Contents

1 Introduction

Every summer since 1996, the Swedish National Road Administration (SNRA) conducts a traffic survey on urban roads. The roads are thought of as partitioned into one-meter road sites, that are the population elements. Data are collected for a random sample of sites (selected by a multi-stage sampling design) by use of a measurement equipment installed on the road. The principal aim of the survey is to estimate the average speed on the roads.

In earlier work by the author ([3], [4], [5]), the isolated impact of some different sources of error on the speed survey estimators is investigated. Here, we adopt a comprehensive view towards those errors, by formulating *survey models* for the two types of estimators in use: an estimator of a population total and of a ratio of totals. 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.

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 US Census Bureau. A review on the model development before 1970 is given in [1]; for later development, see [2, Paper II]. For examples of some quite general models, see [6, Ch. 12] and [8, Ch. 16].

2 Variables and parameters

The main study variables are the traffic flow, y, and the travel time, z. For a given road site and time period, the traffic flow is the number of passing vehicles, and the travel time is the total time the vehicles take to pass the site. Let U denote the target population "in space" – the set of one-meter road sites that make up the urban roads – and U_Υ the target population "in time" – the set of twentyfour-hour periods that make up the time period of study. The population total of y is given by $\sum_{U_\Upsilon} \sum_U y_k^\upsilon$, where y_k^υ equals the traffic flow in site $k \in U$ during twentyfour-hour period $\upsilon \in U_\Upsilon$. Correspondingly, the z total 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.

In this report, 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 $\upsilon \in U_\Upsilon, k \in U$. Hence, we will hereafter drop the time index and refer to $t_y = \sum_U y_k$ as total vehicle mileage, $t_z = \sum_U z_k$ as total travel time, and $R = t_y/t_z$ as the average speed on the roads. Instead of treating population totals for various study variables as separate cases, we will often simply talk about t_a: the population total for an arbitrary study variable a.

3 Sampling design

Road sites are selected for observation by means of a three-stage sampling design. A brief description, based on [3, Ch. 2], of the different stages, will now be given.

The primary sampling units (PSUs) are the N_I population centers in Sweden, labeled $i = 1, ..., N_I$. The ith PSU is represented 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 secondary sampling units (SSUs). The set of SSUs formed by the subdivision of i is denoted $U_{IIi} = \{1, ..., q, ..., N_{IIi}\}$. Finally, the roads in small area q in population center i are viewed as partitioned into N_{iq} one-meter road sites (representing the tertiary sampling units – the TSUs). This set of sites is denoted U_{iq}.

The sample s of road sites is selected from the population U of urban roads in the following way.

Stage I A probability-proportional-to-size (pps) sample of PSUs is drawn with probability proportional to the number of inhabitants. At every draw, p_i is the probability of selecting the ith PSU. Let i_ν denote the PSU selected in the νth draw, $\nu = 1, ..., m_I$, where m_I is the number of draws. The probability of selecting i_ν is denoted p_{i_ν}. If the ith PSU is selected in the νth 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_ν that is a component of os_I, a simple random (SI) sample s_{IIi_ν} of SSUs of size n_{IIi_ν} is selected.

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

To simplify, both in the above description and further, we basically ignore some features of the design. In particular, we ignore that *stratified* sampling is used in each stage. Moreover, we ignore that the three largest PSUs define a take-all stratum in stage I; and finally, that SSUs are selected with pps within one stratum (Residential Areas) in stage II.

In practice, within stratum, the sample sizes in each stage are $m_I = 10$, $n_{IIi_\nu} = 1$ and $n_{i_\nu q} = 1$.

4 The prototype estimators of t_a and R

In this section, we present the 'prototype' estimators of t_a and R; that is, the estimators designed for the ideal situation of no nonsampling errors. Estimators of population entities are denoted by a hat; π estimators (Horvitz-Thompson estimators) are also indicated by subscript π.

