traffic into weighting classes and adjusting the observed traffic flow (within class) either by adding to it the number of imputed vehicles, or by weighting it by an estimated registration probability. In both cases, the adjusted flow figure is then multiplied with the mean of observed travel times to arrive at an estimate of $z_k$. A key issue is whether the estimators of $y_k$ and $z_k$ are unbiased or not. If they are, missing data have no bias effect on any survey estimator. The variances of the survey estimators are then surely larger than those of the corresponding prototype estimators (i.e., the estimators which would be used in the absence of nonsampling errors). We suspect, but cannot theoretically say with certainty, that biased estimators of $y_k$ and $z_k$ will create bias in the estimator of average speed.

Our evaluation of the current strategy for handling missing data, as well as of our two counter-proposals, rested on several models. First, we modeled the registration distribution, which generates the set of observed vehicles. In addition, we formulated models for the errors in the number of imputed vehicles, and for the errors in the estimated registration probabilities. We investigated theoretically the statistical properties of the survey estimators obtained by replacing the true $y_k$'s and $z_k$'s by estimates. We presented both general expressions for the expectations and variances of the resulting estimators, and detailed results for each considered strategy for handling missing data. In our experiment, most of the model assumptions we were able to check seemed to agree reasonably well with data, though we did encounter some model objections that still require further investigation. When we estimated $y_k$ and $z_k$, the missing data adjusted estimates came closer to the true values than the unadjusted estimates did. A measure of the average speed in site $k$ is given by the ratio $u_k = y_k/z_k$. When we estimated $u_k$ by

the ratio of estimated values on $y_k$ and $z_k$, the unadjusted estimates came as close to the true values as the estimates adjusted for missing data did. Since $u_k$ is the counterpart on 'element-level' to the average speed for all roads, our study thus hints that the estimator of average speed does not need to be adjusted for missing data. Still, our main accomplishment in Chapter 4 was the building of a theoretical framework for further evaluations of present and proposed estimation strategies. More experiments or simulations, however, are needed to make full use of the theoretical results.

In Chapter 5, we shifted our attention to the sampling error. Our primary goal was to evaluate the current allocation of the total sample over sampling stages, by estimating the variance components mirroring the variation arising from each sampling stage. This was not a trivial task, since at present the sampling sizes (within stratum) equal one in all sampling stages but the first. We developed the relevant formulae for using a 'fictitious' sampling design to estimate the variance components. The approach presupposes that two or more units are selected in the first and second stages. In order to get access to samples of size two from the final stage, we followed up by performing an experiment. In the experiment, for the estimator of average speed, the variance component due to the final sampling stage seemed to dominate. This result indicates that for unchanged total sample size, the precision of the estimator of average speed is likely to improve if the sample sizes in the third stage are increased, and the number of drawings in the first stage decreased correspondingly. Our variance component estimates are, however, quite uncertain (a sign of this being that they often take on negative values), and the conclusion needs confirmation by a larger experiment. As a support for reallocation of the sample, we gave formulae for choices of optimum sample sizes in different sampling stages.

In Chapter 6, we summarized part of our theoretical work by formulating survey models for the estimators. That is, we derived their expectations and variances, taking into account both the three-stage sampling design and our models for errors due to frame errors and missing data. We learnt that in general, the errors due to missing data and frame imperfections are entwined closely and may influence all components of the model. Results from earlier chapters apply, but only as important special cases. New experiments are

needed to evaluate the full survey models.

In this thesis, we have provided the survey management with theoretical tools for assessing the impact of frame errors and missing data; for evaluating proposals for missing data adjustments; and for reallocating the total sample over sampling stages. We have also reported results from several experiments. The experiments were performed on a small scale and did not cover all relevant aspects of the problems. Our hope is that the speed survey management, given sufficient resources being available, is willing to finish our work by collecting additional data. Also, we wish to call attention to the fact that there are still several research problems associated with the speed survey that remain to be addressed. Among these, we note the effect of the person installing the equipment on the road (the analogue to the 'interviewer effect' known from interviewer surveys) and problems associated with an aging master frame. Furthermore, other missing data problems remain: the effect of incomplete observational data for connected time periods is yet to be investigated, as are the present procedures for handling complete loss of data from a road site (including hot-deck imputation and field substitution in space, time, or both).

# Bibliography


[1] Allogg AB, *The imputation procedure in Metor*. Published by order of the Swedish National Road Administration. SNRA diary number: AL80 B 96:4605, 1996. (In Swedish).

[2] P. Biemer and S. L. Stokes, *Approaches to the modeling of measurement error*, in Measurement Errors in Surveys, P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, eds., Wiley, New York, 1991, pp. 487–516.

[3] A. Bolling and M. Wiklund, *Validation of Metor in urban areas*, VTI Meddelande 814, Statens väg- och transportforskningsinstitut, SE-581 95 Linköping, Sweden, 1997. (In Swedish).

[4] G. Casella and R. L. Berger, *Statistical Inference*, Duxbury Press, Belmont, California, 1990.

[5] C.-M. Cassel, C.-E. Särndal, and J. H. Wretman, *Some uses of statistical models in connection with the nonresponse problem*, in Incomplete Data in Sample Surveys, W. G. Madow and I. Olkin, eds., vol. 3, Academic Press, New York, 1983, pp. 143–160.

