Contents

1 Introduction **3**

2 The missing data problem **3**

3 Proposals for missing data adjustments **4**
3.1 A model of the registration mechanism 4
3.2 Strategy 1 6
3.2.1 Estimator of flow 7
3.2.2 Estimator of travel time 7
3.2.3 A model of the imputation mechanism 8
3.3 Strategy 2 9
3.3.1 Estimator of flow 9
3.3.2 Estimator of travel time 10
3.3.3 An error model for $\hat{\theta}^{(2)}$ 10

4 Sampling and estimation **10**
4.1 Sampling design 11
4.2 Estimation with complete data 11

5 Estimation with missing data **13**
5.1 General results 13
5.2 Results for present and proposed estimators 15
5.2.1 Strategy 0 15
5.2.2 Strategy 1 18
5.2.3 Strategy 2 22
5.3 Summary of theoretical findings 28

6 Empirical study **29**
6.1 Study objectives 29
6.2 Design of the study 29
6.3 Data processing 31
6.4 Estimation 33
6.5 Analysis 37
6.5.1 The forming of registration homogeneity groups 37
6.5.2 Evaluation of the multiplicative imputation error model 37
6.5.3 Evaluation of the error model for $\hat{\theta}^{(2)}$ 38

6.5.4 Empirical behavior of proposed estimators 40
6.6 Summary of empirical findings 42

7 Summary **43**

A Proof of Theorem 5.1 **44**

B A useful proposition **46**

C Experimental data **47**
C.1 Registration probability vs. Flow 47
C.2 Imputation error vs. Number of missing vehicles 49
C.3 Imputation error vs. Registered flow 52
C.4 Observed errors under additive error model for $\hat{\theta}^{(2)}$ 54
C.5 Observed errors under multiplicative error model for $\hat{\theta}^{(2)}$ . . . 57
C.6 Error in $\hat{\theta}^{(2)}$ under additive error model vs. Registered flow . . 59
C.7 Error in $\hat{\theta}^{(2)}$ under multiplicative error model vs. Registered flow 62
C.8 ANOVA tables 64
C.8.1 Under the multiplicative imputation error model 64
C.8.2 Under the additive error model for $\hat{\theta}^{(2)}$ 65
C.8.3 Under the multiplicative error model for $\hat{\theta}^{(2)}$ 65

1 Introduction

Every summer since 1996, the Swedish National Road Administration conducts a traffic survey on urban roads. The roads are thought of as partitioned into one-meter road sites, that are the population elements. The primary study variables are the traffic flow, y, and the travel time, z. The traffic flow for a site equals the number of passing vehicles, and the travel time is the total time all vehicles take to pass the site. Data are collected for a random sample of sites (during randomly selected days) by use of a measurement equipment installed on the road. The totals of y and z, t_y and t_z, are known as *total vehicle mileage* and *total travel time*, respectively. The main survey goal is to estimate their ratio, $R = t_y/t_z$, which is interpreted as the *average speed* on the roads.

Typically, the observational data from a selected road site turn out incomplete. If a large amount of data is missing, the site will be re-measured at a later date. The most common situation, and the one of interest here, is however that some data are missing but not to the extent that the measurement is disqualified. Presently, data are then used in the estimation without any special actions taken. We restrict our attention to incompleteness due to occasional loss of vehicles, thus ignoring cases of lost time periods.

In this paper, two strategies for adjusting for missing data in the estimation stage of the survey are discussed. The strategies differ mainly in the way the flow is estimated when some data are missing. One suggestion is to add the number of vehicles automatically imputed by the measurement device to those properly registered; another to weight the registered number of vehicles by use of estimated registration probabilities. In both cases, a direct weighting type of estimator of the travel time is suggested.

2 The missing data problem

The device used to collect data consists of two pneumatic tubes stretched across the road and connected to a traffic analyzer (a simple computer). When a vehicle wheel crosses a tube, its air pressure changes. 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. Missing data arise when arrived pulses can not unambiguously be translated into vehicles. The ambiguousness may be caused for instance

by vehicles simultaneously crossing the tubes (due to meetings or passings) or dense traffic. At present, missing vehicles are ignored in the estimation stage of the survey. This "do nothing" procedure, henceforth referred to as *Strategy 0*, is bound to result in negatively biased estimators of total vehicle mileage and total travel time. The impact on the estimator of average speed is, on the other hand, 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.

We adopt the following notation. The set of vehicles passing road site k (during a selected day) consists of y_k vehicles labeled $v = 1, ..., y_k$. For simplicity, the vth 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. (In practice, the x_v's are calculated as the inverses of the registered vehicle speeds.) The successfully observed subset of U_k is denoted r_k of size n_{r_k}. 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.

3 Proposals for missing data adjustments

Our suggestions for adjustments for missing data in the speed survey involve weighting adjustment and imputation; the two standard techniques for dealing with survey nonresponse. The registration model in Section 3.1 serves as a common starting-point.

3.1 A model of the registration mechanism

The true registration distribution, that 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 still 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 [16, Eq. (8.1)] or [17, 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 [12, Ch. 8]).

The special features of our model, contrasted with the RHG model, are the following. 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 multistage procedure (a full account of the design is given in Section 4.1). 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 r, is summarized below.

The registration model, r
Assume that the vehicles passing road site $k \in s_{iq}$ during a selected day is 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 paper is applicable on any groups of traffic. In our experiment however (see Section 6), we presume partitioning of the traffic by time intervals. The shortest time unit considered is watch-hour. One reason for this is the common advise (see, e.g., [11]) to avoid too small weighting classes when estimating (response) probabilities θ_{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 θ_{kh} by use of the auxiliary variable *measurement efficiency* (ME); the proportion of registered pulses that have been combined into vehicles. The ME 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 r, conditional on s_{iq}, is denoted $E_r(\cdot\,|s_{iq})$ and $V_r(\cdot\,|s_{iq})$, respectively. In Section 5.2.3, 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,

$$\begin{aligned} E_r(\cdot\,|s_{iq}) &= E_{\mathbf{n}_{r_k}} E_r(\cdot\,|s_{iq}, \mathbf{n}_{r_k}) \\ V_r(\cdot\,|s_{iq}) &= E_{\mathbf{n}_{r_k}} V_r(\cdot\,|s_{iq}, \mathbf{n}_{r_k}) + V_{\mathbf{n}_{r_k}} E_r(\cdot\,|s_{iq}, \mathbf{n}_{r_k}) \end{aligned}$$

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

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

1. Under model 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 a stratified simple random (STSI) selection from U_k.

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 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 1970's 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.

3.2.1 Estimator of flow

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

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

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 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 3.2.3.

3.2.2 Estimator of travel time

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

$$\hat{z}_k^{(1)} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{\hat{\theta}_{kh}^{(1)}} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{n_{r_{kh}}/\hat{y}_{kh}^{(1)}} = \sum_{h=1}^{H_k} (n_{r_{kh}} + n_{I_{kh}})\, \bar{x}_{r_{kh}} \tag{2}$$

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 θ_{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* ([16, Eq. (4.10)], [17, Eq. 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 r.

We do however 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.

3.2.3 A model of the imputation mechanism

In [17, Sec. 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 m, 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 ε_{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.

In the frequency interpretation of the simple measurement model, the observed value for an element varies randomly over repeated (hypothetical) measurements performed under identical survey conditions. Our observations, the $n_{I_{kh}}$'s, do not have this random behavior. The traffic passing a site during a given day will always, for a given $n_{r_{kh}}$, result in the same number of imputed vehicles. Our hope is that the random model still serves as a good approximation of the actual imputation procedure.

The imputation model, m
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 $\mu_{(I|r)_{kh}} = E_m(n_{I_{kh}} | s_{iq}, n_{r_{kh}})$ and variance $\sigma^2_{(I|r)_{kh}} = V_m(n_{I_{kh}} | s_{iq}, n_{r_{kh}})$,
- 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 r and m, are, respectively,

$$\begin{aligned} \mu_{I_{kh}} &= E_{rm}(n_{I_{kh}} | s_{iq}) = E_r E_m(n_{I_{kh}} | s_{iq}, n_{r_{kh}}) \\ \sigma^2_{I_{kh}} &= V_{rm}(n_{I_{kh}} | s_{iq}) = E_r V_m(n_{I_{kh}} | s_{iq}, n_{r_{kh}}) + V_r E_m(n_{I_{kh}} | s_{iq}, n_{r_{kh}}) \end{aligned}$$

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 ε_{kh}. The model is further specified in Section 5.2.2.

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.

3.3.1 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 (2) are no longer an option, but other estimates of the registration probabilities are needed.

The possibility to estimate (response) probabilities from auxiliary data is quite sparsely discussed in the literature. The idea is put forward in [3, Sec. 9]; other references include [5], [6] and [4, Sec. 3.5]. In [7], response probabilities are modeled by logistic regression and estimated from the fitted model. Nonparametric estimation methods are discussed for instance in [9].

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 θ_{kh}:

$$\hat{y}_k^{(2)} = \sum_{h=1}^{H_k} \frac{n_{r_{kh}}}{\hat{\theta}_{kh}^{(2)}} = \sum_{h=1}^{H_k} \frac{n_{r_{kh}}}{(ME)_{kh}} = \sum_{h=1}^{H_k} \hat{y}_{kh}^{(2)} \tag{3}$$

In order to evaluate the statistical properties of $\hat{y}_k^{(2)}$, we need to specify the relationship between θ_{kh} and $\hat{\theta}_{kh}^{(2)}$. A model for this relationship is stated in Section 3.3.3.

3.3.2 Estimator of travel time

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

$$\hat{z}_k^{(2)} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{\hat{\theta}_{kh}^{(2)}} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{(ME)_{kh}} \tag{4}$$

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

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

Our error model for $\hat{\theta}_{kh}^{(2)} = (ME)_{kh}$ as estimator of θ_{kh} has very much in common with the imputation model in Section 3.2.3 (and thus also with the simple measurement model in [17, Sec. 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* θ_{kh}, the $(ME)_{kh}$ is viewed as random; or, more precisely, as composed of the true registration probability, θ_{kh}, and a random error ϵ_{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 $\hat{\theta}_{kh}^{(2)}$, q

- The estimator $\hat{\theta}_{kh}^{(2)} = (ME)_{kh}$ of θ_{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 θ_{kh} and ϵ_{kh}. Two possible relationships, the additive and the multiplicative, are considered in Section 5.2.3.

4 Sampling and estimation

In this section, we acquaint ourselves with the sampling design and estimation procedure of the speed survey. The presentation is based on [10, Ch. 2].

4.1 Sampling design

Road sites are selected for observation by means of a multi-stage sampling design. A brief description of the different stages will now be given. To simplify, we ignore the stratification in each stage. We also ignore the fact that in stage one, the three largest units define a take-all stratum.

Consider the following sets of sampling units. 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. 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 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 $q \in s_{IIi_\nu}$.

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

4.2 Estimation with complete data

We here treat the general problem of estimating a population total $t_a = \sum_U a_k$, where a_k is the true value of study variable a (which may be y or z)

for site $k \in U$, in the speed survey. We also consider estimation of a ratio $R = t_y/t_z$.

Define the population totals $t_{aiq} = \sum_{U_{iq}} a_k$ and $t_{ai} = \sum_{U_{IIi}} t_{aiq}$. Further, define $E_k = y_k - Rz_k$ and the corresponding totals $t_{Eiq} = t_{yiq} - Rt_{ziq}$ and $t_{Ei} = t_{yi} - Rt_{zi}$. Estimators of the totals are denoted by a hat. In addition, π estimators (Horvitz-Thompson estimators) are denoted by a π.

In the ideal situation, in which a_k are known for all $k \in s$, the parameter t_a would be estimated by

$$\hat{t}_a = \frac{1}{m_I} \sum_{\nu=1}^{m_I} \frac{\hat{t}_{\pi a i_\nu}}{p_{i_\nu}} \tag{5}$$

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 $\hat{t}_{\pi a i_\nu} = \hat{t}_{\pi a i}$ and $\hat{t}_{\pi a i_\nu q} = \hat{t}_{\pi a i q}$. The estimator of R would be

$$\hat{R} = \hat{t}_y/\hat{t}_z. \tag{6}$$

The randomness in $\hat{t}_a$ and $\hat{R}$ stems solely from the sample selection. Let E_p and V_p denote expectation and variance with respect to the sampling design. 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 the sampling design.

From [17, Result 4.5.1], $\hat{t}_a$ is design-unbiased for t_a (that is, $E_p(\hat{t}_a) = t_a$). The variance of $\hat{t}_a$ is

$$V_p(\hat{t}_a) = \frac{1}{m_I} \sum_{i=1}^{N_I} p_i \left(\frac{t_{ai}}{p_i} - t_a \right)^2 + \frac{1}{m_I} \sum_{i=1}^{N_I} \frac{V_p(\hat{t}_{\pi a i})}{p_i} \tag{7}$$

where

$$V_p(\hat{t}_{\pi a i}) = N_{II i}^2 \frac{1 - f_{II i}}{n_{II i}} S_{t_a U_i}^2 + \frac{N_{II i}}{n_{II i}} \sum_{U_{II i}} N_{iq}^2 \frac{1 - f_{iq}}{n_{iq}} S_{a U_{iq}}^2$$

with the variances

$$\begin{aligned} S_{t_a U_i}^2 &= \frac{1}{N_{II i} - 1} \sum_{U_{II i}} \left(t_{aiq} - \frac{t_{ai}}{N_{II i}} \right)^2 \\ S_{a U_{iq}}^2 &= \frac{1}{N_{iq} - 1} \sum_{U_{iq}} \left(a_k - \frac{t_{aiq}}{N_{iq}} \right)^2 \end{aligned}$$

and the sampling fractions $f_{IIi} = n_{IIi}/N_{IIi}$ and $f_{iq} = n_{iq}/N_{iq}$.

From [15, Sec. 6.8.2.], $\hat{R}$ is approximately design-unbiased for R. The approximate variance of $\hat{R}$ is

$$AV_p\left(\hat{R}\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{t_{Ei}^2}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_p\left(\hat{t}_{\pi Ei}\right)}{p_i}\right\} \tag{8}$$

where $V_p\left(\hat{t}_{\pi Ei}\right)$ is obtained from $V_p\left(\hat{t}_{\pi ai}\right)$ by replacing a_k with E_k.

5 Estimation with missing data

A more realistic situation than the one dealt with in Section 4.2 is, that some observational data are missing. Then, the true a_k's are unknown. 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* ξ. The estimator obtained by replacing a by $\hat{a}^{(c)}$ in $\hat{t}_a$ is denoted $\hat{t}_{\hat{a}^{(c)}}$. Expectations and variances taken with respect to the model ξ are indicated by subscript ξ. In order to shorten the formulas, we denote $E_\xi\left(\hat{a}_k^{(c)}\,|s_{iq}\right)$ and $V_\xi\left(\hat{a}_k^{(c)}\,|s_{iq}\right)$ by $\gamma\left(\hat{a}^{(c)}\right)_k$ and $\delta\left(\hat{a}^{(c)}\right)_k$, respectively. The population entities $t_{\gamma\left(\hat{a}^{(c)}\right)iq}$, $t_{\gamma\left(\hat{a}^{(c)}\right)i}$, $t_{\gamma\left(\hat{a}^{(c)}\right)}$, $S^2_{t_{\gamma\left(\hat{a}^{(c)}\right)}U_i}$ and $S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{iq}}$ for $\gamma\left(\hat{a}^{(c)}\right)$ are defined in the same manner as the corresponding entities for a in Section 4.2.

5.1 General results

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

Theorem 5.1 *Jointly under the sampling design p in Section 4.1 and the model ξ, the expected value of $\hat{t}_{\hat{a}^{(c)}}$ is given by*

$$E_{p\xi}\left(\hat{t}_{\hat{a}^{(c)}}\right) = \sum_{i=1}^{N_I} t_{\gamma\left(\hat{a}^{(c)}\right)i} = t_{\gamma\left(\hat{a}^{(c)}\right)} \tag{9}$$

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

$$V_{p\xi}\left(\hat{t}_{\hat{a}^{(c)}}\right) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_i\left(\frac{t_{\gamma\left(\hat{a}^{(c)}\right)i}}{p_i} - t_{\gamma\left(\hat{a}^{(c)}\right)}\right)^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{p\xi}\left(\hat{t}_{\pi\hat{a}^{(c)}i}\right)}{p_i} \tag{10}$$

where

$$V_{p\xi}\left(\hat{t}_{\pi\hat{a}^{(c)}i}\right) = N_{IIi}^2\frac{1-f_{IIi}}{n_{IIi}}S^2_{t_{\gamma\left(\hat{a}^{(c)}\right)}U_i} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}} N_{iq}^2\frac{1-f_{iq}}{n_{iq}}S^2_{\gamma\left(\hat{a}^{(c)}\right)U_{iq}} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{a}^{(c)}\right)_k .$$

The proof of Theorem 5.1 is given in Appendix A.

From Theorem 5.1, the bias of $\hat{t}_{\hat{a}^{(c)}}$ as estimator of t_a is

$$E_{p\xi}\left(\hat{t}_{\hat{a}^{(c)}}\right) - t_a = \sum\nolimits_{i=1}^{N_I}\sum\nolimits_{U_{IIi}}\sum\nolimits_{U_{iq}}\left(\gamma\left(\hat{a}^{(c)}\right)_k - a_k\right). \tag{11}$$

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 $\hat{t}_a$, $V_{p\xi}\left(\hat{t}_{\hat{a}^{(c)}}\right) - V_p\left(\hat{t}_{\pi a}\right)$.