Define the totals $t_{aiq} = \sum_{U_{iq}} a_k$ and $t_{ai} = \sum_{U_{IIi}} t_{aiq}$, and note that

$$t_a = \sum_{i=1}^{N_I} t_{ai} = \sum_{i=1}^{N_I} \sum\nolimits_{U_{IIi}} t_{aiq}.$$

The prototype estimator of t_a is given by

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

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 νth draw, then $\hat{t}_{\pi a i_\nu} = \hat{t}_{\pi a i} = (N_{IIi}/n_{IIi}) \sum_{s_{IIi}} \hat{t}_{\pi a i q}$ and $\hat{t}_{\pi a i_\nu q} = \hat{t}_{\pi a i q} = (N_{iq}/n_{iq}) \sum_{s_{iq}} a_k$.)

The prototype estimator of R is

$$\hat{R} = \hat{t}_y / \hat{t}_z. \tag{2}$$

Let E_p and V_p denote the expectation and variance operators with respect to the sampling design p. 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 (for details on this technique, see for instance [8, Section 5.5]). Let AV_p denote such an approximative variance, again with respect to p.

From [8, Result 4.5.1], $\hat{t}_a$ is design-unbiased for t_a (that is, $E_p\big(\hat{t}_a\big) = t_a$).

The variance of $\hat{t}_a$ is given by

$$\begin{aligned} V_p(\hat{t}_a) &= \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{t_{ai}}{p_i} - t_a\right)^2 \\ &+ \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{1}{p_i} N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} S_{t_a U_i}^2 \\ &+ \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} N_{iq}^2 \frac{1-f_{iq}}{n_{iq}} S_{aU_{iq}}^2 \end{aligned} \tag{3}$$

where

$$\begin{aligned} f_{IIi} &= n_{IIi}/N_{IIi}; \quad S_{t_a U_i}^2 = \frac{1}{N_{IIi}-1}\sum\nolimits_{U_{IIi}} \left(t_{aiq} - \frac{t_{ai}}{N_{IIi}}\right)^2; \\ f_{iq} &= n_{iq}/N_{iq}; \quad S_{aU_{iq}}^2 = \frac{1}{N_{iq}-1}\sum\nolimits_{U_{iq}} \left(a_k - \frac{t_{aiq}}{N_{iq}}\right)^2 \end{aligned}$$

From [7, Sec. 6.8.2.], $\hat{R}$ is approximately design-unbiased for R. The approximate variance of $\hat{R}$ is given by

$$AV_p\left(\hat{R}\right) = \frac{V_p(\hat{t}_E)}{t_z^2} \tag{4}$$

where $V_p(\hat{t}_E)$ is obtained from $V_p(\hat{t}_a)$ by replacing a_k with the new variable $E_k = y_k - Rz_k$. Since $t_E = 0$, the first term of $AV_p\left(\hat{R}\right)$ can be simplified as follows:

$$\frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{t_{Ei}}{p_i} - t_E\right)^2 = \frac{1}{m_I}\sum_{i=1}^{N_I} \frac{t_{Ei}^2}{p_i}.$$

5 Problems of frame errors and missing data

In practice, various nonsampling errors prevent the use of the prototype estimators. Estimation with frame errors is treated in [3], estimation with missing data in [5]. In this section, we recapitulate the nature of the errors, how they manifest themselves in the estimators, and how we choose to model them.

5.1 The errors in brief

Frame errors In the third sampling stage, a frame of the small area road network, constructed by manual measurements of road lengths from a map, is employed to select road sites for observation. Let U_{Fiq} (of size N_{Fiq}) denote the set of road sites according to the frame. If U_{Fiq} differs from U_{iq}, the actual sampling procedure differs from the one described in Section 3: in reality an SI sample $s_{Fi_\nu q}$ of sites of size $n_{i_\nu q}$ is then drawn from U_{Fiq}.

Missing data If the measurement instrument fails to observe all vehicles passing a selected road site, the value on study variable a for the site is unknown. Let $\hat{a}_k^{(c)}$ denote an estimator of a_k under estimation strategy c. The present approach, to simply ignore the missing data, corresponds to $c = 0$. Two alternative procedures intended for adjusting for missing data, corresponding to $c = 1$ and $c = 2$, are discussed in [5].

5.2 The error-prone estimators

Due to the errors discussed in Section 5.1, instead of the prototype estimator $\hat{t}_a$, we can only observe

$$\hat{t}_{F\hat{a}^{(c)}} = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{1}{p_{i_\nu}} \frac{N_{IIi_\nu}}{n_{IIi_\nu}} \sum\nolimits_{s_{IIi_\nu}} \hat{t}_{F\pi\hat{a}^{(c)}i_\nu q} \tag{5}$$

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

$$\hat{R}_F^{(c)} = \hat{t}_{F\hat{y}^{(c)}} / \hat{t}_{F\hat{z}^{(c)}}. \tag{6}$$

5.3 Models for specific errors

The following frame error model is drawn from [3, Chapter 3].