[6] D. W. Chapman, *A survey of nonresponse imputation procedures*, in Proceedings of the Social Statistics Section, Washington, 1976, American Statistical Association, pp. 245–251.

[7] S. Danielsson, *Optimum allocation for a class of sampling designs*, PhD thesis, Stockholm University, Stockholm, 1975. (In Swedish).

[8] J. Drew and W. A. Fuller, *Modeling nonresponse in surveys with callbacks*, in Proceedings of the Section on Survey Research Methods, American Statistical Association, 1980, pp. 639–642.

[9] ——, *Nonresponse in complex multiphase surveys*, in Proceedings of the Section on Survey Research Methods, American Statistical Association, 1981, pp. 623–628.

[10] A. Ekholm and S. Laaksonen, *Weighting via response modeling in the Finnish Household Budget Survey*, Journal of Official Statistics, 7 (1991), pp. 325–337.

[11] G. Forsman, *Early survey models and their use in survey quality work*, Journal of Official Statistics, 5 (1989), pp. 41–55.

[12] ——, *Recent advances in survey error modeling*, Jahrbücher für Nationalökonomie und Statistik, 211 (1993), pp. 331–350.

[13] D. L. Gerlough and M. J. Huber, *Traffic flow theory*, Special Report 165, Transportation Research Board, National Academy of Sciences, 2102 Constitution Avenue, N.W., Washington, D.C. 20418, USA, 1975.

[14] A. Giommi, *Nonparametric methods for estimating individual response probabilities*, Survey Methodology, 13 (1987), pp. 127–134.

[15] M. H. Hansen and W. N. Hurwitz, *On the theory of sampling from finite populations*, Annals of Mathematical Statistics, 14 (1943), pp. 333–362.

[16] D. G. Horvitz and D. J. Thompson, *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association, 47 (1952), pp. 663–685.

[17] A. Isaksson, *Speed measurement variations in time and space*, Research report LiU-MAT-R-1999-01, Linköping University, 1999.

[18] ——, *Frame coverage errors in a vehicle speed survey: Effects on the bias and variance of the estimators*, Linköping Studies in Arts & Science, Thesis No. 843, Linköping University, 2000.

[19] ——, *Allocation problems in a three-stage sample survey of vehicle speeds*, Research report LiU-MAT-R-2002-04, Linköping University, 2002.

[20] ——, *Survey models for a vehicle speed survey*, Research report LiU-MAT-R-2002-05, Linköping University, 2002.

[21] ——, *Weighting class adjustments for missing data in a vehicle speed survey*, Research report LiU-MAT-R-2002-01, Linköping University, 2002.

[22] G. Kalton and D. Kasprzyk, *The treatment of missing survey data*, Survey Methodology, 12 (1986), pp. 1–16.

[23] L. Kish, *Survey Sampling*, Wiley, New York, 1965.

[24] J. T. Lessler and W. D. Kalsbeek, *Nonsampling Error in Surveys*, Wiley, New York, 1992.

[25] D. C. Montgomery, *Design and Analysis of Experiments*, Wiley, New York, 4th ed., 1997.

[26] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, Irwin, Chicago, 4th ed., 1996.

[27] D. Raj, *Sampling Theory*, McGraw-Hill, New York, 1968.

[28] C.-E. Särndal and B. Swensson, *A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse*, International Statistical Review, 55 (1987), pp. 279–294.

[29] C.-E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer, New York, 1992.

[30] Swedish Government, *11-point programme for improving road traffic safety.* Memorandum April 9 from Ministry of Industry, Employment and Communications, Stockholm, Sweden, 1999. Available at http://www.vv.se/traf_sak/nollvis/programme_eng.pdf.

[31] Swedish National Road Administration, National Police Board, and Swedish Association of Local Authorities, *National road traffic safety programme for 1995-2000.* Can be ordered from the Swedish National Road Administration, SE-781 87 Borlänge, Sweden, 1994. (In Swedish).

# Appendix A

# Abbreviations

| Abbre-viation | Explanation | Section where introduced |
|---|---|---|
| ANOVA | Analysis of variance | 3.5.4 |
| EVC | Expected variable cost | 5.3.1 |
| iid | Independent and identically distributed | 3.5.4 |
| M50 | Major road with a speed limit of 50 km/h | 2.4.3 |
| M70 | Major road with a speed limit of 70 km/h | 2.4.3 |
| ME | Measurement efficiency | 1.2 |
| MSE | Mean square error | 6.3.3 |
| NID | Normally and independently distributed | 4.5.5 |
| pps | Probability-proportional-to-size sampling with replacement | 2.4.2 |
| PSU | Primary sampling unit | 2.4.1 |
| RHG | Response homogeneity group | 4.3.1 |
| SAMS | Small area market statistics | 2.4.2 |
| SCB | Statistics Sweden | 2.4.2 |
| SI | Simple random sampling without replacement | 2.4.3 |
| SIR | Simple random sampling with replacement | 5.2.2 |
| SNRA | Swedish National Road Administration | 1 |
| SSU | Secondary sampling unit | 2.4.1 |
| STSI | Stratified sampling with SI sampling in each stratum | 4.3.1 |
| TSU | Tertiary sampling unit | 2.4.3 |
| VC | Variable cost | 5.3.1 |

# Appendix B

# Proofs

## B.1 Proofs of Theorems 3.4.1, 4.4.1 and 6.3.1

### B.1.1 Preparatory lemmas

We start by noting that the sampling design $p$ of the speed survey is such that:

i. The PSUs are selected with replacement.

ii. Independent subsampling is conducted from every selection of a PSU (whether a repetition or not).