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

Corollary 5.1 *Assume that $\gamma\left(\hat{a}^{(c)}\right)_k$ 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*

$$V_{p\xi}\left(\hat{t}_{\hat{a}^{(c)}}\right) - V_p\left(\hat{t}_a\right) = \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{a}^{(c)}\right)_k . \tag{12}$$

If data are missing, the estimator of R under Strategy c $(c = 0, 1, 2)$ is $\hat{R}^{(c)} = \hat{t}_{\hat{y}^{(c)}}/\hat{t}_{\hat{z}^{(c)}}$. Its statistical properties are investigated in Theorem 5.2.

Theorem 5.2 *Jointly under the sampling design p in Section 4.1 and the model ξ, the estimator $\hat{R}^{(c)}$ is approximately unbiased for*

$$R^{(c)} = \frac{E_{p\xi}\left(\hat{t}_{\hat{y}^{(c)}}\right)}{E_{p\xi}\left(\hat{t}_{\hat{z}^{(c)}}\right)} = \frac{t_{\gamma\left(\hat{y}^{(c)}\right)}}{t_{\gamma\left(\hat{z}^{(c)}\right)}} \tag{13}$$

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

$$AV_{p\xi}\left(\hat{R}^{(c)}\right) = \frac{1}{t_z^2}\left\{\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{t^2_{\gamma\left(\hat{E}^{(c)}\right)i}}{p_i} + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{p\xi}\left(\hat{t}_{\pi\hat{E}^{(c)}i}\right)}{p_i}\right\} \tag{14}$$

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

The proof of Theorem 5.2 follows by a slight generalization of the results in [15, Sec. 6.8.2.].

From Theorem 5.2, 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 complete-data 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 5.2 *Assume that* $\gamma\big(\hat{y}^{(c)}\big)_k = y_k$ *and* $\gamma\big(\hat{z}^{(c)}\big)_k = z_k$ $(k \in s)$. *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*

$$\begin{aligned} &AV_{p\xi}\Big(\hat{R}^{(c)}\Big) - AV_p\Big(\hat{R}\Big) \\ &= \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\Big(\hat{E}^{(c)}\Big)_k \end{aligned} \tag{15}$$

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

5.2 Results for present and proposed estimators

In Section 5.1, the general statistical properties of $\hat{t}_{\hat{y}^{(c)}}$, $\hat{t}_{\hat{z}^{(c)}}$ and $\hat{R}^{(c)}$ were derived. Here, specific results for the estimators under Strategy 0-2 are derived. This implies presenting the explicit γ and δ 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.

5.2.1 Strategy 0

Model ξ in Section 5.1 is here interpreted as the registration model r. When applying Theorem 5.1 on the estimators $\hat{t}_{\hat{y}^{(0)}}$ and $\hat{t}_{\hat{z}^{(0)}}$, we use

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

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

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

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

$$\gamma\big(\hat{z}^{(0)}\big)_k = \frac{z_k}{y_k}\mu_{r_k} \tag{16}$$

$$\delta\big(\hat{z}^{(0)}\big)_k = E_r\left(n^2_{r_k}\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}\,|s_{iq}\right) + \left(\frac{z_k}{y_k}\right)^2\sigma^2_{r_k} \tag{17}$$

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 (17) simplifies to

$$E_r\left(n^2_{r_k}\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}\,|s_{iq}\right) = \left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right]S^2_{xU_k}. \tag{18}$$

Equations (16)-(17) are derived by use of Proposition B.1 in Appendix B with $(A, B) = (n_{r_k}, \bar{x}_{r_k})$. We also use the fact that

$$E_r(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) = \frac{z_k}{y_k} \tag{19}$$

$$V_r(\bar{x}_{r_k}\,|n_{r_k}, s_{iq}) = \frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k} \tag{20}$$

which follows from implication number 2 of the registration model in Section 3.1. (The moments in Equations (19)-(20) are in fact conditional also on the event that $n_{r_k} \geq 1$. For details, see [17, Section 7.10.1].)

When applying Theorem 5.2 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

$$\gamma\left(\hat{E}^{(0)}\right)_k = \left(1 - R^{(0)}\frac{z_k}{y_k}\right)\mu_{r_k} \tag{21}$$

$$\delta\left(\hat{E}^{(0)}\right)_k = \left(R^{(0)}\right)^2\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right]S^2_{xU_k} + \left(1 - R^{(0)}\frac{z_k}{y_k}\right)^2\sigma^2_{r_k} \tag{22}$$

Equations (21)-(22) are derived by use of Proposition B.1 with $(A, B) = \left(n_{r_k}, 1 - R^{(0)}\bar{x}_{r_k}\right)$ and by applying Equation (18).

We now make an attempt to simplify the results for the Strategy 0 estimators. From implication number 1 of the registration model in Section 3.1, $\mu_{r_k} = y_k\theta_k$ and $\sigma^2_{r_k} = y_k\theta_k\left(1-\theta_k\right)$. It follows that for $\hat{t}_{\hat{y}^{(0)}}$, the model moments are $\gamma\left(\hat{y}^{(0)}\right)_k = y_k\theta_k$ and $\delta\left(\hat{y}^{(0)}\right)_k = y_k\theta_k\left(1-\theta_k\right)$; for $\hat{t}_{\hat{z}^{(0)}}$, they are

$$\begin{aligned}
\gamma\left(\hat{z}^{(0)}\right)_k &= z_k\theta_k && (23)\\
\delta\left(\hat{z}^{(0)}\right)_k &= \theta_k\left(1-\theta_k\right)\left[\frac{z_k^2}{y_k} + \left(y_k - 1\right)S^2_{xU_k}\right] \\
&= \theta_k\left(1-\theta_k\right)\sum\nolimits_{U_k} x_v^2 && (24)
\end{aligned}$$

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

$$\begin{aligned}
\gamma\left(\hat{E}^{(0)}\right)_k &= \theta_k\left(y_k - R^{(0)}z_k\right) && (25)\\
\delta\left(\hat{E}^{(0)}\right)_k &= \theta_k\left(1-\theta_k\right)\left[\left(R^{(0)}\right)^2\left(y_k-1\right)S^2_{xU_k} + \left(1 - R^{(0)}\frac{z_k}{y_k}\right)^2 y_k\right] \\
&= \theta_k\left(1-\theta_k\right)\left[\left(R^{(0)}\right)^2\sum\nolimits_{U_k} x_v^2 - 2R^{(0)}z_k + y_k\right] && (26)
\end{aligned}$$

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

$$\begin{aligned}
t^2_{\gamma\left(\hat{E}^{(0)}\right)i} &= \theta^2 t^2_{Ei} && (27)\\
V_{p\xi}\left(\hat{t}_{\pi\hat{E}^{(0)}i}\right) &= N^2_{IIi}\frac{1-f_{IIi}}{n_{IIi}}\theta^2 S^2_{t_E U_i} + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}} N^2_{iq}\frac{1-f_{iq}}{n_{iq}}\theta^2 S^2_{EU_{iq}} \\
&\quad + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{E}^{(0)}\right)_k \\
&= \theta^2 V_p\left(\hat{t}_{\pi Ei}\right) + \frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\delta\left(\hat{E}^{(0)}\right)_k && (28)
\end{aligned}$$

By insertion of Equations (27) and (28) into Equation (14), and comparison

of the resulting variance expression with the one in Equation (8), we see that

$$\begin{aligned} & AV_{pr}\left(\hat{R}^{(0)}\right) \\ & = \theta^2 AV_p\left(\hat{R}\right) \\ & + \theta\left(1-\theta\right)\frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{U_{iq}}\left(R^2\sum\nolimits_{U_k}x_v^2-2Rz_k+y_k\right) \end{aligned} \tag{29}$$

5.2.2 Strategy 1

Expectations and variances with respect to model ξ in Section 5.1 are here interpreted as taken with respect jointly to the registration model r and the imputation model m. When applying Theorem 5.1 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 5.2 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 γ and δ expressions will first be presented for the general imputation model m in Section 3.2.3 ("general" in the sense that it does not say how n_{I_k} is connected with $y_k - n_{r_k}$ and ε_k); then for a more specified model.

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

$$\gamma\left(\hat{y}^{(1)}\right)_k = \mu_{r_k} + \mu_{I_k} \tag{30}$$

$$\delta\left(\hat{y}^{(1)}\right)_k = \sigma_{r_k}^2 + \sigma_{I_k}^2 + 2Cov_r\left(n_{r_k}, \mu_{(I|r)_k}\,|s_{iq}\right) \tag{31}$$

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

$$\begin{aligned}
\gamma\big(\hat{z}^{(1)}\big)_k &= \frac{z_k}{y_k}\gamma\big(\hat{y}^{(1)}\big)_k \qquad (32)\\
\delta\big(\hat{z}^{(1)}\big)_k &= E_{rm}\left[\left(n_{r_k}+n_{I_k}\right)^2\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}|s_{iq}\right]\\
&\quad+\left(\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k\\
&= E_r\left\{\left[\left(n_{r_k}+\mu_{(I|r)_k}\right)^2+\sigma^2_{(I|r)_k}\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}|s_{iq}\right\}\\
&\quad+\left(\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k. \qquad (33)
\end{aligned}$$

Equations (32)-(33) are derived by use of Proposition B.1 with $(A,B)=(n_{r_k}+n_{I_k},\bar{x}_{r_k})$, and the equalities

$$\begin{aligned}
E_{rm}(\bar{x}_{r_k}|n_{r_k}+n_{I_k},s_{iq}) &= E_mE_r(\bar{x}_{r_k}|n_{r_k}+n_{I_k},n_{I_k},s_{iq})\\
&= E_mE_r(\bar{x}_{r_k}|n_{r_k},s_{iq})=E_r(\bar{x}_{r_k}|n_{r_k},s_{iq})\\
V_{rm}(\bar{x}_{r_k}|n_{r_k}+n_{I_k},s_{iq}) &= E_mV_r(\bar{x}_{r_k}|n_{r_k}+n_{I_k},n_{I_k},s_{iq})\\
&\quad+V_mE_r(\bar{x}_{r_k}|n_{r_k}+n_{I_k},n_{I_k},s_{iq})\\
&= E_mV_r(\bar{x}_{r_k}|n_{r_k},s_{iq})+V_mE_r(\bar{x}_{r_k}|n_{r_k},s_{iq})\\
&= V_r(\bar{x}_{r_k}|n_{r_k},s_{iq})
\end{aligned}$$

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

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

$$\begin{aligned}
\gamma\Big(\hat{E}^{(1)}\Big)_k &= \left(1-R^{(1)}\frac{z_k}{y_k}\right)\gamma\big(\hat{y}^{(1)}\big)_k \qquad (34)\\
\delta\Big(\hat{E}^{(1)}\Big)_k &= \big(R^{(1)}\big)^2E_{rm}\left[\left(n_{r_k}+n_{I_k}\right)^2\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}|s_{iq}\right]\\
&\quad+\left(1-R^{(1)}\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k\\
&= \big(R^{(1)}\big)^2E_r\left\{\left[\left(n_{r_k}+\mu_{(I|r)_k}\right)^2+\sigma^2_{(I|r)_k}\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S^2_{xU_k}|s_{iq}\right\}\\
&\quad+\left(1-R^{(1)}\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k. \qquad (35)
\end{aligned}$$

Equations (34)-(35) are obtained by use of Proposition B.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 5.1, $\hat{t}_{\hat{y}^{(1)}}$ and $\hat{t}_{\hat{z}^{(1)}}$ are unbiased for t_y and t_z, respectively, and from Corollary 5.2, $\hat{R}^{(1)}$ 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:

$$n_{I_k} = (y_k - n_{r_k})\,\varepsilon_k. \tag{36}$$

Let the conditional mean and variance of ε_k given s_{iq} and n_{r_k} be denoted $\mu_\varepsilon = E_m(\varepsilon_k\,|s_{iq}, n_{r_k})$ and $\sigma^2_\varepsilon = V_m(\varepsilon_k\,|s_{iq}, n_{r_k})$, respectively. As the notation suggests, the conditional moments μ_ε and σ^2_ε 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,

$$\mu_{(I|r)_k} = (y_k - n_{r_k})\,\mu_\varepsilon \tag{37}$$

$$\sigma^2_{(I|r)_k} = (y_k - n_{r_k})^2\,\sigma^2_\varepsilon \tag{38}$$

and

$$\mu_{I_k} = E_r[(y_k - n_{r_k})\,\mu_\varepsilon\,|s_{iq}] = \big(y_k - \mu_{r_k}\big)\,\mu_\varepsilon \tag{39}$$

$$\begin{aligned}\sigma^2_{I_k} &= E_r\big[(y_k - n_{r_k})^2\,\sigma^2_\varepsilon\,|s_{iq}\big] + V_r[(y_k - n_{r_k})\,\mu_\varepsilon\,|s_{iq}] \\ &= \Big[\big(y_k - \mu_{r_k}\big)^2 + \sigma^2_{r_k}\Big]\,\sigma^2_\varepsilon + \sigma^2_{r_k}\mu^2_\varepsilon \end{aligned} \tag{40}$$

We now modify the γ and δ expressions presented earlier (Equations (30)-(35)) in compliance with Equations (37)-(40). The resulting model moments for $\hat{t}_{\hat{y}^{(1)}}$ are

$$\gamma\big(\hat{y}^{(1)}\big)_k = \mu_{r_k}\,(1 - \mu_\varepsilon) + y_k\mu_\varepsilon \tag{41}$$

$$\delta\big(\hat{y}^{(1)}\big)_k = \sigma^2_{r_k}\left[(1 - \mu_\varepsilon)^2 + \sigma^2_\varepsilon\right] + \big(y_k - \mu_{r_k}\big)^2\,\sigma^2_\varepsilon \tag{42}$$

In the derivation of Equation (42), we use the fact that

$$Cov_r[n_{r_k}, (y_k - n_{r_k})\,\mu_\varepsilon\,|s_{iq}] = -\mu_\varepsilon V_r(n_{r_k}\,|s_{iq}) = -\mu_\varepsilon \sigma^2_{r_k}.$$

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

$$\gamma\big(\hat{z}^{(1)}\big)_k = \frac{z_k}{y_k}\gamma\big(\hat{y}^{(1)}\big)_k \tag{43}$$

$$\begin{aligned}\delta\big(\hat{z}^{(1)}\big)_k &= E_r\Big\{\big[(n_{r_k}(1-\mu_\varepsilon) + y_k\mu_\varepsilon)^2 + (y_k - n_{r_k})^2\,\sigma^2_\varepsilon\big]\cdot \\ &\quad \cdot \frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k}\,|s_{iq}\Big\} + \left(\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(1)}\big)_k\end{aligned} \tag{44}$$

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

$$\gamma\Big(\hat{E}^{(1)}\Big)_k = \left(1 - R^{(1)}\frac{z_k}{y_k}\right)\gamma\big(\hat{y}^{(1)}\big)_k \tag{45}$$

$$\begin{aligned}\delta\Big(\hat{E}^{(1)}\Big)_k &= \big(R^{(1)}\big)^2 E_r\Big\{\big[(n_{r_k}(1-\mu_\varepsilon) + y_k\mu_\varepsilon)^2 + (y_k - n_{r_k})^2\,\sigma^2_\varepsilon\big]\cdot \\ &\quad \cdot \frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k}\,|s_{iq}\Big\} + \left(1 - R^{(1)}\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(1)}\big)_k\end{aligned} \tag{46}$$

with

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

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 δ expressions for $\hat{t}_{\hat{y}^{(1)}}$ and $\hat{t}_{\hat{z}^{(1)}}$ (to be inserted in Equation (12)) are

$$\delta\big(\hat{y}^{(1)}\big)_k = \Big[\big(y_k - \mu_{r_k}\big)^2 + \sigma^2_{r_k}\Big]\,\sigma^2_\varepsilon \tag{47}$$

and

$$\begin{aligned}\delta\big(\hat{z}^{(1)}\big)_k &= E_r\left\{\left[y_k^2 + (y_k - n_{r_k})^2\,\sigma^2_\varepsilon\right]\frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k}\,|s_{iq}\right\} \\ &\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(1)}\big)_k\end{aligned} \tag{48}$$

respectively. The δ expression for $\hat{R}^{(1)}$ (to be inserted in Equation (15)) is

$$\begin{aligned}\delta\Big(\hat{E}^{(1)}\Big)_k &= R^2 E_r\left\{\left[y_k^2+(y_k-n_{r_k})^2\,\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\}\\ &\quad+\left(1-R\frac{z_k}{y_k}\right)^2\delta\big(\hat{y}^{(1)}\big)_k\,. \end{aligned}\tag{49}$$

The expectation occurring in Equations (48) and (49) can of course be worked out. Some straight-forward algebra gives:

$$\begin{aligned}&E_r\left\{\left[y_k^2+(y_k-n_{r_k})^2\,\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\}\\ &=\left\{y_k\left(1+\sigma_\varepsilon^2\right)\left[y_kE_r\left(\frac{1}{n_{r_k}}\,|s_{iq}\right)-1\right]+\sigma_\varepsilon^2\left(3\mu_{r_k}-2y_k-\frac{\sigma_{r_k}^2+\mu_{r_k}^2}{y_k}\right)\right\}S_{xU_k}^2\\ &\approx\left\{y_k\left(1+\sigma_\varepsilon^2\right)\left(\frac{y_k}{\mu_{r_k}}-1\right)+\sigma_\varepsilon^2\left(3\mu_{r_k}-2y_k-\frac{\sigma_{r_k}^2+\mu_{r_k}^2}{y_k}\right)\right\}S_{xU_k}^2\end{aligned}\tag{50}$$

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

Can Equations (47) and (50) be additionally simplified? From implication number 1 of the registration model in Section 3.1, $\mu_{r_k} = y_k\theta_k$ and $\sigma_{r_k}^2 = y_k\theta_k\,(1-\theta_k)$. Insertion in Equation (47) gives

$$\delta\big(\hat{y}^{(1)}\big)_k = y_k\,(1-\theta_k)\,[\theta_k\,(1-y_k)+y_k]\,\sigma_\varepsilon^2\tag{51}$$

and in Equation (50)

$$\begin{aligned}&E_r\left\{\left[y_k^2+(y_k-n_{r_k})^2\,\sigma_\varepsilon^2\right]\frac{1-n_{r_k}/y_k}{n_{r_k}}S_{xU_k}^2\,|s_{iq}\right\}\\ &\approx\left\{y_k\left[\left(\frac{1}{\theta_k}-1\right)+\sigma_\varepsilon^2\left(\frac{1}{\theta_k}-3+3\theta_k-\theta_k^2\right)\right]-\sigma_\varepsilon^2\theta_k\,(1-\theta_k)\right\}S_{xU_k}^2.\end{aligned}\tag{52}$$

5.2.3 Strategy 2

Model ξ in Section 5.1 is here interpreted as taken with respect jointly to the registration model r and the error model q. When applying Theorem 5.1 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 5.2 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 γ and δ expressions will first be presented by use of the general error model q in Section 3.3.3 ("general" in the sense that it does not say how $\hat{\theta}_k^{(2)}$ is connected with θ_k and ϵ_k), then two special cases will be treated.