A frame error model, m_1
For small area $q \in s_{IIi}$,

- the sample s_{Fiq} is an SI sample from U_{iq},
- the frame road length N_{Fiq} is a function of the true length and a random error, and
- the N_{Fiq}'s are independent random variables.

In mathematical terms, the first statement of m_1 means that s_{Fiq} is assumed to equal s_{iq}. Then, the error-prone estimators are subjected to frame errors only through the weighting by N_{Fiq} instead of N_{iq}.

The joint probability distribution (conditional on s_{iq}) of the independent random variables $\hat{a}_k^{(c)}$ $(c = 0, 1, 2)$ for site $k \in s_{iq}$, is called **model $\mathbf{m}_2$**. This missing data model, taken from [5, Section 5], is very general. In the cited source, special cases of m_2 are however formulated for different estimation strategies.

We have no reason to believe that the two types of error are somehow related. Therefore, to keep things simple, we assume that the mechanism that generates $\hat{a}^{(c)}$ is unconfounded with the one that generates N_F. That is, the probability of a certain outcome of $\hat{a}_k^{(c)}$ is assumed to be unaffected by N_{Fiq}.

Remark 5.1 *The notations for the models differ from those originally used. The frame error model is denoted 'm' in [3, Chapter 3]; the missing data model 'ξ' in [5, Section 5]*

6 Survey models for the error-prone estimators

The expectations and variances of $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$ are now derived. In this, we need to take several sources of randomness into account: three stages of sampling as well as the procedures that generate N_{Fiq} and $\hat{a}_k^{(c)}$. Thus not very surprisingly, we are caught in some quite intricate variance expressions. By assuming N_{Fiq} and $\hat{a}_k^{(c)}$ to be unbiased for their true counterparts, we are able to simplify the formulas to some extent. We finish off with a note on the relation between our work and a customary decomposition of mean square error (MSE) by error source.

6.1 Notation

Let expectations and variances with respect to the error models m_1 and m_2 be indicated, respectively, by subscript m_1 and m_2. Let us also refine the notation regarding the sampling design p described in Section 3. Expectations and variances are indicated by subscript I if taken with respect to the design used in stage one; II if taken with respect to the design used in stage two, given os_I; and III if taken with respect to the design used in stage three, given os_I and s_{IIi_v}.

In the following, we make diligent use of conditioning. Conditional expectations and variances are then indicated by '|' (for instance, $E_{\mathrm{m}_2|\mathrm{m}_1}$ denote expectation with respect to model m_2, conditional on model m_1). In order to shorten the formulas, we use special notation for the conditional expectations and variances of N_{Fiq} and $\hat{a}_k^{(c)}$: for $q \in s_{IIi}$,

$$E_{\mathrm{m}_1|p}(N_{Fiq}) = E_{\mathrm{m}_1}(N_{Fiq}\,|s_{IIi}) = \mu_{iq} \tag{7}$$

$$V_{\mathrm{m}_1|p}(N_{Fiq}) = V_{\mathrm{m}_1}(N_{Fiq}\,|s_{IIi}) = \sigma_{iq}^2 \tag{8}$$

and for $k \in s_{iq}$,

$$E_{\mathrm{m}_2|p}\left(\hat{a}_k^{(c)}\right) = E_{\mathrm{m}_2}\left(\hat{a}_k^{(c)}\,|s_{iq}\right) = \gamma\big(\hat{a}^{(c)}\big)_k \tag{9}$$

$$V_{\mathrm{m}_2|p}\left(\hat{a}_k^{(c)}\right) = V_{\mathrm{m}_2}\left(\hat{a}_k^{(c)}\,|s_{iq}\right) = \delta\big(\hat{a}^{(c)}\big)_k \tag{10}$$

The population entities $t_{\gamma(\hat{a}^{(c)})iq}$ and $S^2_{\gamma(\hat{a}^{(c)})U_{iq}}$ for $\gamma\big(\hat{a}^{(c)}\big)$ are defined as in Section 4, only with $a = \gamma\big(\hat{a}^{(c)}\big)$.