Also, the principal appearances of the prototype estimators of $t_a$ and $R$ are

$$\hat{t}_a = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{\hat{t}_{ai_\nu}}{p_{i_\nu}} \tag{B.1}$$

and $\hat{R} = \hat{t}_y/\hat{t}_z$, respectively. The expected values and variances of these estimators are investigated in the following lemma:

**Lemma B.1.1** *Under a sampling design $p$ satisfying the specifications (i)-(ii), the expected value of $\hat{t}_a$ is $E_p\left(\hat{t}_a\right) = \sum_{i=1}^{N_I} E_p\left(\hat{t}_{ai}\right)$. The variance of $\hat{t}_a$ is*

$$V_p\big(\hat{t}_a\big) = \frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left( \frac{E_p\big(\hat{t}_{ai}\big)}{p_i} - \sum_{i=1}^{N_I} E_p\big(\hat{t}_{ai}\big) \right)^2 + \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_p\big(\hat{t}_{ai}\big)}{p_i}. \tag{B.2}$$

*The estimator* $\hat{R}$ *has the approximate expected value* $E_p\left(\hat{t}_y\right)/E_p\left(\hat{t}_z\right)$*, and the approximate variance*

$$AV_p\left(\hat{R}\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{\left[E_p\left(\hat{t}_{Ei}\right)\right]^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_p\left(\hat{t}_{Ei}\right)}{p_i}\right\}. \tag{B.3}$$

The part of Lemma B.1.1 that refers to $\hat{t}_a$ is a slight generalization of Result 4.5.1 in [29] or Theorem 6.4 in [27]; the part that refers to $\hat{R}$ a slight generalization of results presented in [27, Section 6.8.2.]. Unlike the cited sources, we do not presuppose that the estimators $\hat{t}_{ai}$ and $\hat{t}_{Ei}$ are unbiased for $t_{ai}$ and $t_{Ei}$, respectively.

Now assume that some fix entity $\kappa$, included in $\hat{t}_a$ and $\hat{R}$ and associated with the subsampling from selected PSUs, is unknown. Instead of $\kappa$, we only have access to a substitute value $\hat{\kappa}$ which is impaired by a random error. Let the estimator of $t_a$ based on $\hat{\kappa}$ instead of $\kappa$ be denoted $\hat{t}_a^*$; the corresponding estimator of $R$ is $\hat{R}^* = \hat{t}_y^*/\hat{t}_z^*$. The stochastic properties of $\hat{\kappa}$ are regulated by model $\xi$.

Let expectations and variances be indicated by subscript $\xi$ if taken with respect to model $\xi$; by $p\xi$ if taken with respect jointly to the sampling design $p$ and model $\xi$. The following lemma is a straightforward expansion of Lemma B.1.1:

**Lemma B.1.2** *Jointly under a sampling design* $p$ *satisfying the specifications (i)-(ii), and model* $\xi$*, the expected value of* $\hat{t}_a^*$ *is given by* $E_{p\xi}\left(\hat{t}_a^*\right) = \sum_{i=1}^{N_I} E_{p\xi}\left(\hat{t}_{ai}^*\right)$*. The variance of* $\hat{t}_a^*$ *is given by*

$$\begin{aligned} V_{p\xi}\left(\hat{t}_a^*\right) &= \frac{1}{m_I}\sum_{i=1}^{N_I} p_i\left(\frac{E_{p\xi}\left(\hat{t}_{ai}^*\right)}{p_i} - \sum\nolimits_{i=1}^{N_I} E_{p\xi}\left(\hat{t}_{ai}^*\right)\right)^2 \\ &\quad + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{p\xi}\left(\hat{t}_{ai}^*\right)}{p_i}. \end{aligned} \tag{B.4}$$

*The estimator* $\hat{R}^*$ *has the approximate expected value* $E_{p\xi}\left(\hat{t}_y^*\right)/E_{p\xi}\left(\hat{t}_z^*\right)$*, and the approximate variance*

$$AV_{p\xi}\left(\hat{R}^*\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{\left[E_{p\xi}\left(\hat{t}_{Ei}^*\right)\right]^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{p\xi}\left(\hat{t}_{Ei}^*\right)}{p_i}\right\}. \tag{B.5}$$

Now consider the speed survey sampling design $p$ as described in detail in Section 2.4. Let conditional expectations and variances be indicated by '|'; for instance, $E_{\xi|p}$ denotes expectation with respect to model $\xi$, conditional on the sampling design $p$. By the use of conditioning, we can write

$$\begin{aligned} E_{p\xi}\left(\hat{t}^*_{ai}\right) &= E_{II}E_{III}E_{\xi|p}\left(\hat{t}^*_{ai}\right) = E_i \qquad \text{(B.6)} \\ V_{p\xi}\left(\hat{t}^*_{ai}\right) &= E_{II}E_{III}V_{\xi|p}\left(\hat{t}^*_{ai}\right) + E_{II}V_{III}E_{\xi|p}\left(\hat{t}^*_{ai}\right) \\ &\quad + V_{II}E_{III}E_{\xi|p}\left(\hat{t}^*_{ai}\right) = V_{i1} + V_{i2} + V_{i3} \qquad \text{(B.7)} \end{aligned}$$

which turns out to be useful in the proofs which follow.