Under general error model assumptions The estimator $\hat{y}_k^{(2)}$ is theoretically complicated, it being a ratio of random variables. By use of Taylor's theorem (see, e.g., [2, 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

$$\gamma\left(\hat{y}^{(2)}\right)_k \approx \frac{\mu_{r_k}}{\mu_{\hat{\theta}_k^{(2)}}} \tag{53}$$

$$\delta\left(\hat{y}^{(2)}\right)_k \approx \left(\frac{\mu_{r_k}}{\mu_{\hat{\theta}_k^{(2)}}}\right)^2 \left(\frac{\sigma_{r_k}^2}{\mu_{r_k}^2} + \frac{\sigma_{\hat{\theta}_k^{(2)}}^2}{\mu_{\hat{\theta}_k^{(2)}}^2}\right) \tag{54}$$

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

$$\begin{aligned}
E_{rq}\left(\bar{x}_{r_k} \middle| n_{r_k}/\hat{\theta}_k^{(2)}, s_{iq}\right) &= E_q E_r\left(\bar{x}_{r_k} \middle| n_{r_k}/\hat{\theta}_k^{(2)}, \hat{\theta}_k^{(2)}, s_{iq}\right) \\
&= E_q E_r(\bar{x}_{r_k} | n_{r_k}, s_{iq}) = E_r(\bar{x}_{r_k} | n_{r_k}, s_{iq}) \\
V_{rq}\left(\bar{x}_{r_k} \middle| n_{r_k}/\hat{\theta}_k^{(2)}, s_{iq}\right) &= E_q V_r\left(\bar{x}_{r_k} \middle| n_{r_k}/\hat{\theta}_k^{(2)}, \hat{\theta}_k^{(2)}, s_{iq}\right) \\
&\quad + V_q E_r\left(\bar{x}_{r_k} \middle| n_{r_k}/\hat{\theta}_k^{(2)}, \hat{\theta}_k^{(2)}, s_{iq}\right) \\
&= E_q V_r(\bar{x}_{r_k} | n_{r_k}, s_{iq}) + V_q E_r(\bar{x}_{r_k} | n_{r_k}, s_{iq}) \\
&= V_r(\bar{x}_{r_k} | n_{r_k}, s_{iq})
\end{aligned}$$

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

$$\begin{aligned}
\gamma\left(\hat{z}^{(2)}\right)_k &\approx \frac{z_k}{y_k}\frac{\mu_{r_k}}{\mu_{\hat{\theta}_k^{(2)}}} \qquad (55)\\
\delta\left(\hat{z}^{(2)}\right)_k &\approx E_{rq}\left[\left(\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\right)^2 \frac{1-n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k}\,|s_{iq}\right] \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\
&= E_q\left[\frac{1}{\left(\hat{\theta}_k^{(2)}\right)^2}\right]\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k}+\sigma^2_{r_k}\right)\right] S^2_{xU_k} \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\
&\approx \frac{1}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}}\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k}+\sigma^2_{r_k}\right)\right] S^2_{xU_k} \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \qquad (56)
\end{aligned}$$

The second equality for δ is derived using the independency of n_{r_k} and $\hat{\theta}_k^{(2)}$, and Equation (18). The final equality arises from the (first-order) Taylor approximation

$$E_q\left[\frac{1}{\left(\hat{\theta}_k^{(2)}\right)^2}\right] \approx \frac{1}{E_q\left[\left(\hat{\theta}_k^{(2)}\right)^2\right]}. \qquad (57)$$

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

$$\begin{aligned}
\gamma\left(\hat{E}^{(2)}\right)_k &\approx \left(1 - R^{(2)}\frac{z_k}{y_k}\right)\gamma\left(\hat{y}^{(2)}\right)_k \qquad (58)\\
\delta\left(\hat{E}^{(2)}\right)_k &\approx \left(R^{(2)}\right)^2 E_{rq}\left[\left(\frac{n_{r_k}}{\hat{\theta}_k^{(2)}}\right)^2 \frac{1 - n_{r_k}/y_k}{n_{r_k}} S^2_{xU_k}\,|s_{iq}\right]\\
&\quad + \left(1 - R^{(2)}\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k\\
&= \left(R^{(2)}\right)^2 E_q\left[\frac{1}{\left(\hat{\theta}_k^{(2)}\right)^2}\right]\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right] S^2_{xU_k}\\
&\quad + \left(1 - R^{(2)}\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k\\
&\approx \frac{\left(R^{(2)}\right)^2}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}}\left[\mu_{r_k} - \frac{1}{y_k}\left(\mu^2_{r_k} + \sigma^2_{r_k}\right)\right] S^2_{xU_k}\\
&\quad + \left(1 - R^{(2)}\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \qquad (59)
\end{aligned}$$

Equations (58)-(59) are derived by use of Proposition B.1 with $(A, B) = \left(n_{r_k}/\hat{\theta}_k^{(2)}, 1 - R^{(2)}\bar{x}_{r_k}\right)$, and (for δ) the independency of n_{r_k} and $\hat{\theta}_k^{(2)}$, Equation (18), and the approximation in Equation (57).

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 (53)-(54) simplify to

$$\begin{aligned}
\gamma\left(\hat{y}^{(2)}\right)_k &= y_k\frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}} \qquad (60)\\
\delta\left(\hat{y}^{(2)}\right)_k &= y_k^2\left(\frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}}\right)^2\left(\frac{1 - \theta_k}{y_k\theta_k} + \frac{\sigma^2_{\hat{\theta}_k^{(2)}}}{\mu^2_{\hat{\theta}_k^{(2)}}}\right); \qquad (61)
\end{aligned}$$

Equations (55)-(56) to

$$\begin{aligned}
\gamma\left(\hat{z}^{(2)}\right)_k &\approx z_k \frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}} && (62)\\
\delta\left(\hat{z}^{(2)}\right)_k &\approx \frac{1}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \theta_k \left(1-\theta_k\right)\left(y_k - 1\right) S^2_{xU_k} \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k \\
&= \frac{1}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \theta_k \left(1-\theta_k\right)\left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) \\
&\quad + \left(\frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k ; && (63)
\end{aligned}$$

and Equations (58)-(59) to

$$\begin{aligned}
\gamma\left(\hat{E}^{(2)}\right)_k &\approx y_k \left(1 - R^{(2)} \frac{z_k}{y_k}\right) \frac{\theta_k}{\mu_{\hat{\theta}_k^{(2)}}} && (64)\\
\delta\left(\hat{E}^{(2)}\right)_k &\approx \frac{\left(R^{(2)}\right)^2}{\sigma^2_{\hat{\theta}_k^{(2)}} + \mu^2_{\hat{\theta}_k^{(2)}}} \theta_k \left(1-\theta_k\right)\left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) \\
&\quad + \left(1 - R^{(2)} \frac{z_k}{y_k}\right)^2 \delta\left(\hat{y}^{(2)}\right)_k && (65)
\end{aligned}$$

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 (60) and (62), $\gamma\left(\hat{y}^{(2)}\right)_k = y_k$ and $\gamma\left(\hat{z}^{(2)}\right)_k = z_k$. It follows from Corollary 5.1 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 5.2, $\hat{R}^{(2)}$ is approximately unbiased for R.

Under some special cases of the error model for $\hat{\theta}_k^{(2)}$ Consider the error model q for $\hat{\theta}_k^{(2)}$ stated in Section 3.3.3. Two possible functional relationships between $\hat{\theta}_k^{(2)}$ and θ_k are the additive error model,

$$\hat{\theta}_k^{(2)} = \theta_k + \epsilon_k \qquad (66)$$

and the multiplicative error model,

$$\hat{\theta}_k^{(2)} = \theta_k \epsilon_k. \tag{67}$$

Let the mean and variance of ϵ_k be denoted μ_ϵ and σ_ϵ^2, respectively. According to model q, 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 (66)-(67), by letting $\mu_{\hat{\theta}_k^{(2)}} = \theta_k + \mu_\epsilon$ and $\sigma^2_{\hat{\theta}_k^{(2)}} = \sigma_\epsilon^2$, results are obtained for the additive model. In the same manner, by letting $\mu_{\hat{\theta}_k^{(2)}} = \theta_k \mu_\epsilon$ and $\sigma^2_{\hat{\theta}_k^{(2)}} = \theta_k^2 \sigma_\epsilon^2$, we get results for the multiplicative model. Consider in particular the latter model. For this, Equations (60)-(61) modify to

$$\gamma\big(\hat{y}^{(2)}\big)_k = \frac{y_k}{\mu_\epsilon} \tag{68}$$

$$\delta\big(\hat{y}^{(2)}\big)_k = \left(\frac{y_k}{\mu_\epsilon}\right)^2 \left(\frac{1-\theta_k}{y_k \theta_k} + \frac{\sigma_\epsilon^2}{\mu_\epsilon^2}\right); \tag{69}$$

Equations (62)-(63) to

$$\gamma\big(\hat{z}^{(2)}\big)_k \approx \frac{z_k}{\mu_\epsilon} \tag{70}$$

$$\delta\big(\hat{z}^{(2)}\big)_k \approx \frac{1}{\sigma_\epsilon^2 + \mu_\epsilon^2} \frac{1-\theta_k}{\theta_k} \left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) + \left(\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(2)}\big)_k \tag{71}$$

and Equations (64)-(65) to

$$\gamma\Big(\hat{E}^{(2)}\Big)_k \approx \frac{y_k}{\mu_\epsilon}\left(1 - R\frac{z_k}{y_k}\right) \tag{72}$$

$$\delta\Big(\hat{E}^{(2)}\Big)_k \approx \frac{R^2}{\sigma_\epsilon^2 + \mu_\epsilon^2} \frac{1-\theta_k}{\theta_k} \left(\sum\nolimits_{U_k} x_v^2 - \frac{z_k^2}{y_k}\right) + \left(1 - R^2\frac{z_k}{y_k}\right)^2 \delta\big(\hat{y}^{(2)}\big)_k \tag{73}$$

For the multiplicative model, the propitious case $\mu_{\hat{\theta}_k^{(2)}} = \theta_k$, 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 ($\mu_\epsilon = 1$).

5.3 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 the one of the corresponding complete-data estimator (the estimator based on the true variable values) 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 complete-data estimators.

The statistical properties of the Strategy 0 estimators are determined jointly by the sampling design and the registration model. Under this strategy, the values of both y and z are 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 θ_k are equal for all sites, the Strategy 0 estimator of R is however not biased by missing data.

Evaluation of the statistical properties of the Strategy 1 estimators requires not only that the sampling design and the registration model are taken into consideration, but also the imputation model. 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 θ_k as well as the error variance.

Finally consider the results for Strategy 2. Under this strategy, estimates of the registration probabilities θ_k are used to adjust for missing data. Expectations and variances of the Strategy 2 estimators are taken with respect jointly to the sampling design, the registration model and the error model for the estimator of θ_k. Due to the fact that the y values are estimated by ratios of random variables, our results are not exact, but rely on Taylor approximations. The estimators of y and z are unbiased if the estimator of θ_k is. If the error in the estimator of θ_k is multiplicative, like under Strategy 1, this allows us to simplify the variance expressions somewhat.

6 Empirical study

6.1 Study objectives

In Section 5, the statistical properties of the various estimators are investigated. The results however rely on model assumptions whose realism remains to be checked. Also, the results do 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 motivates the collection of some empirical data.

The main objectives of our study are to investigate:

- The **forming of registration homogeneity groups**. For reasons stated in Section 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 this way be avoided?

- The assumptions of the **multiplicative imputation error model**. Is the error in the number of imputed vehicles multiplicative (as suggested in Section 5.2.2)? 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 θ_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.

6.2 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

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

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 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 equipment is outlined in Figure 1. One pair of tubes (A_0, B_0) connected to a traffic analyzer M_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 the 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 center line 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 1, lane 1 is measured by the tube parts (A_{11}, B_{11}) connected to traffic analyzer M_1; lane 2 by (A_{12}, B_{12}) connected to traffic analyzer M_2.

The plugging method has been developed at the SNRA as a means of improving data quality. The registration task facing M_1 and M_2 is much easier (and hence less subject to measurement errors) than that of M_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.

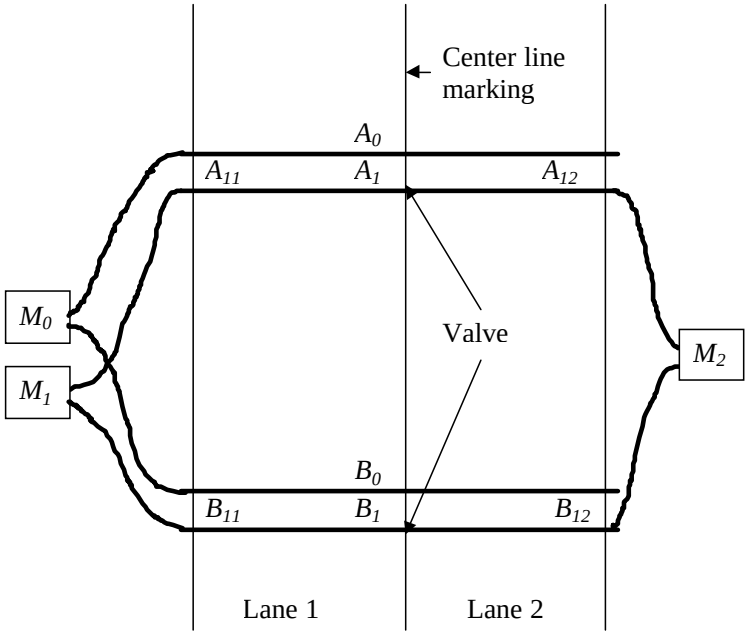


Figure 1: Installation of the measurement equipment.

According to plan, all experimental data were to be collected August 21-22, 2001. Due to valve malfunctioning, site 1, 3 and 4 were however re-measured September 24-25, 2001.

In the experiment, the data set produced by M_0 is intended to represent the output one would expect from a measurement performed within the regular survey. The data set produced jointly by M_1 and M_2, on the other hand, is intended to represent the 'truth'.

6.3 Data processing

We start by introducing some notation. Consider site k during hour h as measured by traffic analyzer $M_d; k = 1, ..., 5; h = 1, ..., 24; d = 0, 1, 2$. For (k, h, M_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 measures for a 24-hour period are $n_{r_{k(d)}} = \sum_{h=1}^{24} n_{r_{kh(d)}}$, $n_{I_{k(d)}} = \sum_{h=1}^{24} n_{I_{kh(d)}}$ and $(ME)_{k(d)} = \sum_{h=1}^{24} (ME)_{kh(d)} /24$.

The observational data from M_0 is 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 $n_{r_{k(1+2)}} =$

	M_0			M_1			M_2		
Site no.	n_r	n_I	ME	n_r	n_I	ME	n_r	n_I	ME
1	5690	69	98.9	2976	13	99.5	2763	13	99.5
2	14314	747	96.5	7924	249	97.9	6989	58	96.9
3	10850	2856	87.3	5772	2038	84.8	6546	527	94.3
4	10948	181	98.4	5363	47	98.6	5730	80	98.5
5	11259	338	97.9	5660	8	99.8	5907	66	98.4

Table 2: The number of registered vehicles (n_r) and imputed vehicles (n_I), and the measurement efficiency (ME) in per cent, by site and traffic analyzer.