6.2 Expectations and variances

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

$$E_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = E_p E_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\hat{a}^{(c)}}\right) \tag{11}$$

and its total variance as

$$\begin{aligned} V_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= V_p E_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\hat{a}^{(c)}}\right) + E_p V_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\hat{a}^{(c)}}\right) \\ &\quad + E_p E_{\mathrm{m}_1|p} V_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\hat{a}^{(c)}}\right) \\ &= V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big)\,. \end{aligned} \tag{12}$$

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 $\hat{t}_{F\pi\gamma(\hat{a}^{(c)})i_\nu q}$ 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 [6, Eq. (12.9)] or the "simple measurement variance" in [8, Eq. (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:

$$\begin{aligned} V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= V_I E_{II} E_{III} E_{\mathrm{m_1 m_2}|p}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + E_I V_{II} E_{III} E_{\mathrm{m_1 m_2}|p}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &\quad + E_I E_{II} V_{III} E_{\mathrm{m_1 m_2}|p}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &= V_{1,\mathrm{PSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,\mathrm{SSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,\mathrm{TSU}}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \end{aligned} \tag{13}$$

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 6.1 *Jointly under the sampling design p in Section 3 and the error models m_1 and m_2 in Section 5.3, the expected value of $\hat{t}_{\hat{a}^{(c)}}$ is given by*

$$E_{pm_1m_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \sum_{i=1}^{N_I} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq}. \tag{14}$$

The variance of $\hat{t}_{F\hat{a}^{(c)}}$ is given by

$$\begin{aligned} V_{pm_1m_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= V_1\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &= V_{1,PSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,SSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_{1,\,TSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \\ &\quad + V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) + V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) \end{aligned} \tag{15}$$

where

$$V_{1,PSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left(\frac{1}{p_i} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq} - E_{pm_1m_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) \right)^2 \tag{16}$$

$$\begin{aligned} V_{1,SSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{1}{p_i} N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \\ &\quad \times \sum\nolimits_{U_{IIi}} \left(\frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq} - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \frac{\mu_{iq}}{N_{iq}} t_{\gamma\left(\hat{a}^{(c)}\right)iq} \right)^2 \end{aligned} \tag{17}$$

$$V_{1,\,TSU}\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{1}{p_i} \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \mu_{iq}^2 \frac{1-f_{iq}}{n_{iq}} S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{iq}} \tag{18}$$

$$V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\sigma^2_{iq}}{N^2_{iq}}\left(N^2_{iq}\frac{1-f_{iq}}{n_{iq}}S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{iq}} + t^2_{\gamma\left(\hat{a}^{(c)}\right)_{iq}}\right) \tag{19}$$

and

$$V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\mu^2_{iq}+\sigma^2_{iq}}{N_{iq}n_{iq}}\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k\,. \tag{20}$$

The proof of Theorem 6.1 is given in Appendix A.

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 [7, Sec. 6.8.2.].

Theorem 6.2 *Jointly under the sampling design p in Section 3 and the error models m_1 and m_2 in Section 5.3, the estimator $\hat{R}_F^{(c)}$ is approximately unbiased for*

$$R_F^{(c)} = \frac{E_{pm_1m_2}\big(\hat{t}_{F\hat{y}^{(c)}}\big)}{E_{pm_1m_2}\big(\hat{t}_{F\hat{z}^{(c)}}\big)}. \tag{21}$$

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

$$\begin{aligned} AV_{pm_1m_2}\Big(\hat{R}_F^{(c)}\Big) &= V_1\Big(\hat{R}_F^{(c)}\Big) + V_2\Big(\hat{R}_F^{(c)}\Big) + V_3\Big(\hat{R}_F^{(c)}\Big) \\ &= \frac{V_1\big(\hat{t}_{F\pi\hat{E}^{(c)}}\big)}{t_z^2} + \frac{V_2\big(\hat{t}_{F\pi\hat{E}^{(c)}}\big)}{t_z^2} + \frac{V_3\big(\hat{t}_{F\pi\hat{E}^{(c)}}\big)}{t_z^2} \end{aligned} \tag{22}$$

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 6.1 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 Simplifications