## B.1.2 Proof of Theorem 3.4.1

If we interpret $\kappa$ as $N_{iq}$, $\hat{\kappa}$ as $N_{Fiq}$, model $\xi$ as the frame error model $\mathrm{m}_1$, and the estimators $\hat{t}^*_a$ and $\hat{R}^*$ as $\hat{t}_{Fa}$ and $\hat{R}_F$, respectively, Lemma B.1.2 is applicable. Hence it suffices to show that $E_{p\mathrm{m}_1}\left(\hat{t}_{F\pi ai}\right)$ and $V_{p\mathrm{m}_1}\left(\hat{t}_{F\pi ai}\right)$ equal the stated expressions. To do this, we make use of the conditioning in Equations (B.6) and (B.7).

We start with the expectation:

$$\begin{aligned} E_i &= E_{II}E_{III}E_{\mathrm{m}_1|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}a_k\right) \\ &= E_{II}E_{III}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}a_k\right) \\ &= E_{II}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{iq}}{N_{iq}}t_{aiq}\right) \\ &= \sum\nolimits_{U_{IIi}}\frac{\mu_{iq}}{N_{iq}}t_{aiq}. \end{aligned}$$

which equals the stated expression for $E_{p\mathrm{m}_1}\left(\hat{t}_{F\pi ai}\right)$.

Now we turn to the variance. First,

$$\begin{aligned} V_{i1} &= E_{II}E_{III}V_{\mathrm{m}_1|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}a_k\right) \\ &= E_{II}E_{III}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\frac{\sigma^2_{iq}}{n^2_{iq}}\left(\sum\nolimits_{s_{iq}}a_k\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \frac{\sigma_{iq}^2}{N_{iq}^2}\left(V_{aiq}+t_{aiq}^2\right)\right] \\
&= \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\sigma_{iq}^2}{N_{iq}^2}\left(V_{aiq}+t_{aiq}^2\right).
\end{aligned}
$$

Second,

$$
\begin{aligned}
V_{i2} &= E_{II}V_{III}E_{\mathrm{m}_1|p}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi_\nu}} \frac{N_{Fiq}}{n_{iq}} \sum\nolimits_{s_{iq}} a_k\right) \\
&= E_{II}V_{III}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \frac{\mu_{iq}}{n_{iq}} \sum\nolimits_{s_{iq}} a_k\right) \\
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}}\right)^2 V_{aiq}\right] \\
&= \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}}\right)^2 V_{aiq}
\end{aligned}
$$

and finally,

$$
\begin{aligned}
V_{i3} &= V_{II}E_{III}E_{\mathrm{m}_1|p}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi_\nu}} \frac{N_{Fiq}}{n_{iq}} \sum\nolimits_{s_{iq}} a_k\right) \\
&= V_{II}E_{III}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi_\nu}} \frac{\mu_{iq}}{n_{iq}} \sum\nolimits_{s_{iq}} a_k\right) \\
&= V_{II}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{aiq}\right) \\
&= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}} t_{aiq} - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{aiq}\right)^2.
\end{aligned}
$$

Addition of $V_{i1}$, $V_{i2}$ and $V_{i3}$ gives the stated expression for $V_{p\mathrm{m}_1}\left(\hat{t}_{F\pi ai}\right)$.

### B.1.3 Proof of Theorem 4.4.1

If we interpret $\kappa$ as $a_k$, $\hat{\kappa}$ as $\hat{a}_k^{(c)}$, model $\xi$ as model $\mathrm{m}_2$, and the estimators $\hat{t}_a^*$ and $\hat{R}^*$ as $\hat{t}_{\hat{a}^{(c)}}$ and $\hat{R}^{(c)}$, respectively, Lemma B.1.2 is again applicable. Hence it suffices to show that $E_{p\mathrm{m}_2}\left(\hat{t}_{\pi\hat{a}^{(c)}i}\right)$ and $V_{p\mathrm{m}_2}\left(\hat{t}_{\pi\hat{a}^{(c)}i}\right)$ equal the stated expressions. To do this, we make use of the conditioning in Equations (B.6) and (B.7).

We start with the expectation:

$$\begin{aligned} E_i &= E_{II}E_{III}E_{\mathrm{m}_2|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right) \\ &= E_{II}E_{III}\left[\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right] \\ &= E_{II}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}t_{\gamma\left(\hat{a}^{(c)}\right)iq}\right) \\ &= t_{\gamma\left(\hat{a}^{(c)}\right)i}. \end{aligned}$$

which equals the stated expression for $E_{p\mathrm{m}_2}\big(\hat{t}_{\hat{a}^{(c)}i}\big)$.