$\sum_{h=1}^{24} n_{r_{kh(1+2)}}$, 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 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 2. If the data collection had turned out perfectly, the table had contained nothing but zeroes in the n_I columns for analyzer M_1 and M_2 (the measurement efficiencies for M_1 and M_2 had then also been 100 per cent.) Table 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 is approaching 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 measurement efficiency, makes the M_1 data useless for our purposes. For this reason, only the lane 2 part of the M_0 data, and the M_2 data, is used in the coming analysis of site 3.

In certain cases, imputations in the M_1 or M_2 data can be matched with vehicles properly registered by M_0. These situations are most likely to occur when passing vehicles straddle the valves. For each site, we compared the data files from M_0, M_1 and M_2, looking for imputations in M_1 and M_2 which with reasonable certainty could be matched with registered vehicles in M_0. These imputations were then substituted by the registered vehicles. Table 3 shows the number of imputed vehicles that were substituted, how many registered vehicles they were substituted by, and how many unsubsti-

	M_1			M_2		
Site no.	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 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.

tuted vehicles there are left in the adjusted data files. We see that the number of substituted vehicles 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 M_1 and M_2.

6.4 Estimation

In the estimation, for M_1 and M_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 6.3) with no distinction made between 'truly registered' and 'substitute' vehicles. From Table 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 of knowing 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 θ_{kh}

for (k,h) is estimated by

$$\hat{\theta}_{kh,\text{woi}} = \frac{n_{r_{kh(0)}}}{n_{r_{kh(1+2)}}} \tag{74}$$

$$\hat{\theta}_{kh,\text{wi}} = \frac{n_{r_{kh(0)}}}{n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}}} \tag{75}$$

In both Equation (74) and (75), the denominator is intended to represent the true flow.

Estimates of multiplicative imputation error Consider the multiplicative imputation error model in Section 5.2.2. For (k,h), the multiplicative error ε_{kh} is estimated by

$$\hat{\varepsilon}_{kh,\text{woi}} = \frac{n_{I_{kh(0)}}}{n_{r_{kh(1+2)}} - n_{r_{kh(0)}}} \tag{76}$$

$$\hat{\varepsilon}_{kh,\text{wi}} = \frac{n_{I_{kh(0)}}}{n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}} - n_{r_{kh(0)}}} \tag{77}$$

In both Equation (76) and (77), the denominator is intended to represent the number of vehicles missing in the M_0 data.

Estimates of error in $\hat{\theta}^{(2)}$ Consider the additive error model for $\hat{\theta}^{(2)}$ in Equation (66). For (k,h), the error ϵ_{kh} is estimated by

$$\hat{\epsilon}_{kh,\text{woi}} = (ME)_{kh(0)} - \hat{\theta}_{kh,\text{woi}} \tag{78}$$

$$\hat{\epsilon}_{kh,\text{wi}} = (ME)_{kh(0)} - \hat{\theta}_{kh,\text{wi}} \tag{79}$$

Further consider the multiplicative model for $\hat{\theta}^{(2)}$ in Equation (67). Under this model, the error ϵ_{kh} is estimated by

$$\hat{\epsilon}_{kh,\text{woi}} = \frac{(ME)_{kh(0)}}{\hat{\theta}_{kh,\text{woi}}} \tag{80}$$

$$\hat{\epsilon}_{kh,\text{wi}} = \frac{(ME)_{kh(0)}}{\hat{\theta}_{kh,\text{wi}}} \tag{81}$$

The resulting estimates are presented in Appendix C. Throughout the appendix, one graph refers to one site, and one data point to one hour. Estimates for which the imputations are retained are indexed by plus signs; estimates for which they are removed by dots.

In Appendix C.1, the estimated registration probabilities $\hat{\theta}_{kh}$ are plotted against the 'true' flows. The estimates $\hat{\theta}_{kh,\mathrm{woi}}$ are plotted against $n_{r_{kh(1+2)}}$; the estimates $\hat{\theta}_{kh,\mathrm{wi}}$ against $n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}}$.

The estimated imputation errors are presented in Appendix C.2-C.3. In C.2, the errors are plotted against the 'true' number of missing vehicles. Then, the estimates $\hat{\varepsilon}_{kh,\mathrm{woi}}$ are plotted against $n_{r_{kh(1+2)}} - n_{r_{kh(0)}}$; the estimates $\hat{\varepsilon}_{kh,\mathrm{wi}}$ against $n_{r_{kh(1+2)}} + n_{I_{kh(1+2)}} - n_{r_{kh(0)}}$. In C.3 however, both $\hat{\varepsilon}_{kh,\mathrm{woi}}$ and $\hat{\varepsilon}_{kh,\mathrm{wi}}$ are plotted against the registered flows $n_{r_{kh(0)}}$.

The observed errors in $\hat{\theta}^{(2)}_{kh}$ under the *additive* error model are presented in Appendix C.4 and C.6. The graphs include a horizontal reference line at level zero: the desired expected value of these errors. Correspondingly, the errors under the multiplicative model are presented in Appendix C.5 and C.7 with a horizontal reference line at level one. In Appendix C.4 and C.5, the errors are plotted against the estimated registration probabilities $\hat{\theta}_{kh,\mathrm{woi}}$ and $\hat{\theta}_{kh,\mathrm{wi}}$. The diagonal patterns in the observations are a result of the measurement efficiencies (used in the calculations of the errors) only being available as integers. Note the occurrence of considerable errors for registration probabilities equal to one. Even after thorough examination of the raw data, we have not been able to come up with an explanation for this. In Appendix C.6 and C.7, the errors are plotted against the registered flows $n_{r_{kh(0)}}$.

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 formulas in Section 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 M_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

$$\tilde{y}^{(c)}_{k,\mathrm{woi}} = \frac{\hat{y}^{(c)}_{k(0)}}{n_{r_{k(1+2)}}} \tag{82}$$

$$\tilde{y}^{(c)}_{k,\mathrm{wi}} = \frac{\hat{y}^{(c)}_{k(0)}}{n_{r_{k(1+2)}} + n_{I_{kh(1+2)}}} \tag{83}$$

whereas the standardized estimate of travel time is

$$\tilde{z}^{(c)}_{k,\mathrm{woi}} = \frac{\hat{z}^{(c)}_{k(0)}}{\sum_{r_{k(1+2)}} x_\upsilon} \tag{84}$$

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* [8, Sec. 2.2.2])

$$u_k = \frac{1}{\frac{1}{y_k}\sum_{\upsilon=1}^{y_k}\frac{1}{u_\upsilon}} = \frac{y_k}{z_k} \tag{85}$$

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

$$\hat{u}^{(c)}_k = \frac{\hat{y}^{(c)}_{k(0)}}{\hat{z}^{(c)}_{k(0)}}. \tag{86}$$

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

$$\tilde{u}^{(c)}_{k,\mathrm{woi}} = \frac{\hat{u}^{(c)}_k}{n_{r_{k(1+2)}} / \sum_{r_{k(1+2)}} x_\upsilon} \tag{87}$$

$$\tilde{u}^{(c)}_{k,\mathrm{wi}} = \frac{\hat{u}^{(c)}_k}{\left(n_{r_{k(1+2)}} + n_{I_{kh(1+2)}}\right) / \sum_{r_{k(1+2)}} x_\upsilon} \tag{88}$$

In both Equation (87) and (88), the denominator is intended to represent the true average speed.

The standardized estimates of y_k and z_k are presented in Table 6 and 7, respectively, whereas the standardized estimates of u_k are given in Table 8.

6.5 Analysis

6.5.1 The forming of registration homogeneity groups

Consider the estimated registration probabilities within any site in Appendix C.1. The probability estimates are often fairly constant for adjacent hours, which speaks in favor of merging hours into larger homogeneity groups by flow. Since the relationship between registration probability and flow typically is quite smooth over the 24 hours, it seems however as unnecessary work.

6.5.2 Evaluation of the multiplicative imputation error model

If the multiplicative imputation error model is correct, the estimated errors $\hat{\varepsilon}_{kh,\mathrm{woi}}$ and $\hat{\varepsilon}_{kh,\mathrm{wi}}$ should not reveal any obvious patterns if plotted against other variables. When we plot the errors against the number of missing vehicles (Appendix C.2), for some sites, we discern however a tendency of the error variance to decrease as the number of missing vehicles increases. Due to the scarcity of observations for large numbers of missing vehicles, it is hard though to draw any certain conclusions. When the errors are plotted against the number of registered vehicles (Appendix C.3), on the other hand, no unusual structures are apparent.

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

$$\hat{\varepsilon}_{kh} = \alpha + \beta_k + e_{kh} \begin{cases} k = 1, 2, ..., b \\ h = 1, 2, ..., c \end{cases} \tag{89}$$

where $\hat{\varepsilon}_{kh}$ may be either $\hat{\varepsilon}_{kh,\mathrm{woi}}$ or $\hat{\varepsilon}_{kh,\mathrm{wi}}$, 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 α is an overall mean, β_k is the random effect of the kth site, and e_{kh} is a random error. We assume that the β_k's are $NID\left(0, \sigma_\beta^2\right)$, the e_{kh}'s $NID\left(0, \sigma_e^2\right)$, and that β_k and e_{kh} are independent. This *random effects model* (see, for instance, [13, Sec. 3-7], [14, Ch. 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 C.8.1. We see that our

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

Table 4: Confidence intervals for $\mu_{\hat{\varepsilon}}$, calculated with the imputations in the valve measurements removed and retained, respectively.

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 if 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 [14, Eq. (24.15)], a $100(1-\alpha)$ percent confidence interval on $\mu_{\hat{\varepsilon}}$ is given by

$$\bar{\bar{\hat{\varepsilon}}} \pm t_{1-\alpha/2,b-1}\sqrt{\frac{MS_{\mathrm{site}}}{bc}} \tag{90}$$

where $\bar{\bar{\hat{\varepsilon}}} = \sum_{k=1}^{b}\sum_{h=1}^{c}\hat{\varepsilon}_{kh}$ and MS_{site} is the mean square due to sites. By use of Equation (90) and the ANOVA's in Appendix C.8.1, the interval estimates of $\mu_{\hat{\varepsilon}}$ in Table 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, the hypothesis is rejected. Whether the number of imputed vehicles is conditionally unbiased for the true number of missing vehicles or not thus remains an open question.

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

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

We start by the observed errors under the *additive* model (Equations (78)-(79)). In the graphs in Appendix C.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 (Appendix C.6), no unusual structures are apparent.

Now consider the observed errors under the *multiplicative* model (Equations (80)- (81)). In the graphs in Appendix C.5, we see again the tendency of the two 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 (Appendix C.7), we see no clear signs of dependency between the variables.

In summary, so far, both models seem to fit our data quite well. In neither case have we found strong evidence against assuming constant error variance within site. Possible bias in $\hat{\theta}_{kh}^{(2)}$ as estimator of $\hat{\theta}_{kh}$ is hard to evaluate, since our results are sensitive to the choice of treatment of the imputations in the valve measurements.

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 (89) – only with $\hat{\varepsilon}_{kh}$ replaced by $\hat{\epsilon}_{kh}$ (which may represent either $\hat{\epsilon}_{kh,\mathrm{woi}}$ in Equation (78) or (80), or $\hat{\epsilon}_{kh,\mathrm{wi}}$ in Equation (79) or (81)). 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 C.8.2-C.8.3. 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 (90) with $\hat{\varepsilon}_{kh}$ replaced by $\hat{\epsilon}_{kh}$, and the ANOVA's in Appendices C.8.2 and C.8.3, the interval estimates of $\mu_{\hat{\epsilon}}$ in Table 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.

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

Table 5: Confidence intervals for $\mu_{\hat{\epsilon}}$ by error model, calculated with the imputations in the valve measurements removed and retained, respectively.

Site no.	$\tilde{y}^{(0)}_{\mathrm{woi}}$	$\tilde{y}^{(1)}_{\mathrm{woi}}$	$\tilde{y}^{(2)}_{\mathrm{woi}}$	$\tilde{y}^{(0)}_{\mathrm{wi}}$	$\tilde{y}^{(1)}_{\mathrm{wi}}$	$\tilde{y}^{(2)}_{\mathrm{wi}}$
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 6: Standardized estimates of flow, by site. (For site 3, only data from M_2 are used.)

6.5.4 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 6 and 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 travel time estimates are standardized with a too small figure (since the imputations are ignored). In all, from Tables 6 and 7, it is far from obvious which adjustment strategy (1 or 2) that ought to be recommended.

Now consider the standardized estimates of average speed in Table 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

Site no.	$\tilde{z}^{(0)}_{\text{woi}}$	$\tilde{z}^{(1)}_{\text{woi}}$	$\tilde{z}^{(2)}_{\text{woi}}$
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 7: Standardized estimates of travel time, by site. (For site 3, only data from M_2 are used.)

Site no.	$\tilde{u}^{(0)}_{k,\text{woi}}$	$\tilde{u}^{(1)}_{k,\text{woi}}$	$\tilde{u}^{(2)}_{k,\text{woi}}$	$\tilde{u}^{(0)}_{k,\text{wi}}$	$\tilde{u}^{(1)}_{k,\text{wi}}$	$\tilde{u}^{(1)}_{k,\text{wi}}$
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 8: Standardized estimates of average speed, by site. (For site 3, only data from M_2 are used.)

on "element-level" to the average speed R for all roads. The estimates in Table 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.

6.6 Summary of empirical findings

Since we could not calculate registration probabilities for individual vehicles, but only by hour, we were not really able to check the assumptions of the registration model. We seized however the opportunity to see if data spoke in favor of merging hours into larger homogeneity groups. In our opinion, this was not the case.

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 if the Strategy 1 estimators are unbiased or not.)

Under both the additive and the multiplicative error model for the estimator of the registration probability, the errors seemed independent of the 'true' probability. Both error models state that the errors are site independent. In our ANOVA tests however, throughout, 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 got 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, the travel time and their ratio average speed, and 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.

7 Summary

We have put forward two possible strategies for missing data adjustments in the speed survey. Both strategies 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 the estimation. Still, the implementation is only worth it, 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 against 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 enough 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. Also, the adjustment estimators seemed to produce better (less biased) estimates of the totals t_y and t_z than today's unadjusted estimators. We were not able to tell how their variances stand in comparison. None of the adjustment strategies showed its clear superiority to the other. Also, as already mentioned, it remains an open question if the estimator of average speed really needs any missing data adjustments.

A Proof of Theorem 5.1

By a slight generalization of [17, Result 4.5.1], the expected value of $\hat{t}_{\hat{a}^{(c)}}$ is

$$E_{p\xi}\big(\hat{t}_{\hat{a}^{(c)}}\big) = \sum_{i=1}^{N_I} E_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big)$$

and the $p\xi$-variance of $\hat{t}_{\hat{a}^{(c)}}$ is

$$V_{p\xi}\big(\hat{t}_{\hat{a}^{(c)}}\big) = \frac{1}{m_I}\sum_{i=1}^{N_I} p_i \left(\frac{E_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big)}{p_i} - E_{p\xi}\big(\hat{t}_{\hat{a}^{(c)}}\big)\right)^2 + \frac{1}{m_I}\sum_{i=1}^{N_I}\frac{V_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big)}{p_i}.$$

Hence, it suffices to show the stated expressions for $E_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big)$ and $V_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big)$.