The expectations and variances of $\hat{t}_{F\hat{a}^{(c)}}$ and $\hat{R}_F^{(c)}$, as presented in Theorems 6.1 and 6.2, are quite complicated and thus hard to evaluate. We now try to simplify the expressions by making a few assumptions:

i. The frame road lengths are unbiased for the true road lengths

ii. The missing data adjusted estimators $\hat{a}_k^{(1)}$ and $\hat{a}_k^{(2)}$ are unbiased for a_k

Assumption i, which is confirmed by our frame errors investigation in [3], implies that we can replace μ_{iq} by N_{iq}. Our results in [5] give some support for the second assumption, which means that we can substitute $\gamma\big(\hat{a}^{(c)}\big)_k$ by a_k if $c = 1$ or 2.

Under assumptions i-ii, 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). Furthermore, for $\hat{t}_{F\hat{a}^{(c)}}$,

$$V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\sigma_{iq}^2}{N_{iq}^2}\left(N_{iq}^2\frac{1-f_{iq}}{n_{iq}}S_{aU_{iq}}^2 + t_{aiq}^2\right) \quad (23)$$

$$V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\left(\frac{N_{iq}}{n_{iq}} + \frac{\sigma_{iq}^2}{N_{iq}n_{iq}}\right)\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k \quad (24)$$

(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.

6.4 Decompositions of MSE

In the literature, a survey model for an estimator is often formulated as a decomposition of its MSE – see for instance [6, Sec. 12.2] or [8, Ch. 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 3 and the error models m_1 and m_2 in Section 5.3, is given by

$$\begin{aligned} MSE_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= E_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}} - t_a\big)^2 \\ &= V_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) + \big(B\big(\hat{t}_{F\hat{a}^{(c)}}\big)\big)^2 \end{aligned} \tag{25}$$

where $B\big(\hat{t}_{F\hat{a}^{(c)}}\big) = E_{p\mathrm{m}_1\mathrm{m}_2}\big(\hat{t}_{F\hat{a}^{(c)}}\big) - t_a$ is the bias of $\hat{t}_{F\hat{a}^{(c)}}$ as estimator of $\hat{t}_{F\hat{a}^{(c)}}$. The total variance is, in turn, composed of variances arising from different sources. The relevant components are derived in Theorem 6.1, as is the expectation of $\hat{t}_{F\hat{a}^{(c)}}$. Hence, in this theorem, all the tools for making an MSE decomposition for $\hat{t}_{F\hat{a}^{(c)}}$ is provided. Likewise, an MSE decomposition for $\hat{R}_F^{(c)}$ can be made by use of Theorem 6.2.

7 Discussion and final remarks

In this report, we have tried to take a complete grip of the impact of some error sources on the speed survey estimators. Thus, in some respects, it summarizes earlier work presented in [3], [4] and [5]. In the cited sources, however, experimental data are utilized to evaluate model assumptions and estimate model expectations and variances. Most of those results have not been mentioned yet. The main reason for this is that they are produced not taking the entwinement of various errors properly into account. Estimation of the expectations and variances stated in Theorems 6.1 and 6.2 would, in general, require additional data collection, conducted in such a way that the various errors are simultaneously controlled for. Assume however, as in Section 6.3, that $\mu_{iq} = N_{iq}$ and $\gamma\big(\hat{a}^{(c)}\big)_k = a_k$. 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 [4] of their relative sizes applies. The investigation suggests that, for $\hat{R}_F^{(c)}$, the $V_{1,\mathrm{TSU}}$ component prevails among the three.

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

- from [3, Corollary 3.2.3], the V_2 term for $\hat{R}_F^{(c)}$ can be written as

$$V_2\Big(\hat{R}_F^{(c)}\Big) = \tau^2 A V_p\Big(\hat{R}\Big). \tag{26}$$

  In the same source, from some experimental data, a 95 percent upper bounded confidence interval for τ^2 is worked out to equal $[0, 0.00848]$.

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

$$\begin{aligned} V_3\Big(\hat{R}_F^{(c)}\Big) &= \big(1+\tau^2\big)\,\frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}} \\ &\quad \times \sum\nolimits_{U_{iq}} \delta\Big(\hat{E}^{(c)}\Big)_k \end{aligned} \tag{27}$$

  where $\hat{E}^{(c)} = \hat{y}^{(c)} - R\hat{z}^{(c)}$. In [5], 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.