Now we turn to the variance. First,

$$\begin{aligned} V_{i1} &= E_{II}E_{III}V_{\mathrm{m}_2|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right) \\ &= E_{II}E_{III}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\left(\frac{N_{iq}}{n_{iq}}\right)^2\sum\nolimits_{s_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\right] \\ &= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\right] \\ &= \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k. \end{aligned}$$

Second,

$$\begin{aligned} V_{i2} &= E_{II}V_{III}E_{\mathrm{m}_2|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right) \\ &= E_{II}V_{III}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right) \\ &= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}V_{\gamma\left(\hat{a}^{(c)}\right)iq}\right] \\ &= \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}V_{\gamma\left(\hat{a}^{(c)}\right)iq} \end{aligned}$$

and finally,

$$\begin{aligned} V_{i3} &= V_{II}E_{III}E_{\mathrm{m}_2|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right) \\ &= V_{II}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}t_{\gamma\left(\hat{a}^{(c)}\right)iq}\right) \\ &= V_{\gamma\left(\hat{a}^{(c)}\right)IIi}. \end{aligned}$$

Addition of $V_{i2}$ and $V_{i3}$ gives $V_{\gamma\left(\hat{a}^{(c)}\right)i}$; addition of $V_{\gamma\left(\hat{a}^{(c)}\right)i}$ and $V_{i1}$ gives the stated expression for $V_{p\mathrm{m}_2}\left(\hat{t}_{\pi\hat{a}^{(c)}i}\right)$.

### B.1.4 Proof of Theorem 6.3.1

We now have *two* fixed but unknown entities: $N_{iq}$ and $a_k$. Lemma B.1.2 however still applies, if $\kappa$ is interpreted as the vector $(N_{iq}, a_k)$ and $\hat{\kappa}$ as $\left(N_{Fiq}, \hat{a}_k^{(c)}\right)$, and if expectations and variances with respect to model $\xi$ are taken with respect jointly to model $\mathrm{m}_1$ and $\mathrm{m}_2$. The estimators $\hat{t}_a^*$ and $\hat{R}^*$ now correspond to $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$, respectively. In the following, we make use of slightly expanded versions of Equations (B.6) and (B.7):

$$
\begin{aligned}
E_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right) &= E_{II}E_{III}E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right) = E_i && \text{(B.8)}\\
V_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right) &= E_{II}E_{III}V_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
&\quad +E_{II}E_{III}E_{\mathrm{m}_1|p}V_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
&\quad +E_{II}V_{III}E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
&\quad +V_{II}E_{III}E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
&= V_{i11} + V_{i12} + V_{i2} + V_{i3} && \text{(B.9)}
\end{aligned}
$$

We start by deriving the model expectations and variances. First,

$$
\begin{aligned}
& E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
= \; & E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{N_{III}}{n_{III}}\sum\nolimits_{s_{III}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right)\\
= \; & E_{\mathrm{m}_1|p}\left(\frac{N_{III}}{n_{III}}\sum\nolimits_{s_{III}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
= \; & \frac{N_{III}}{n_{III}}\sum\nolimits_{s_{III}}\frac{\mu_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\left(\hat{a}^{(c)}\right)_k,
\end{aligned}
$$

second,

$$
\begin{aligned}
& V_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
= \; & V_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{N_{III}}{n_{III}}\sum\nolimits_{s_{III}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= V_{\mathrm{m}_1|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right)\\
&= \left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\frac{\sigma^2_{iq}}{n^2_{iq}}\left(\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right)^2
\end{aligned}
$$

and finally

$$
\begin{aligned}
&E_{\mathrm{m}_1|p}V_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)\\
=\;& E_{\mathrm{m}_1|p}V_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}^{(c)}_k\right)\\
=\;& E_{\mathrm{m}_1|p}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\left(\frac{N_{Fiq}}{n_{iq}}\right)^2\sum\nolimits_{s_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\right]\\
=\;& \left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\frac{\sigma^2_{iq}+\mu^2_{iq}}{n^2_{iq}}\sum\nolimits_{s_{iq}}\delta\big(\hat{a}^{(c)}\big)_k.
\end{aligned}
$$

Now, the expectation is given by

$$
\begin{aligned}
E_i &= E_{II}E_{III}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{\mu_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right)\\
&= E_{II}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{\mu_{iq}}{N_{iq}}t_{\gamma\left(\hat{a}^{(c)}\right)iq}\right)\\
&= \sum\nolimits_{U_{IIi}}\frac{\mu_{iq}}{N_{iq}}t_{\gamma\left(\hat{a}^{(c)}\right)iq}.
\end{aligned}
$$

which equals the stated expression for $E_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\pi\hat{a}^{(c)}i}\right)$.