Let subscript II indicate conditional expected value or conditional variance with respect to the design used in stage two, given os_I, and subscript III indicate conditional expected value or conditional variance with respect to the design used in stage three, given os_I and s_{IIi_v}. Then we can write

$$\begin{aligned}
E_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big) &= E_{II}E_{III}\big[E_\xi\big(\hat{t}_{\pi\hat{a}^{(c)}i}\,|s_{iq}\big)\big] = E\\
V_{p\xi}\big(\hat{t}_{\pi\hat{a}^{(c)}i}\big) &= E_{II}E_{III}\big[V_\xi\big(\hat{t}_{\pi\hat{a}^{(c)}i}\,|s_{iq}\big)\big] + E_{II}V_{III}\big[E_\xi\big(\hat{t}_{\pi\hat{a}^{(c)}i}\,|s_{iq}\big)\big]\\
&\quad + V_{II}E_{III}\big[E_\xi\big(\hat{t}_{\pi\hat{a}^{(c)}i}\,|s_{iq}\big)\big]\\
&= V_1 + V_2 + V_3.
\end{aligned}$$

The expectation is given by

$$\begin{aligned}
E &= E_{II}E_{III}\left[\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right]\\
&= E_{II}\left[\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}E_{III}\left(\frac{N_{iq}}{n_{iq}}\sum\nolimits_{s_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right)\right]\\
&= E_{II}\left[\frac{N_{IIi}}{n_{IIi}}\sum\nolimits_{s_{IIi}}\sum\nolimits_{U_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k\right]\\
&= \sum\nolimits_{U_{IIi}}\sum\nolimits_{U_{iq}}\gamma\big(\hat{a}^{(c)}\big)_k.
\end{aligned}$$

Now we turn to the variance. First,

$$\begin{aligned}
V_1 &= E_{II}E_{III}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \left(\frac{N_{iq}}{n_{iq}}\right)^2 \sum\nolimits_{s_{iq}} \delta\big(\hat{a}^{(c)}\big)_k\right] \\
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} E_{III}\left(\left(\frac{N_{iq}}{n_{iq}}\right)^2 \sum\nolimits_{s_{iq}} \delta\big(\hat{a}^{(c)}\big)_k\right)\right] \\
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} \frac{N_{iq}}{n_{iq}} \sum\nolimits_{U_{iq}} \delta\big(\hat{a}^{(c)}\big)_k\right] \\
&= \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} \frac{N_{iq}}{n_{iq}} \sum\nolimits_{U_{iq}} \delta\big(\hat{a}^{(c)}\big)_k .
\end{aligned}$$

Second,

$$\begin{aligned}
V_2 &= E_{II}V_{III}\left[\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \frac{N_{iq}}{n_{iq}} \sum\nolimits_{s_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right] \\
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} V_{III}\left(\frac{N_{iq}}{n_{iq}} \sum\nolimits_{s_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right)\right] \\
&= E_{II}\left[\left(\frac{N_{IIi}}{n_{IIi}}\right)^2 \sum\nolimits_{s_{IIi}} N_{iq}^2 \frac{1-f_{iq}}{n_{iq}} \frac{1}{N_{iq}-1} \cdot \right. \\
&\qquad \left. \cdot \sum\nolimits_{U_{iq}} \left(\gamma\big(\hat{a}^{(c)}\big)_k - \frac{1}{N_{iq}} \sum\nolimits_{U_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right)^2\right] \\
&= \frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{U_{IIi}} N_{iq}^2 \frac{1-f_{iq}}{n_{iq}} \frac{1}{N_{iq}-1} \cdot \\
&\qquad \cdot \sum\nolimits_{U_{iq}} \left(\gamma\big(\hat{a}^{(c)}\big)_k - \frac{1}{N_{iq}} \sum\nolimits_{U_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right)^2
\end{aligned}$$

and finally,

$$\begin{aligned}
V_3 &= V_{II}\left[\frac{N_{IIi}}{n_{IIi}} \sum\nolimits_{s_{IIi}} \sum\nolimits_{U_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right] \\
&= N_{IIi}^2 \frac{1-f_{IIi}}{n_{IIi}} \frac{1}{N_{IIi}-1} \cdot \\
&\qquad \cdot \sum\nolimits_{U_{IIi}} \left(\sum\nolimits_{U_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k - \frac{1}{N_{IIi}} \sum\nolimits_{U_{IIi}} \sum\nolimits_{U_{iq}} \gamma\big(\hat{a}^{(c)}\big)_k\right)^2 .
\end{aligned}$$

B A useful proposition

Derivations of γ and δ for various cases are facilitated by the following proposition (which is easily proven).

Proposition B.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*

$$E(AB) = \alpha E(A)\,,$$

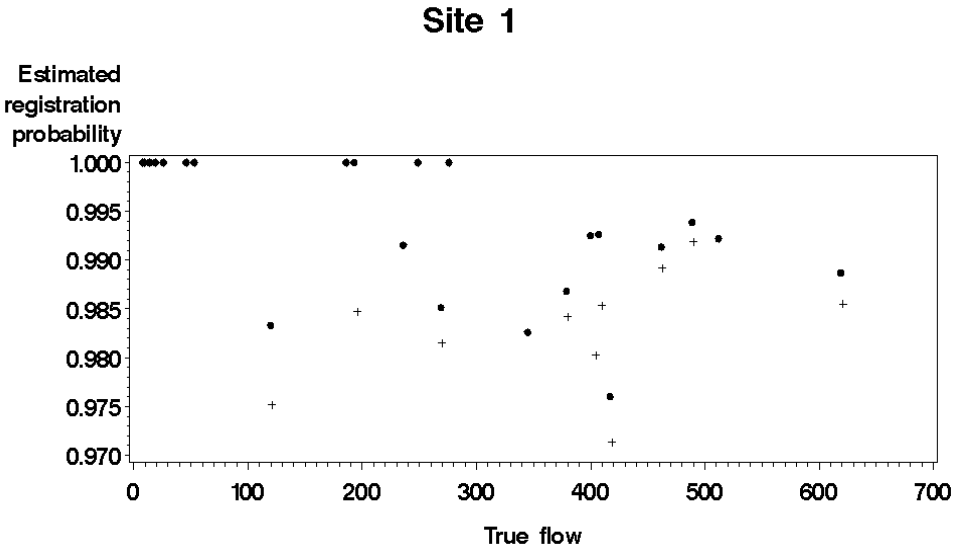
and the variance of AB by

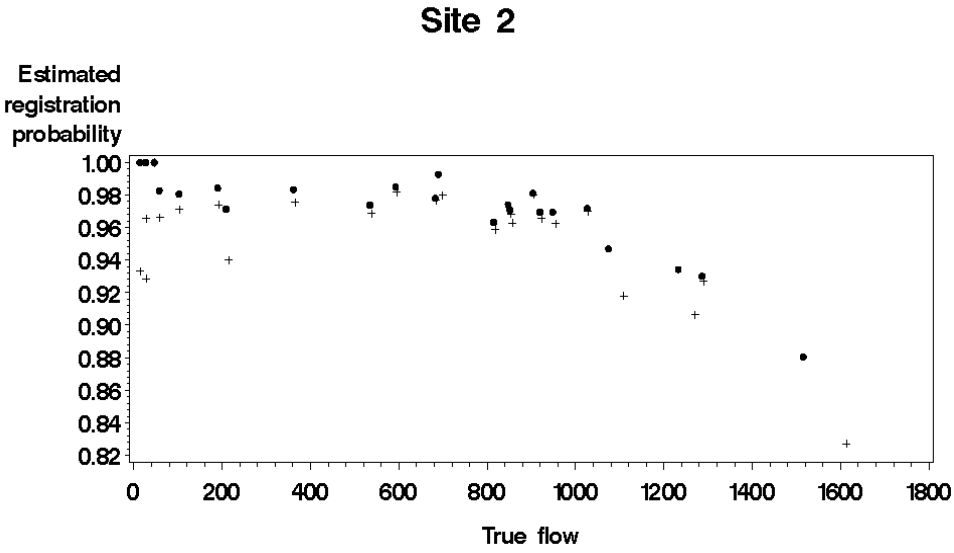
$$V(AB) = E\big[A^2 V(B\,|A)\big] + \alpha^2 V(A)\,.$$

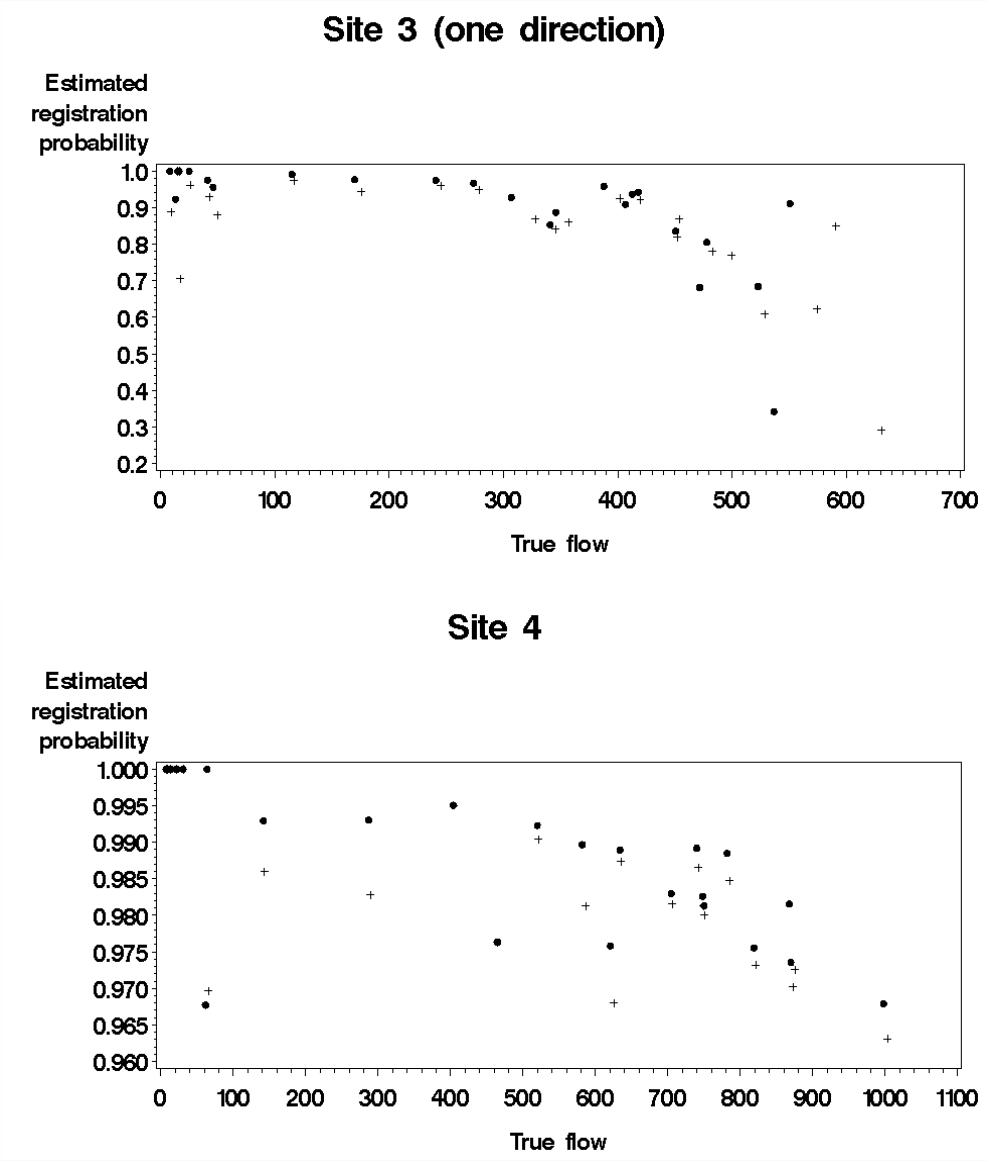
where $E(A)$ and $V(A)$ is the expectation and variance of A, respectively, and $V(B\,|A)$ is the variance of B given A.