The survey models presented in this report are quite complex and hard to grasp. Still, they support only the impact of a few, selected, sources of error. Among the possible errors that are not accounted for, we note

- the effect of the person installing the equipment on the road (the analogue to the 'interviewer effect' known from interviewer surveys),

- the effects of present procedures for handling the complete loss of observational data from a road site (including hot-deck imputation and field substitution in space or time or both), and
- the effect of incomplete data from a road site for connected time periods.

We do however entertain hopes that we have taken the most influential errors into account.

References

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

[2] ——, *Survey Models–A Review and Some Applications to Reinterview Data*, PhD thesis, Lund University, Department of Statistics, Lund, 1991.

[3] A. Isaksson, *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, Linköpings universitet, 2000.

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

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

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

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

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

A Proof of Theorem 6.1

We start with the expectation $E_{p\mathrm{m}_1\mathrm{m}_2}\left(\hat{t}_{F\hat{a}^{(c)}}\right)$. It is derived exactly as the expected value of $\hat{t}_a$, only with t_{ai} replaced by

$$\begin{aligned}
E_{II}E_{III}E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\hat{t}_{F\hat{a}^{(c)}i}\right) &= E_{II}E_{III}E_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\hat{a}_k^{(c)}\right)\\
&= E_{II}E_{III}E_{\mathrm{m}_1|p}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fiq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
&= E_{II}E_{III}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
&= E_{II}\left(\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{iq}}{N_{iq}}t_{\gamma\left(\hat{a}^{(c)}\right)iq}\right)\\
&= \sum\nolimits_{U_{IIi}}\frac{\mu_{iq}}{N_{iq}}t_{\gamma\left(\hat{a}^{(c)}\right)iq}.
\end{aligned}$$

We now turn to the variances.

$$\begin{aligned}
V_1\left(\hat{t}_{F\hat{a}^{(c)}}\right) &= V_pE_{\mathrm{m}_1|p}E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fi_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\hat{a}_k^{(c)}\right)\\
&= V_pE_{\mathrm{m}_1|p}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fi_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
&= V_IE_{II}E_{III}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{i_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
&\quad +E_IV_{II}E_{III}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{i_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
&\quad +E_IE_{II}V_{III}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{\mu_{i_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\gamma\left(\hat{a}^{(c)}\right)_k\right)\\
&= V_{1,\mathrm{PSU}}\left(\hat{t}_{F\hat{a}^{(c)}}\right)+V_{1,\mathrm{SSU}}\left(\hat{t}_{F\hat{a}^{(c)}}\right)+V_{1,\mathrm{TSU}}\left(\hat{t}_{F\hat{a}^{(c)}}\right)
\end{aligned}$$

The $V_{1,\mathrm{PSU}}\left(\hat{t}_{F\hat{a}^{(c)}}\right)$ term is derived as the first term in Equation (3), only with t_{ai} replaced by $\sum_{U_{IIi}}\left(\mu_{iq}/N_{iq}\right)t_{\gamma\left(\hat{a}^{(c)}\right)iq}$ and t_a by $\sum_{i=1}^{N_I}\sum_{U_{IIi}}\left(\mu_{iq}/N_{iq}\right)t_{\gamma\left(\hat{a}^{(c)}\right)iq}$. The $V_{1,\mathrm{SSU}}\left(\hat{t}_{F\hat{a}^{(c)}}\right)$ term is derived as the second term in Equation (3), only with t_{aiq} replaced by $\left(\mu_{iq}/N_{iq}\right)t_{\gamma\left(\hat{a}^{(c)}\right)iq}$ and t_{ai} (again) by $\sum_{U_{IIi}}\left(\mu_{iq}/N_{iq}\right)t_{\gamma\left(\hat{a}^{(c)}\right)iq}$. The $V_{1,\mathrm{TSU}}\left(\hat{t}_{F\hat{a}^{(c)}}\right)$ term, finally, is derived as the third term in Equation (3),

only with a_k replaced by $\gamma\big(\hat{a}^{(c)}\big)_k$ and t_{aiq} (again) by $\big(\mu_{iq}/N_{iq}\big)\, t_{\gamma\left(\hat{a}^{(c)}\right)iq}$.