Now we turn to the variance:

$$
\begin{aligned}
V_{i11} &= E_{II}E_{III}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\frac{\sigma^2_{iq}}{n^2_{iq}}\left(\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right)^2\right]\\
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2\sum\nolimits_{s_{IIi}}\frac{\sigma^2_{iq}}{N^2_{iq}}\left(V_{\gamma\left(\hat{a}^{(c)}\right)iq}+t^2_{\gamma\left(\hat{a}^{(c)}\right)iq}\right)\right]\\
&= \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\sigma^2_{iq}}{N^2_{iq}}\left(V_{\gamma\left(\hat{a}^{(c)}\right)iq}+t^2_{\gamma\left(\hat{a}^{(c)}\right)iq}\right);
\end{aligned}
$$

$$\begin{aligned} V_{i12} &= E_{II}E_{III}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \frac{\sigma_{iq}^2+\mu_{iq}^2}{n_{iq}^2} \sum\nolimits_{s_{iq}} \delta\big(\hat{a}^{(c)}\big)_k\right] \\ &= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \frac{\sigma_{iq}^2+\mu_{iq}^2}{N_{iq}n_{iq}} \sum\nolimits_{U_{iq}} \delta\big(\hat{a}^{(c)}\big)_k\right] \\ &= \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\sigma_{iq}^2+\mu_{iq}^2}{N_{iq}n_{iq}} \sum\nolimits_{U_{iq}} \delta\big(\hat{a}^{(c)}\big)_k; \end{aligned}$$

$$\begin{aligned} V_{i2} &= E_{II}V_{III}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \frac{\mu_{iq}}{n_{iq}} \sum\nolimits_{s_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right) \\ &= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}}\right)^2 V_{\gamma\left(\hat{a}^{(c)}\right)iq}\right] \\ &= \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}}\right)^2 V_{\gamma\left(\hat{a}^{(c)}\right)iq}; \end{aligned}$$

$$\begin{aligned} V_{i3} &= V_{II}E_{III}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \frac{\mu_{iq}}{n_{iq}} \sum\nolimits_{s_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right) \\ &= V_{II}\left(\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq}\right) \\ &= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq} - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq}\right)^2. \end{aligned}$$

Addition of $V_{i11}$, $V_{i12}$, $V_{i2}$ and $V_{i3}$ gives $V_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\pi\hat{a}^{(c)}i}\big)$. An expression for $V_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\pi\hat{a}^{(c)}i}\big)$ is not explicitly stated in Theorem 6.3.1. Note however that

$$\begin{aligned} \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{i11}}{p_i} &= V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big); \qquad \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{i12}}{p_i} = V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big); \\ \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{i2}}{p_i} &= V_{1,\mathrm{TSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big); \qquad \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{i3}}{p_i} = V_{1,\mathrm{SSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big). \end{aligned}$$

## B.2 Proof of Corollary 3.4.3

The variance increase due to the use of $\hat{t}_{Fa}$ instead of $\hat{t}_a$ as estimator of $t_a$ is immediately obtained from Corollary 3.4.1 by replacing $\sigma_{iq}^2$ with $N_{iq}^2\tau^2$ in

B.2. Proof of Corollary 3.4.3

Equation (3.6). In the same manner, we obtain

$$AV_{p\mathrm{m}_1}\left(\hat{R}_F\right) - AV_p\left(\hat{R}\right) = \tau^2\lambda$$

where

$$\lambda = \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\left(V_{Eiq} + t_{Eiq}^2\right).$$

It remains to show that $\lambda = AV_p\left(\hat{R}\right)$. From Section 2.6,

$$AV_p\left(\hat{R}\right) = \frac{1}{t_z^2}\left[\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{t_{Ei}^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\left(V_{EIIi} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}} V_{Eiq}\right)\right].$$

We note that $V_{EIIi}$ can be written as

$$V_{EIIi} = \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}} t_{Eiq}^2 - \frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}} t_{Eiq}^2 - \frac{1}{n_{IIi}}\frac{N_{IIi}-n_{IIi}}{N_{IIi}-1}t_{Ei}^2.$$

Hence,

$$\begin{aligned}
AV_p\left(\hat{R}\right) - \lambda &= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\left[\frac{t_{Ei}^2}{p_i} - \frac{1}{p_i}\left(\frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}} t_{Eiq}^2\right.\right.\\
&\quad \left.\left. + \frac{N_{IIi}-n_{IIi}}{n_{IIi}\left(N_{IIi}-1\right)}t_{Ei}^2\right)\right]\\
&= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\left[\frac{t_{Ei}^2}{p_i}\left(1 - \frac{N_{IIi}-n_{IIi}}{n_{IIi}\left(N_{IIi}-1\right)}\right)\right.\\
&\quad \left. - \frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}} t_{Eiq}^2\right]\\
&= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\frac{n_{IIi}-1}{N_{IIi}-1}\left(t_{Ei}^2 - \sum\nolimits_{U_{IIi}} t_{Eiq}^2\right).
\end{aligned}$$

Since, in practice, $n_{IIi}$ equals one for all $i$, the derived expression is zero, and we are ready.

# Appendix C

# A useful proposition

The derivations of $\gamma$ and $\delta$ in Section 4.4.2 are facilitated by the following proposition (which is easily proven):

**Proposition C.1** *For any random variables $A$ and $B$ such that the expected value of $B$ given $A$ is constant; $E(B\,|A) = \alpha$, the expected value of $AB$ is given by*

$$E(AB) = \alpha E(A)\,,$$

*and the variance of $AB$ by*

$$V(AB) = E\big[A^2 V(B\,|A)\big] + \alpha^2 V(A)$$

*where $E(A)$ and $V(A)$ is the expectation and variance of $A$, respectively, and $V(B\,|A)$ is the variance of $B$ given $A$.*

# Appendix D

# Derivations of fictitious first-stage inclusion probabilities

The fictitious inclusion probabilities stated in Section 5.2.2 are here derived.