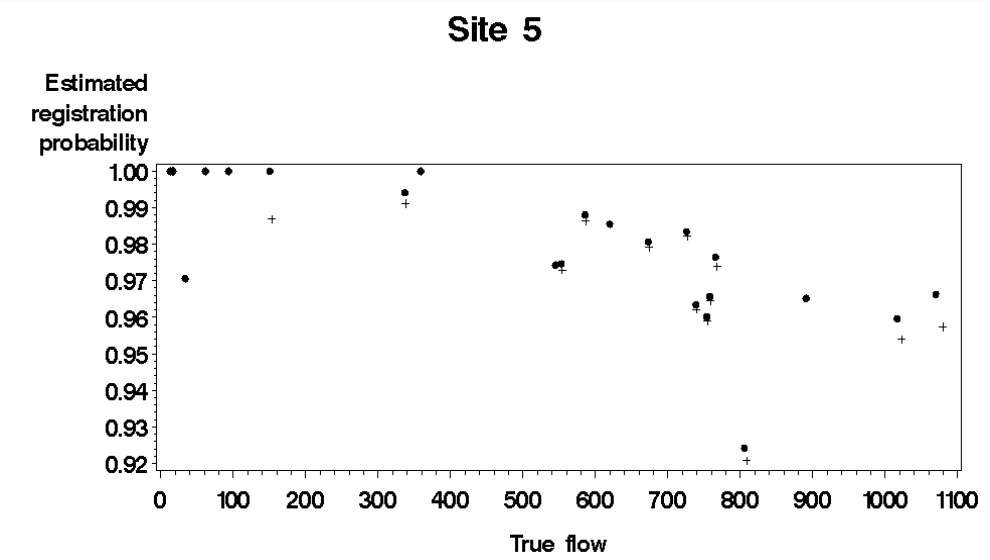
C Experimental data

C.1 Registration probability vs. Flow

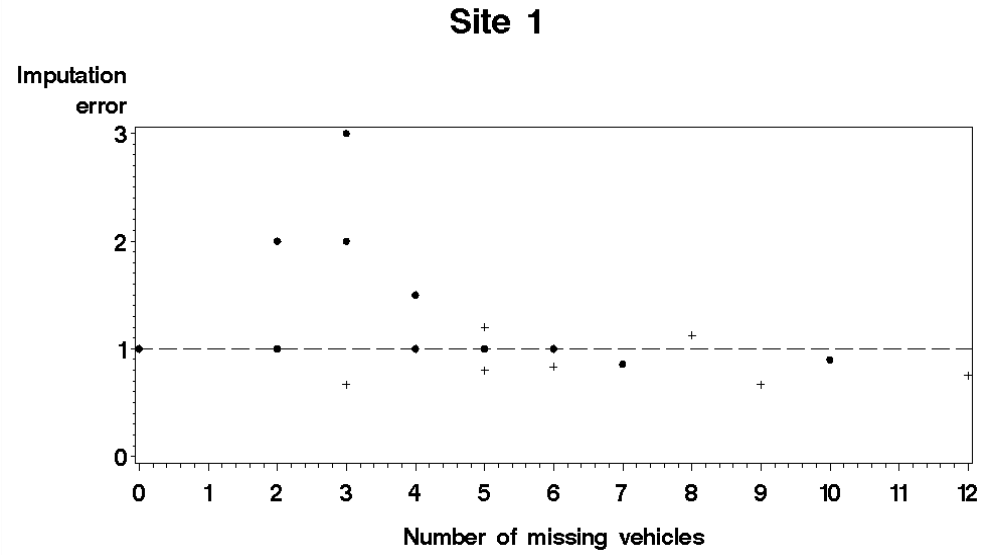


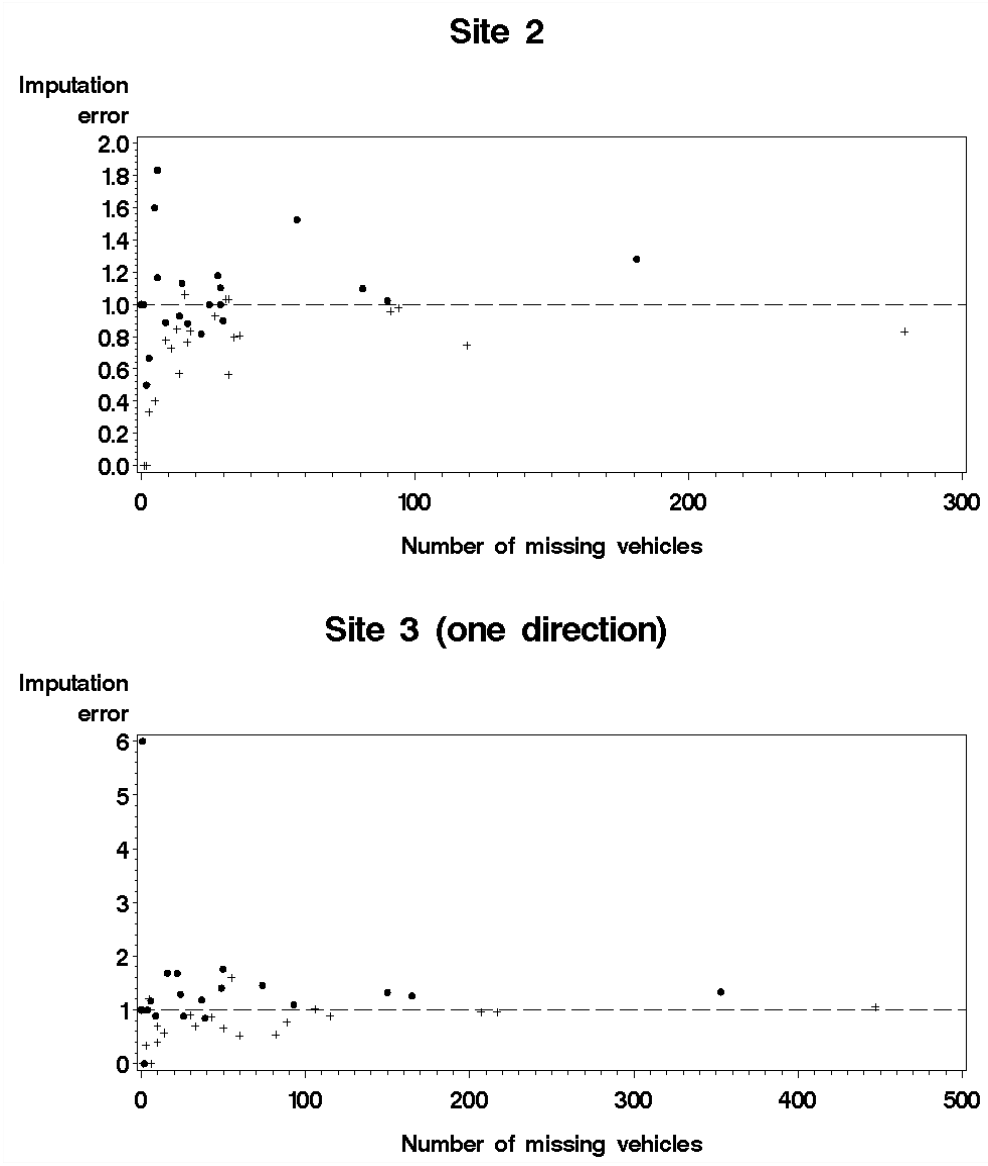


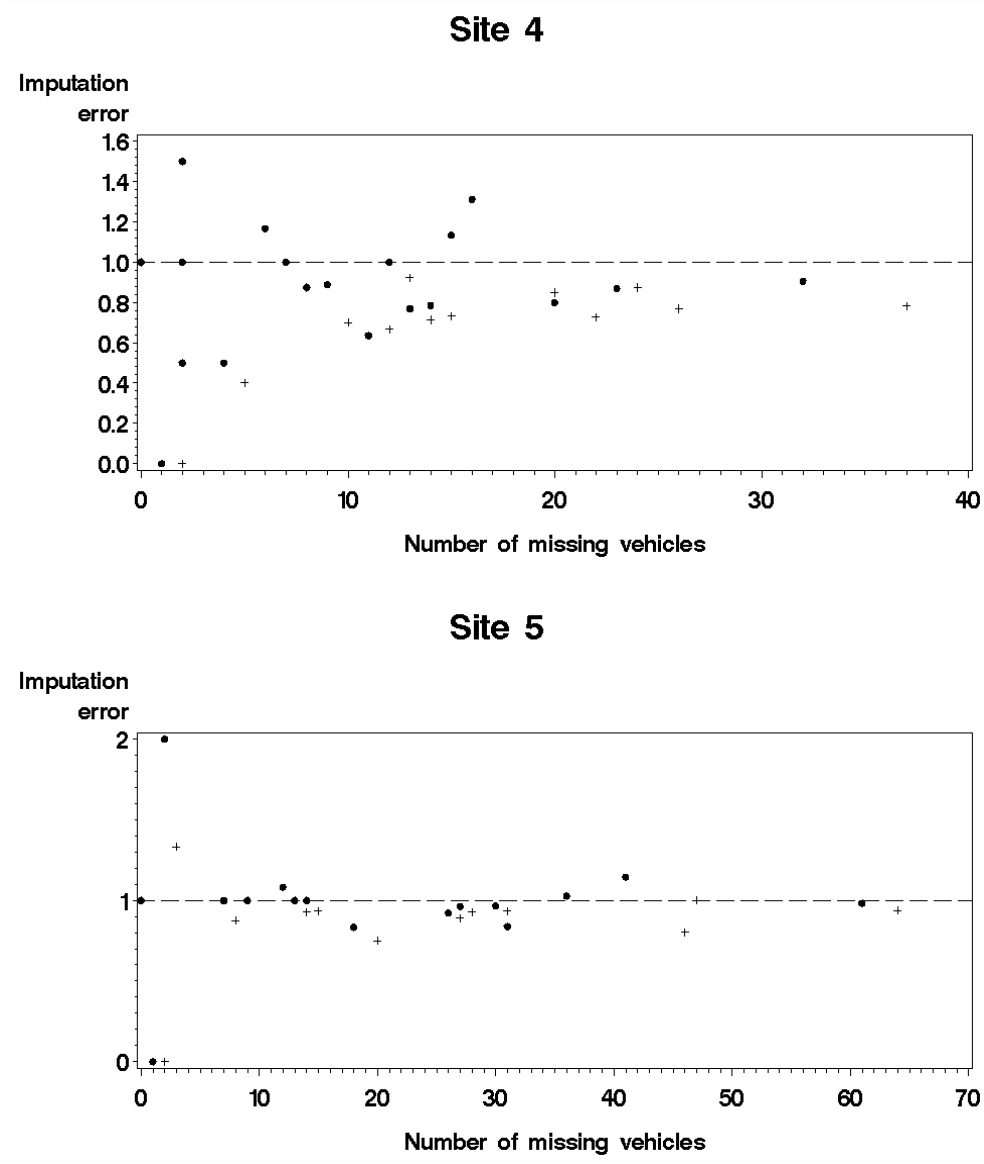

Site 3 (one direction)
Estimated registration probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
100
200
300
400
500
600
700
True flow
Site 4
Estimated registration probability
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.960
0
100
200
300
400
500
600
700
800
900
1000
1100
True flow



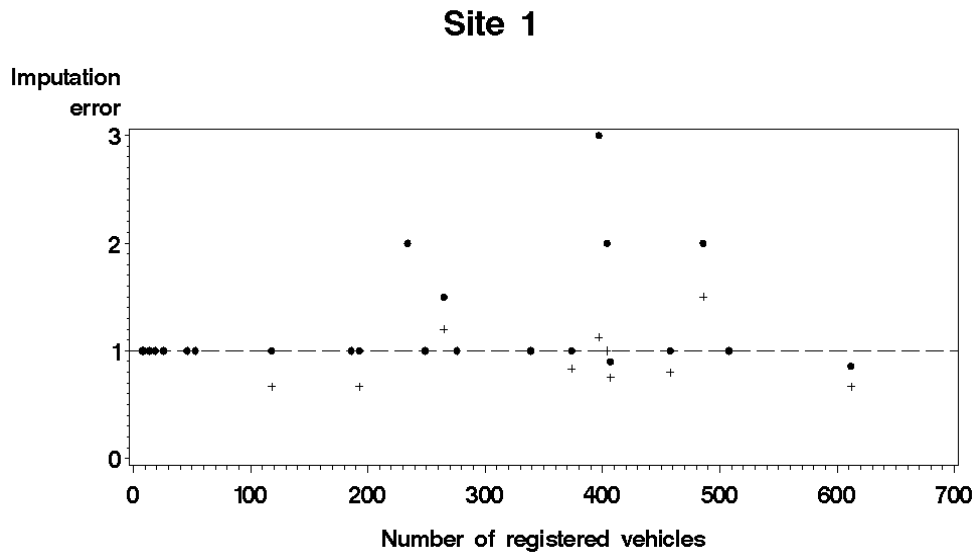
C.2 Imputation error vs. Number of missing vehicles

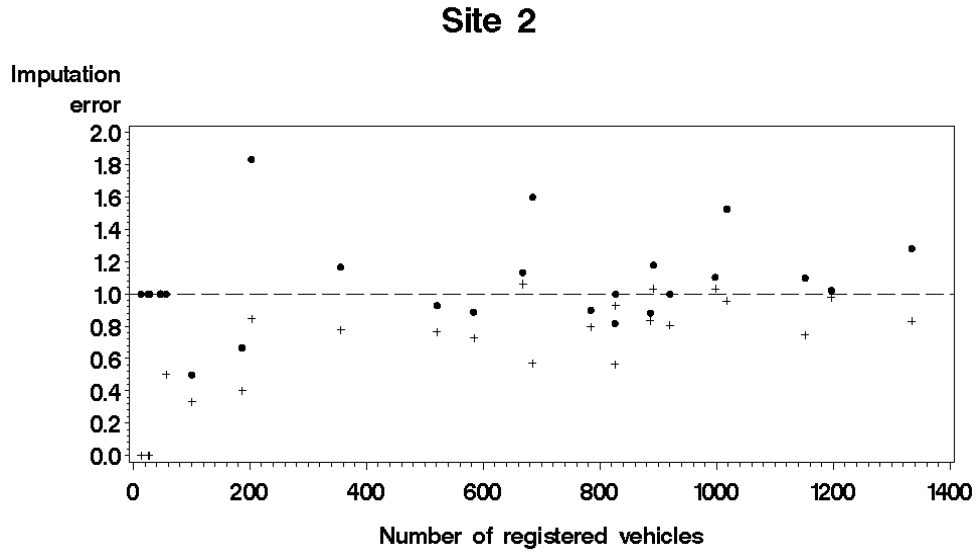


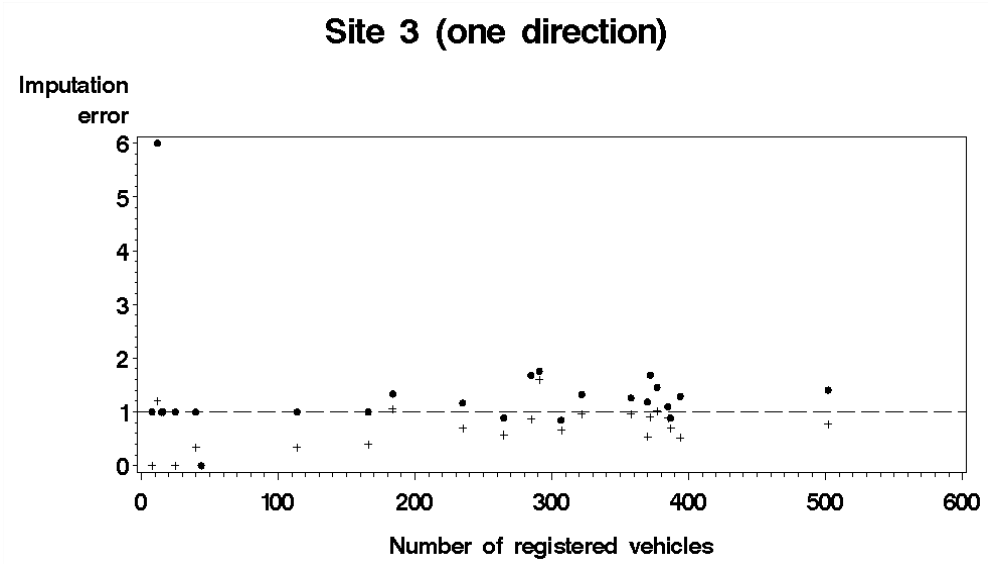

Site 2
Imputation error
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
Number of missing vehicles
Site 3 (one direction)
Imputation error
6
5
4
3
2
1
0
0
100
200
300
400
500
Number of missing vehicles

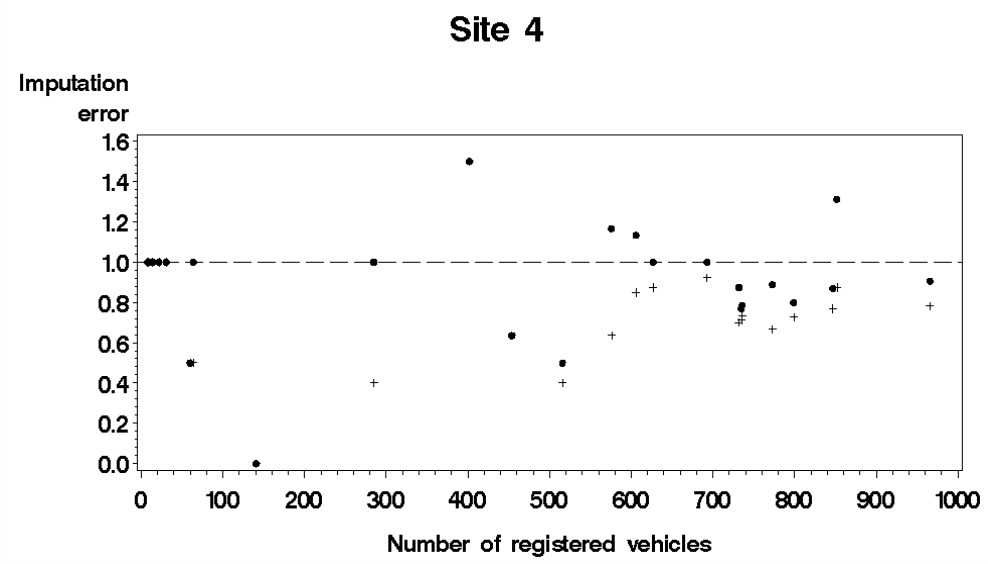

Site 4
Imputation error
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
Number of missing vehicles
Site 5
Imputation error
2
1
0
0
10
20
30
40
50
60
70
Number of missing vehicles

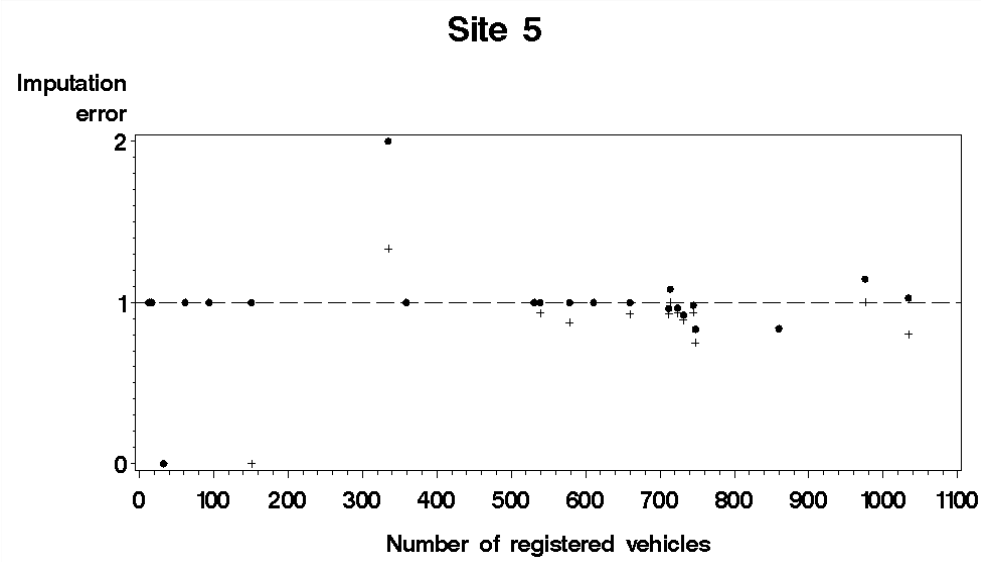
C.3 Imputation error vs. Registered flow



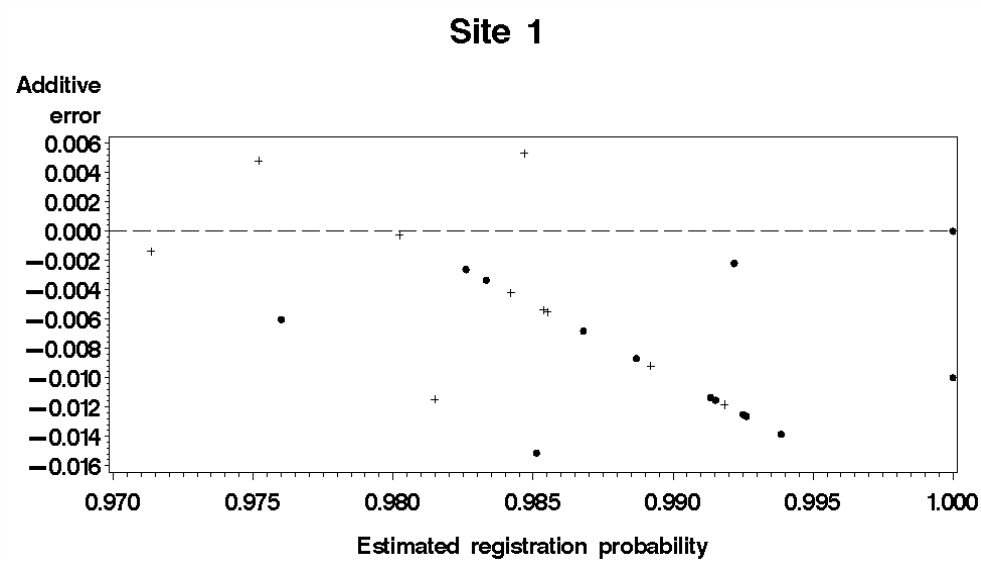


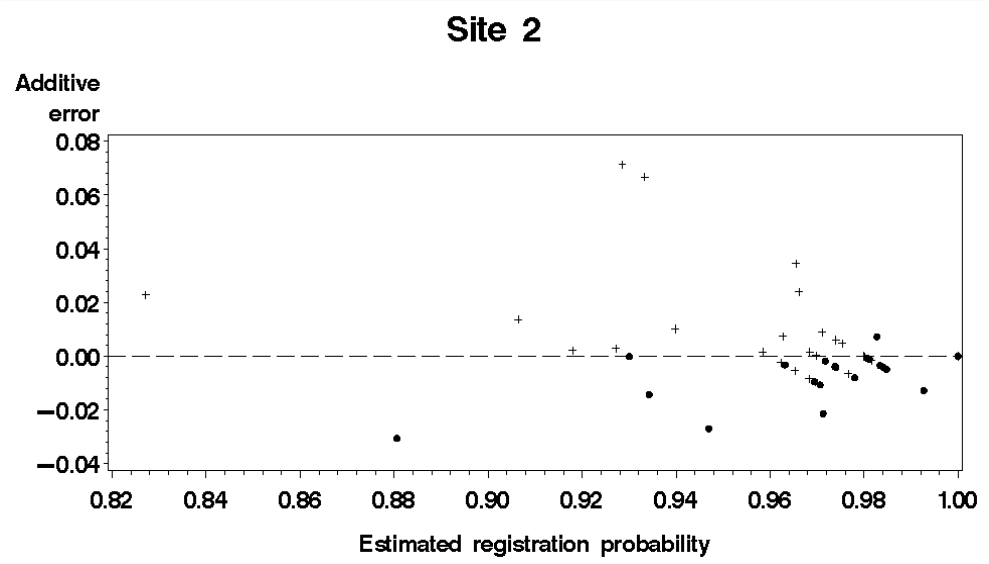

Site 3 (one direction)
Imputation error
6
5
4
3
2
1
0
0
100
200
300
400
500
600
Number of registered vehicles