$$
\begin{aligned}
V_2\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= E_p V_{\mathrm{m}_1|p} E_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fi_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\hat{a}_k^{(c)}\right)\\
&= E_p V_{\mathrm{m}_1|p}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fi_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\gamma\big(\hat{a}^{(c)}\big)_k\right)\\
&= E_I E_{II} E_{III}\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\left(\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\right)^2\sum\nolimits_{s_{IIi_\nu}}\frac{\sigma_{i_\nu q}^2}{n_{i_\nu q}^2}\left(\sum\nolimits_{s_{i_\nu q}}\gamma\big(\hat{a}^{(c)}\big)_k\right)^2\right]\\
&= E_I E_{II}\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\left(\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\right)^2\sum\nolimits_{s_{IIi_\nu}}\frac{\sigma_{i_\nu q}^2}{N_{i_\nu q}^2}\left(N_{i_\nu q}^2\frac{1-f_{i_\nu q}}{n_{i_\nu q}}S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{i_\nu q}}\right.\right.\\
&\qquad \left.\left. + t^2_{\gamma\left(\hat{a}^{(c)}\right)_{i_\nu q}}\right)\right]\\
&= E_I\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{U_{IIi_\nu}}\frac{\sigma_{i_\nu q}^2}{N_{i_\nu q}^2}\left(N_{i_\nu q}^2\frac{1-f_{i_\nu q}}{n_{i_\nu q}}S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{i_\nu q}}\right.\right.\\
&\qquad \left.\left. + t^2_{\gamma\left(\hat{a}^{(c)}\right)_{i_\nu q}}\right)\right]\\
&= \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\sigma_{iq}^2}{N_{iq}^2}\left(N_{iq}^2\frac{1-f_{iq}}{n_{iq}}S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{iq}} + t^2_{\gamma\left(\hat{a}^{(c)}\right)_{iq}}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
V_3\big(\hat{t}_{F\hat{a}^{(c)}}\big) &= E_p E_{\mathrm{m}_1|p} V_{\mathrm{m}_2|p\mathrm{m}_1}\left(\frac{1}{m_I}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{s_{IIi_\nu}}\frac{N_{Fi_\nu q}}{n_{i_\nu q}}\sum\nolimits_{s_{i_\nu q}}\hat{a}_k^{(c)}\right)\\
&= E_p E_{\mathrm{m}_1|p}\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\left(\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\right)^2\sum\nolimits_{s_{IIi_\nu}}\left(\frac{N_{Fi_\nu q}}{n_{i_\nu q}}\right)^2\sum\nolimits_{s_{i_\nu q}}\delta\big(\hat{a}^{(c)}\big)_k\right]\\
&= E_I E_{II} E_{III}\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\left(\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\right)^2\sum\nolimits_{s_{IIi_\nu}}\frac{\sigma_{i_\nu q}^2+\mu_{i_\nu q}^2}{n_{i_\nu q}^2}\sum\nolimits_{s_{i_\nu q}}\delta\big(\hat{a}^{(c)}\big)_k\right]\\
&= E_I E_{II}\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\left(\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\right)^2\sum\nolimits_{s_{IIi_\nu}}\frac{\sigma_{i_\nu q}^2+\mu_{i_\nu q}^2}{N_{i_\nu q}n_{i_\nu q}}\sum\nolimits_{U_{i_\nu q}}\delta\big(\hat{a}^{(c)}\big)_k\right]\\
&= E_I\left[\frac{1}{m_I^2}\sum_{\nu=1}^{m_I}\frac{1}{p_{i_\nu}^2}\frac{N_{IIi_\nu}}{n_{IIi_\nu}}\sum\nolimits_{U_{IIi_\nu}}\frac{\sigma_{i_\nu q}^2+\mu_{i_\nu q}^2}{N_{i_\nu q}n_{i_\nu q}}\sum\nolimits_{U_{i_\nu q}}\delta\big(\hat{a}^{(c)}\big)_k\right]\\
&= \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{\mu_{iq}^2+\sigma_{iq}^2}{N_{iq}n_{iq}}\sum\nolimits_{U_{iq}}\delta\big(\hat{a}^{(c)}\big)_k.
\end{aligned}
$$