Let $r_i$ denote the number of times PSU $i$ occurs in the ordered sample $os_I; i \in U_I$. Note that $r_i$ is a binomial$(m_I, p_i)$-distributed variable. Since only PSUs which occur at least twice in $os_I$ are included in $s_{I'}$, the first-order inclusion probability $\pi_{I'i}$ in Equation (5.9) is derived as

$$\begin{aligned}
\pi_{I'i} &= \Pr(i \in s_{I'}) = \Pr(r_i \geq 2) = 1 - \Pr(r_i = 0) - \Pr(r_i = 1) \\
&= 1 - \frac{m_I!}{0! m_I!} p_i^0 (1-p_i)^{m_I} - \frac{m_I!}{1!(m_I-1)!} p_i^1 (1-p_i)^{m_I - 1} \\
&= 1 - (1-p_i)^{m_I} \left(1 + m_I \frac{p_i}{1-p_i}\right)
\end{aligned}$$

and the second-order inclusion probability $\pi_{I'ij}$ in Equation (5.10) as

$$\begin{aligned}
\pi_{I'ij} &= \Pr(i \& j \in s_{I'}) = \Pr[(i \in s_{I'}) \cap (j \in s_{I'})] \\
&= 1 - \Pr\left[\overline{(i \in s_{I'}) \cap (j \in s_{I'})}\right] \\
&= 1 - \Pr[(i \notin s_{I'}) \cup (j \notin s_{I'})] \\
&= 1 - \{\Pr(i \notin s_{I'}) + \Pr(j \notin s_{I'}) - \Pr[(i \notin s_{I'}) \cap (j \notin s_{I'})]\} \\
&= 1 - \{\Pr(r_i < 2) + \Pr(r_j < 2) - \Pr[(r_i < 2) \cap (r_j < 2)]\}
\end{aligned}$$

$$
\begin{aligned}
= \; & 1 - \{\Pr(r_i = 0) + \Pr(r_i = 1) + \Pr(r_j = 0) + \Pr(r_j = 1) \\
& - \Pr(r_i = 0, r_j = 0) - \Pr(r_i = 0, r_j = 1) \\
& - \Pr(r_i = 1, r_j = 0) - \Pr(r_i = 1, r_j = 1)\} \\
= \; & 1 - \left\{ \frac{m_I!}{0! m_I!} \left[ p_i^0 (1 - p_i)^{m_I} + p_j^0 (1 - p_j)^{m_I} \right] \right. \\
& + \frac{m_I!}{1! (m_I - 1)!} \left[ p_i^1 (1 - p_i)^{m_I - 1} + p_j^1 (1 - p_j)^{m_I - 1} \right] \\
& - \frac{m_I!}{0!0! m_I!} p_i^0 p_j^0 (1 - p_i - p_j)^{m_I} \\
& - \frac{m_I!}{0!1! (m_I - 1)!} \left[ p_i^0 p_j^1 (1 - p_i - p_j)^{m_I - 1} + p_i^1 p_j^0 (1 - p_i - p_j)^{m_I - 1} \right] \\
& \left. - \frac{m_I!}{1!1! (m_I - 2)!} p_i^1 p_j^1 (1 - p_i - p_j)^{m_I - 2} \right\} \\
= \; & 1 - (1 - p_i)^{m_I} \left( 1 + m_I \frac{p_i}{1 - p_i} \right) - (1 - p_j)^{m_I} \left( 1 + m_I \frac{p_j}{1 - p_j} \right) \\
& - (1 - p_i - p_j)^{m_I - 1} m_I \left[ p_i + p_j + (m_I - 1) \frac{p_i p_j}{1 - p_i - p_j} \right].
\end{aligned}
$$

# Appendix E

# ANOVA tables

## E.1 Based on the frame error experiment

### E.1.1 Under the additive error model for $N_{Fiq}$

Test for SNRA region effect:

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| SNRA region | 416735.9 | 6 | 69456.0 | 0.80 | 0.5761 |
| Error | 5054106.6 | 58 | 87139.8 | | |
| Total | 5470842.6 | 64 | | | |

Test for population center size class effect:

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Size class | 89074.4 | 2 | 44537.2 | 0.51 | 0.6012 |
| Error | 5381768.1 | 62 | 86802.7 | | |
| Total | 5470842.6 | 64 | | | |

Test for small area stratum effect:

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Small area stratum | 101240.4 | 3 | 33746.8 | 0.38 | 0.7653 |
| Error | 5369602.2 | 61 | 88026.3 | | |
| Total | 5470842.6 | 64 | | | |

### E.1.2 Under the multiplicative error model for $N_{Fiq}$

Test for SNRA region effect:

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| SNRA region | 0.0341 | 6 | 0.0057 | 0.90 | 0.5036 |
| Error | 0.3683 | 58 | 0.0063 | | |
| Total | 0.4024 | 64 | | | |

Test for population center size class effect:

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Size class | 0.0238 | 2 | 0.0119 | 1.95 | 0.1507 |
| Error | 0.3786 | 62 | 0.0061 | | |
| Total | 0.4024 | 64 | | | |

Test for small area stratum effect:

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Small area stratum | 0.0169 | 3 | 0.0056 | 0.89 | 0.4508 |
| Error | 0.3855 | 61 | 0.0063 | | |
| Total | 0.4024 | 64 | | | |

# E.2 Based on the missing data experiment

## E.2.1 Under the multiplicative error model for $n_{I_{kh}}$

Test for site effect with imputations in the valve measurements *removed* ($\hat{\varepsilon}_{kh} = \hat{\varepsilon}_{kh,\text{woi}}$):

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Site | 3.04343 | 4 | 0.76086 | 2.33 | 0.0603 |
| Error | 37.58069 | 115 | 0.32679 | | |
| Total | 40.62412 | 119 | | | |

Test for site effect with imputations in the valve measurements *retained* ($\hat{\varepsilon}_{kh} = \hat{\varepsilon}_{kh,\text{wi}}$):

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Site | 1.75452 | 4 | 0.43863 | 4.30 | 0.0028 |
| Error | 11.74024 | 115 | 0.10209 | | |
| Total | 13.49476 | 119 | | | |

## E.2.2 Under the additive error model for $\hat{\theta}^{(2)}_{kh}$

Test for site effect with imputations in the valve measurements *removed* ($\hat{\epsilon}_{kh} = \hat{\epsilon}_{kh,\text{woi}}$):

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Site | 0.00518 | 4 | 0.00129 | 9.38 | $< 0.0001$ |
| Error | 0.01587 | 115 | 0.00014 | | |
| Total | 0.02104 | 119 | | | |

Test for site effect with imputations in the valve measurements *retained* ($\hat{\epsilon}_{kh} = \hat{\epsilon}_{kh,\mathrm{wi}}$):

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Site | 0.01809 | 4 | 0.00452 | 8.52 | $< 0.0001$ |
| Error | 0.06107 | 115 | 0.00053 | | |
| Total | 0.07916 | 119 | | | |

### E.2.3 Under the multiplicative error model for $\hat{\theta}_{kh}^{(2)}$

Test for site effect with imputations in valve measurements *removed* ($\hat{\epsilon}_{kh} = \hat{\epsilon}_{kh,\mathrm{woi}}$):

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Site | 0.01050 | 4 | 0.00263 | 9.45 | $< 0.0001$ |
| Error | 0.03198 | 115 | 0.00028 | | |
| Total | 0.04248 | 119 | | | |

Test for site effect with imputations in valve measurements *retained* ($\hat{\epsilon}_{kh} = \hat{\epsilon}_{kh,\mathrm{wi}}$):

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ | $P$-value |
|---|---|---|---|---|---|
| Site | 0.02750 | 4 | 0.00688 | 7.96 | $< 0.0001$ |
| Error | 0.09933 | 115 | 0.00086 | | |
| Total | 0.12683 | 119 | | | |

# Appendix F

# Variance component estimates for $\hat{t}_y$ and $\hat{t}_z$

In the following tables, estimates of the approximate variances of $\hat{t}_y$ and $\hat{t}_z$ and their components, for various combinations of SSU and TSU strata, are presented. The estimates for $\hat{t}_y$ are given in thousands of kilometers; the estimates for $\hat{t}_z$ in thousands of hours.

| Development type | Road type | $\hat{V}_{3\text{st}}(\hat{t}_y)$ | $\hat{V}_{\text{TSU}}(\hat{t}_y)$ | $\hat{V}_{\text{SSU}}(\hat{t}_y)$ | $\hat{V}_{\text{PSU}}(\hat{t}_y)$ |
|---|---|---|---|---|---|
| City | M70 | 2005.9518 | 562.4455 | $*$ | 1443.5063 |
| City | Other | 6.8357 | 5.8788 | 0.9569 | $*$ |
| Industrial | M70 | 600.3100 | 384.2024 | $*$ | 216.1076 |
| Industrial | Other | 4245.1520 | 4564.1312 | $-$ | $-$ |
| Residential | M70 | 17167.6676 | 606.4022 | 16561.2654 | $*$ |
| Residential | Other | 28.8047 | 5.3361 | 23.4686 | $*$ |
| Other | M70 | 672.9285 | 72.2111 | 600.7174 | $*$ |
| Other | Other | 0.4747 | 0.2510 | 0.2237 | $*$ |

| Development type | Road type | $\hat{V}_{3\text{st}}(\hat{t}_z)$ | $\hat{V}_{\text{TSU}}(\hat{t}_z)$ | $\hat{V}_{\text{SSU}}(\hat{t}_z)$ | $\hat{V}_{\text{PSU}}(\hat{t}_z)$ |
|---|---|---|---|---|---|
| City | M70 | 0.9143 | 0.2054 | $*$ | 0.7089 |
| City | Other | 0.0039 | 0.0032 | 0.0007 | $*$ |
| Industrial | M70 | 0.2697 | 0.1963 | $*$ | 0.0734 |
| Industrial | Other | 1.4209 | 1.3942 | $*$ | 0.0267 |
| Residential | M70 | 6.5975 | 0.1195 | 6.4780 | $*$ |
| Residential | Other | 0.0215 | 0.0019 | 0.0196 | $*$ |
| Other | M70 | 0.3748 | 0.0291 | 0.3457 | $*$ |
| Other | Other | 0.0003 | 0.0002 | 0.0001 | $*$ |