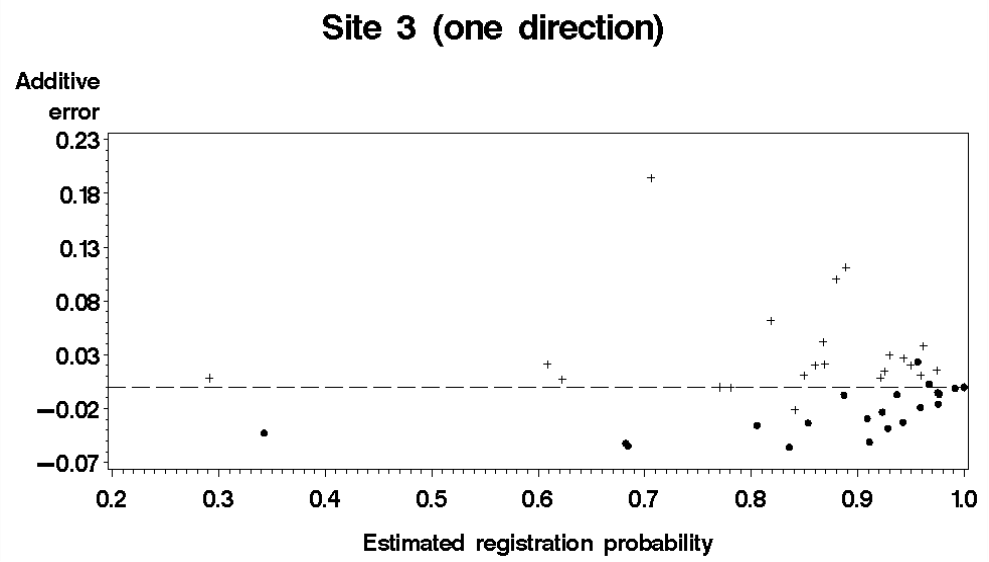

Site 4
Imputation error
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
700
800
900
1000
Number of registered vehicles

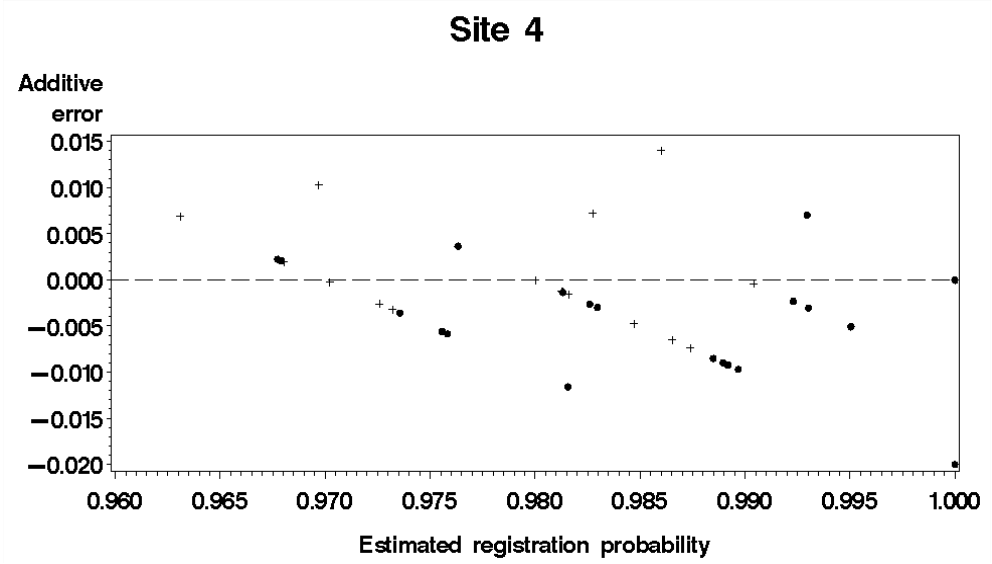


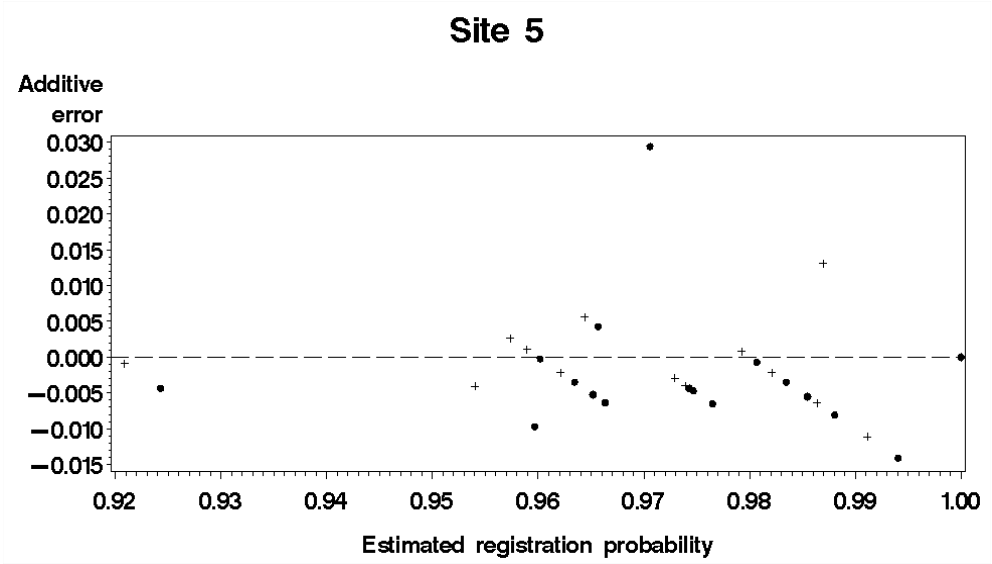
C.4 Observed errors under additive error model for $\hat{\theta}^{(2)}$



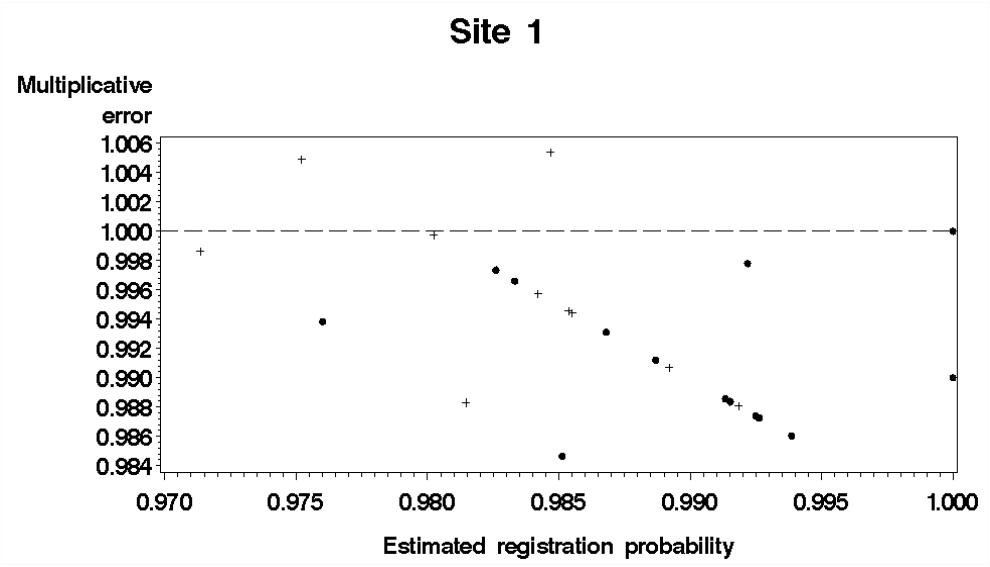

Site 2
Additive
error
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
0.82
0.84
0.86
0.88
0.90
0.92
0.94
0.96
0.98
1.00
Estimated registration probability

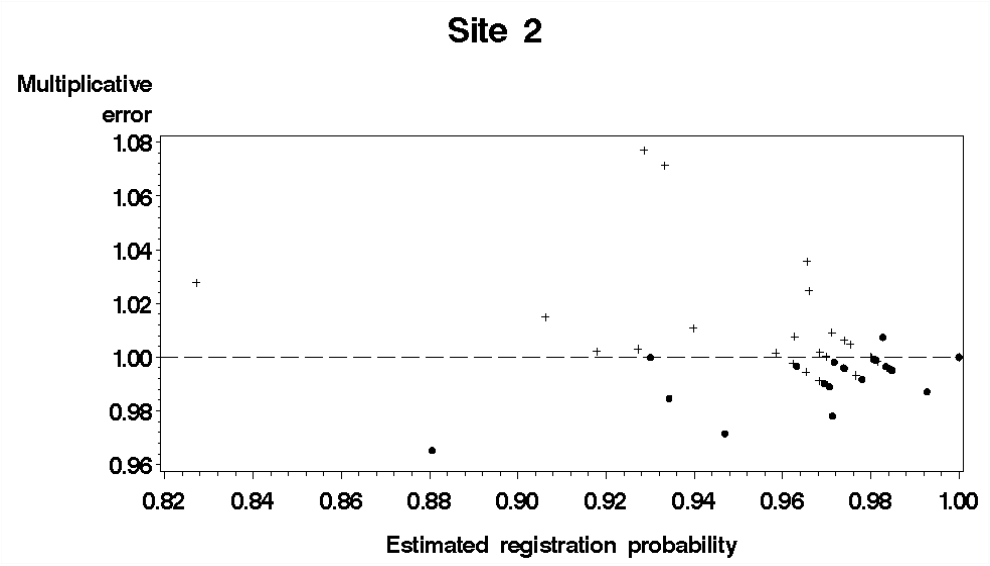

Site 3 (one direction)
Additive
error
0.23
0.18
0.13
0.08
0.03
−0.02
−0.07
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Estimated registration probability

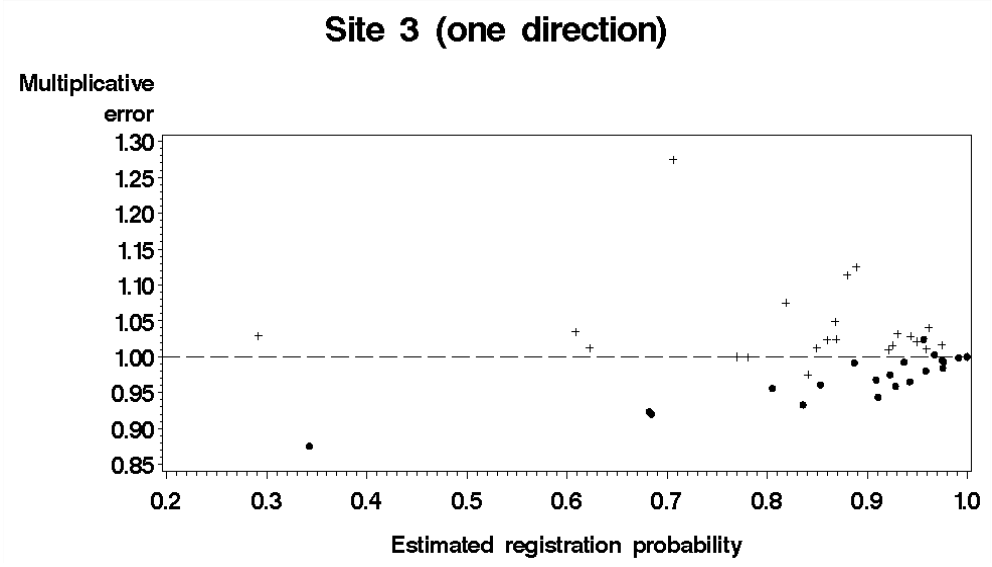

Site 4
Additive
error
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
0.960
0.965
0.970
0.975
0.980
0.985
0.990
0.995
1.000
Estimated registration probability

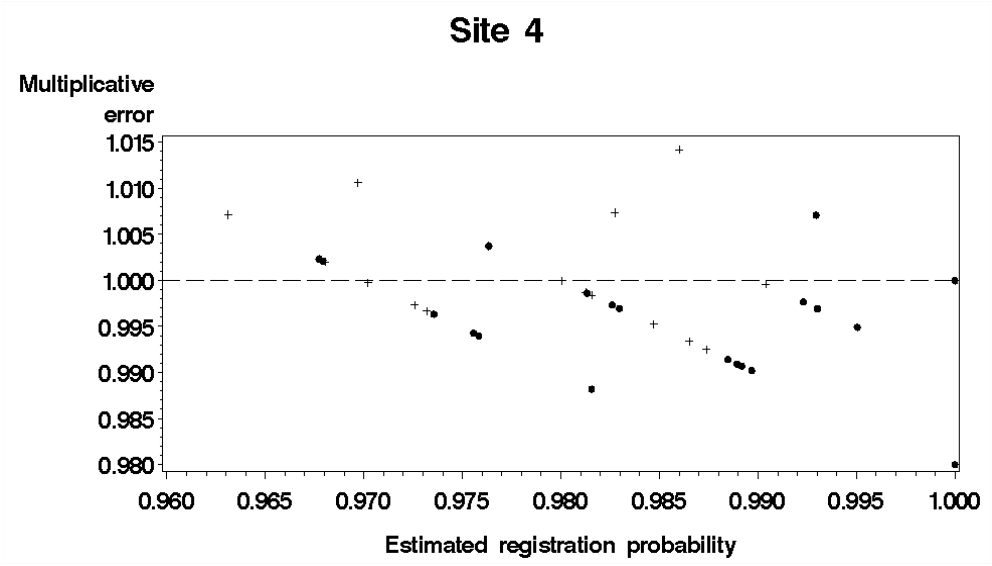

Site 5
Additive
error
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1.00
Estimated registration probability

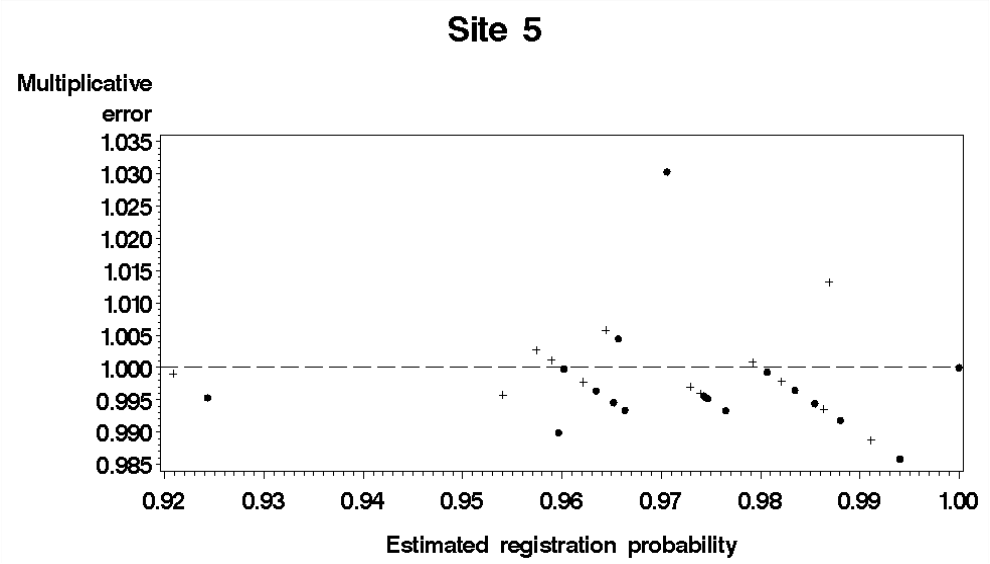
C.5 Observed errors under multiplicative error model for $\hat{\theta}^{(2)}$



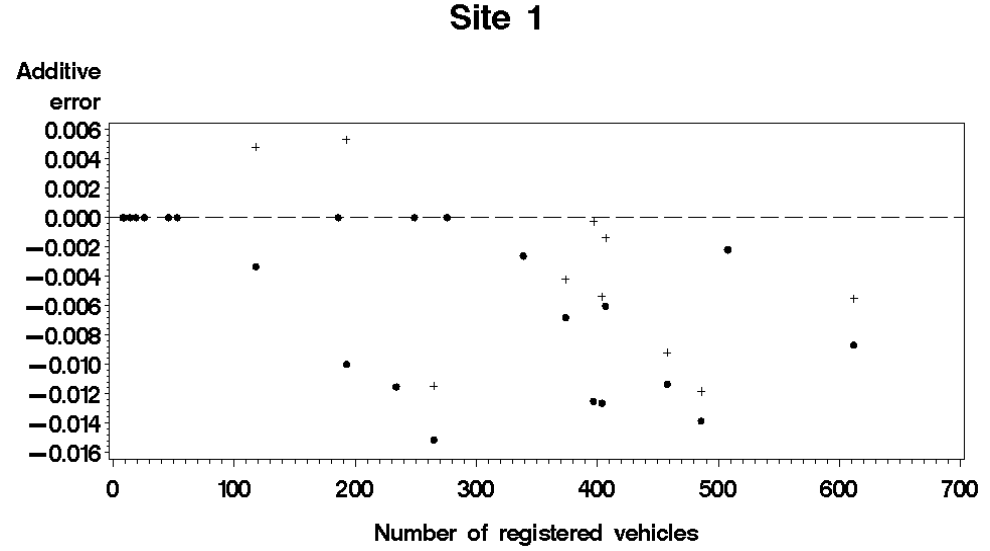


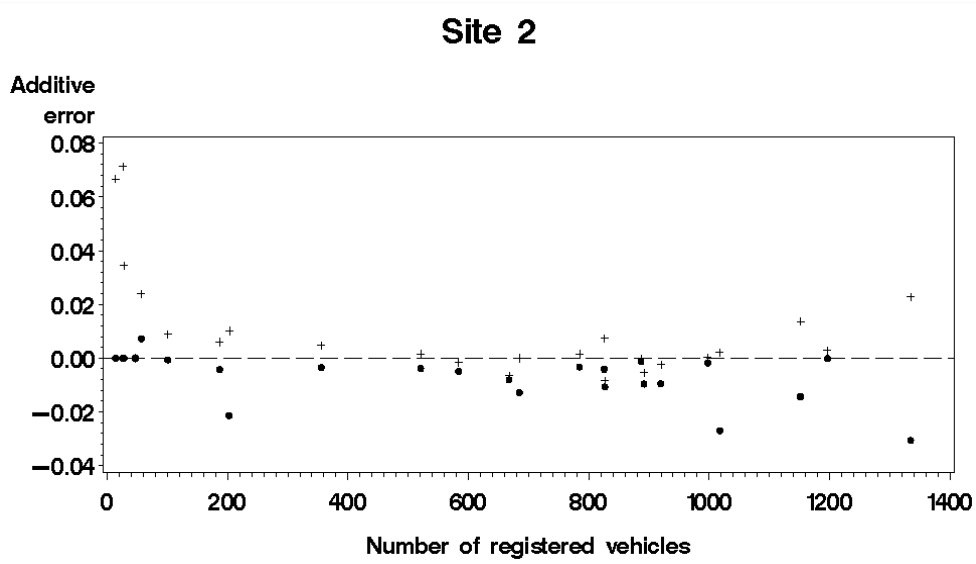

Site 3 (one direction)
Multiplicative
error
1.30
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Estimated registration probability

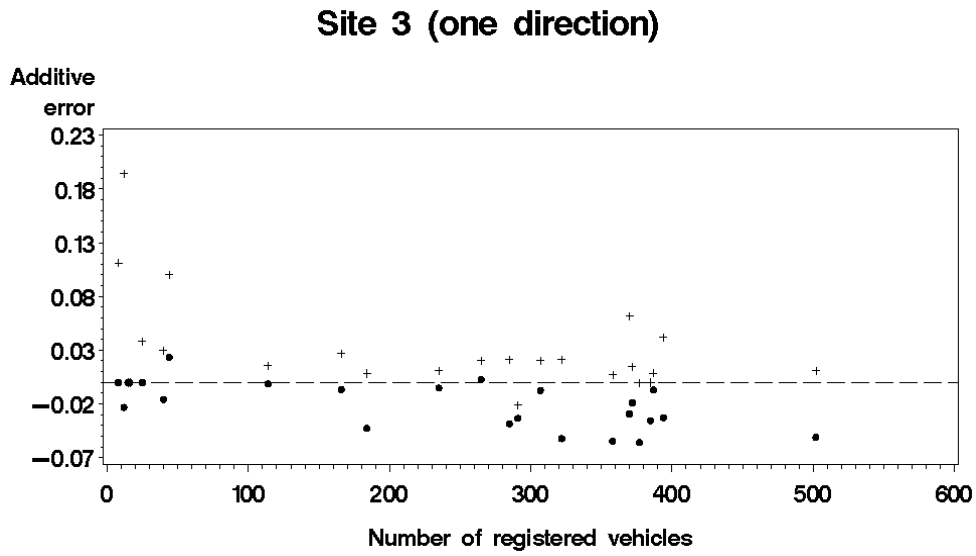

Site 4
Multiplicative
error
1.015
1.010
1.005
1.000
0.995
0.990
0.985
0.980
0.960
0.965
0.970
0.975
0.980
0.985
0.990
0.995
1.000
Estimated registration probability

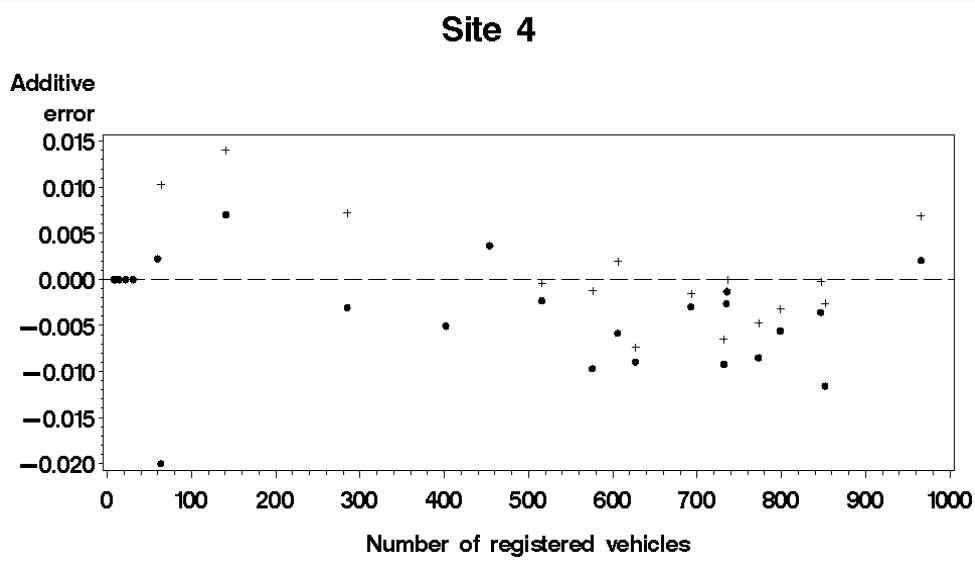


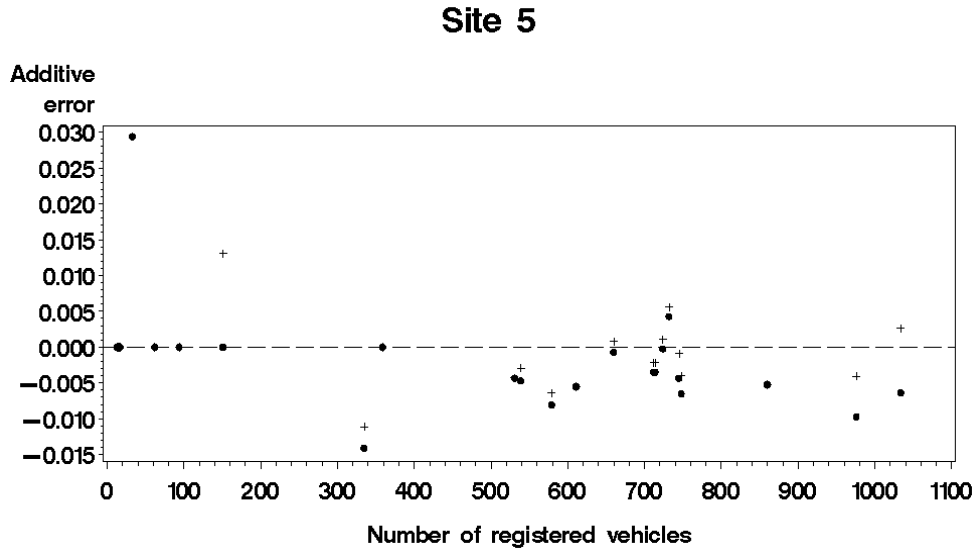
C.6 Error in $\hat{\theta}^{(2)}$ under additive error model vs. Registered flow



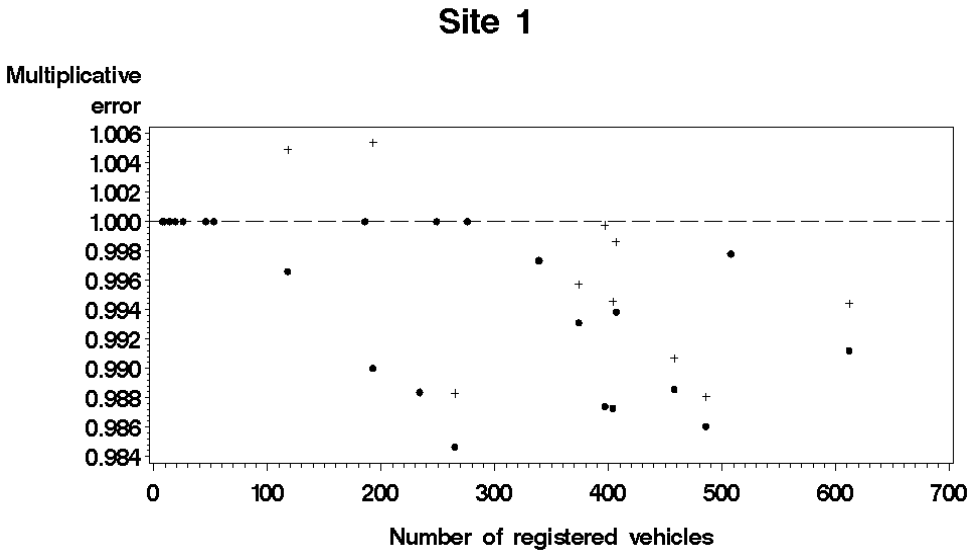

Site 2
Additive
error
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
0
200
400
600
800
1000
1200
1400
Number of registered vehicles

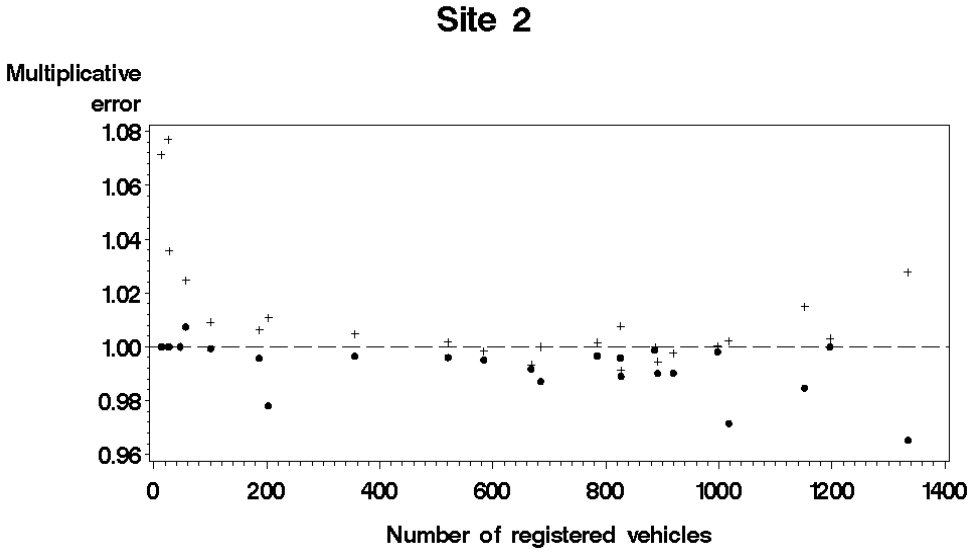

Site 3 (one direction)
Additive
error
0.23
0.18
0.13
0.08
0.03
−0.02
−0.07
0
100
200
300
400
500
600
Number of registered vehicles


Site 4
Additive
error
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
0
100
200
300
400
500
600
700
800
900
1000
Number of registered vehicles

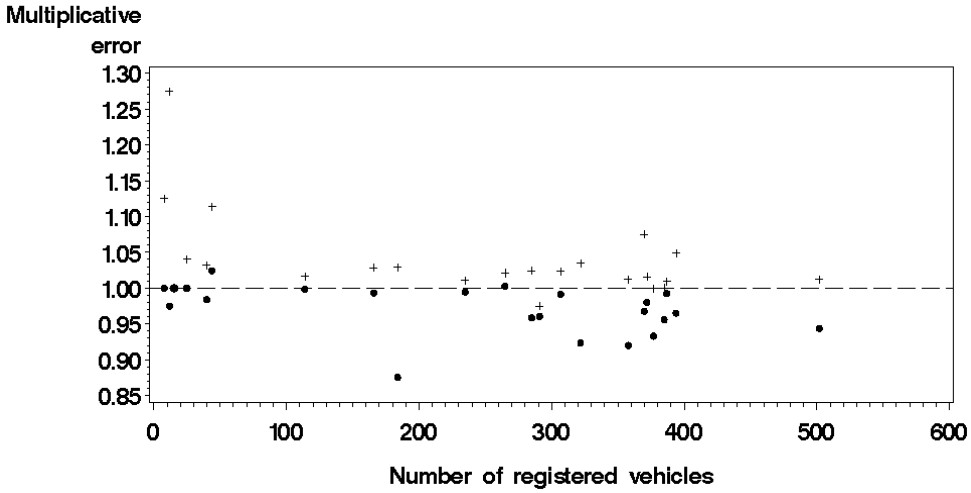

Site 5
Additive
error
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
0
100
200
300
400
500
600
700
800
900
1000
1100
Number of registered vehicles

C.7 Error in $\hat{\theta}^{(2)}$ under multiplicative error model vs. Registered flow

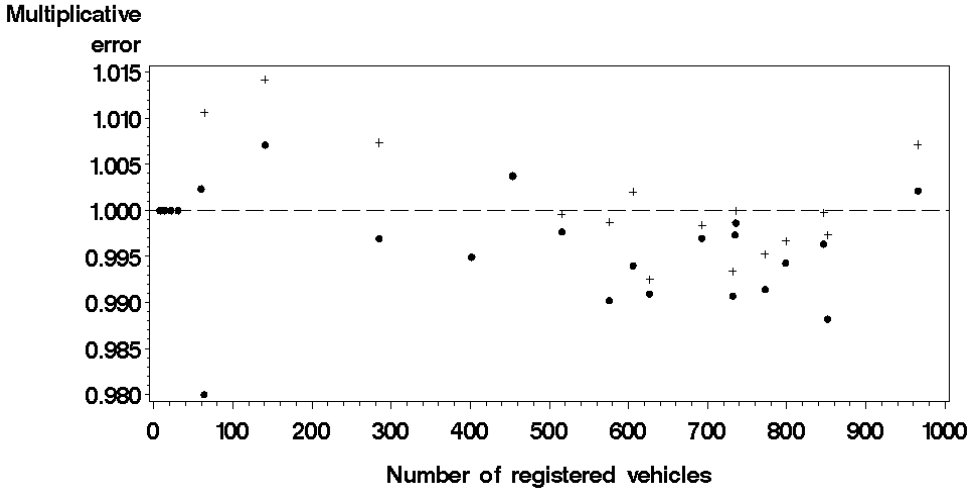


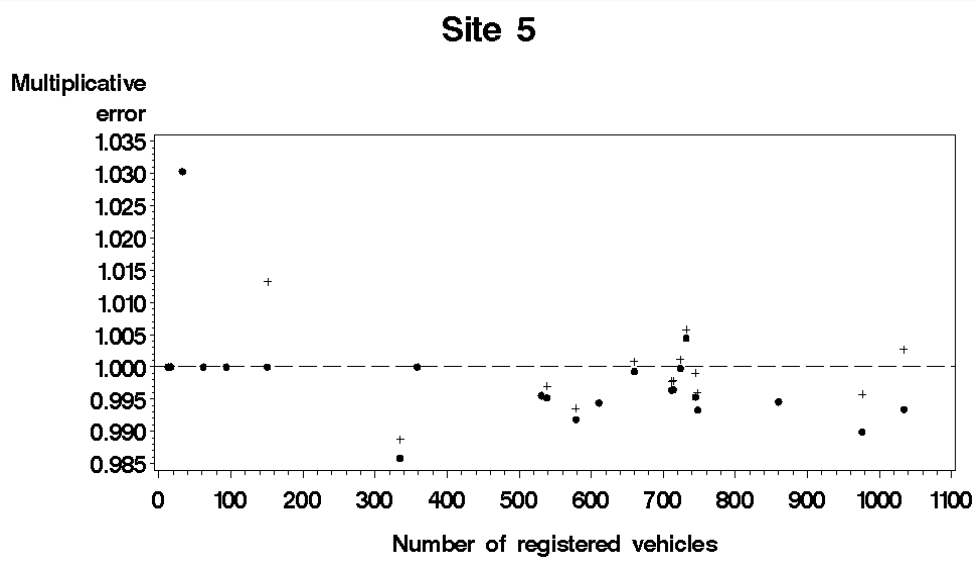


Site 3 (one direction)



Site 4





C.8 ANOVA tables

C.8.1 Under the multiplicative imputation error model

With imputations in valve measurements *removed* ($\hat{\varepsilon}_{kh} = \hat{\varepsilon}_{kh,\mathrm{woi}}$):

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			

With imputations in valve measurements *retained* ($\hat{\varepsilon}_{kh} = \hat{\varepsilon}_{kh,\mathrm{wi}}$):

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			

C.8.2 Under the additive error model for $\hat{\theta}^{(2)}$

With imputations in 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			

With imputations in valve measurements *retained* ($\hat{\epsilon}_{kh} = \hat{\epsilon}_{kh,\text{wi}}$):

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			

C.8.3 Under the multiplicative error model for $\hat{\theta}^{(2)}$

With imputations in 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	0.01050	4	0.00263	9.45	< 0.0001
Error	0.03198	115	0.00028		
Total	0.04248	119			

With imputations in valve measurements *retained* ($\hat{\epsilon}_{kh} = \hat{\epsilon}_{kh,\text{wi}}$):

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			

References

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

[3] 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.

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

[5] 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.

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

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

[8] 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.

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

[10] 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, Linköping University, SE-581 83 Linköping, Sweden, 2000.

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

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

[13] D. C. Montgomery, *Design and Analysis of Experiments*, Wiley, New York, fourth ed., 1997.

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

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

[16] 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.

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