Journal of Official Statistics, Vol. 21, No. 4, 2005, pp. 605–615

Adjustments for Missing Data in a Swedish Vehicle Speed Survey

Annica Isaksson[1]

In a Swedish vehicle speed survey for a multi-stage sample of road sites, data are collected by use of a measurement device installed on the road. Typically, some of the vehicles passing a chosen site will remain unobserved. Therefore, we suggest dividing the traffic into weighting classes. The main difficulty is to adjust the observed number of vehicles for missing data. Within class, one proposal is to add vehicles imputed by the device; another to use registration probability weighting. Models for the errors in the number of imputed vehicles, and in the estimated registration probabilities, enable theoretical evaluations of the proposals. From some empirical data, the models are evaluated.

Key words: Weighting classes; imputation; response probability; error model.

1. Introduction

Swedish traffic safety work is guided by a vision of a desirable future society, in which no one is killed or seriously injured in road traffic. To turn vision into reality, large resources are spent on changing the attitudes and behaviors of the road users. A strong emphasis is put on road-user responsibility, including compliance with speed limits. Current measures to reduce speed include physical changes of the traffic environment (for instance, converting intersections into roundabouts) and campaigns directed towards the public. In order to assess the results of these measures, the Swedish National Road Administration (SNRA) has since 1996 conducted an annual survey of vehicle speeds.

In the speed survey, for a stratified multi-stage sample of road sites, data are collected by a measurement device installed on the road. Vehicle wheel passages are here registered as pulses, and the pulses put together into vehicles. The main study variables are traffic flow, y, and 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. The main survey goal is to estimate the average speed, $R = t_y/t_z$, where t_y and t_z are the population totals of y and z, respectively.

Typically, some of the vehicles passing a chosen site remain unobserved. The failure to observe some vehicles is indicated on one hand by imputations automatically created by the device, on the other by the measurement efficiency (ME) – the proportion of registered pulses successfully combined into vehicles – being small. An undercount of vehicles is bound to bias the estimators of the totals, whereas the impact on the estimator of R is unclear.

At present, the missing data problem is simply ignored – an approach henceforth referred to as Strategy 0. In this article, which to a great extent is based on Isaksson (2003,

[1]Statistics Sweden, Department of Research and Development, SE-701 89 Örebro, Sweden. Email: annica.isaksson@scb.se

Acknowledgment: The financial support from the Swedish National Road Administration and from the Bank of Sweden Tercentenary Foundation (Grant no. 2000-5063) is gratefully acknowledged.

Chapter 4), two strategies for adjusting for missing data in the estimation stage of the survey are introduced. Both are designed for easy implementation: they do not require simulations or the collection of new auxiliary data, but only minor modifications of the computer programs presently used for estimation. One strategy utilizes the imputations for adjustments; the other utilizes the ME for the same purpose. The two strategies rest however on a common model for the vehicle registration mechanism.

2. Main Survey Operations

Here, brief descriptions of some important operations of the speed survey – the sample selection, the collection of data and the estimation – are provided.

2.1. Sample selection

Road sites are selected for observation by a three-stage design with stratification in each stage. In particular, in the final stage, road sites are stratified according to speed limit and priority (major road or not). In the final report of the survey, domain estimates are presented for various strata. To simplify, we choose however to ignore all stratification in this article. We also ignore the fact that in stage one, the three largest units define a take-all stratum.

The primary sampling units (PSUs) are the N_I population centers in Sweden. The ith PSU is represented by its label i. Thus, we denote the set of PSUs as $U_I = \{1, \ldots, i, \ldots, N_I\}$. Population center $i \in U_I$ is partitioned into N_{IIi} small areas, labeled $q = 1, \ldots, N_{IIi}$, that represent the secondary sampling units (SSUs). The set of SSUs formed by the subdivision of i is denoted $U_{IIi} = \{1, \ldots, q, \ldots, N_{IIi}\}$. Finally, the roads in small area q in population center i are viewed as partitioned into N_{iq} one-meter road sites.

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

Stage I. A probability-proportional-to-size sample of PSUs is drawn with replacement and with probability proportional to the number of inhabitants. Let i_ν denote the PSU selected in the νth draw, $\nu = 1, \ldots, m_I$, and p_{i_ν} the probability of selecting i_ν. The vector of selected PSUs, $(i_1, \ldots, i_\nu, \ldots, i_{m_I})$, constitutes the 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 $q \in s_{IIi_\nu}$.

2.2. Data collection

The measurement device 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 pulse stream, the analyzer creates vehicles and assigns speeds to them.

Missing data arise when arrived pulses cannot unambiguously be translated into vehicles. On the basis of excess pulses, by a stepwise, basically non-random, procedure, vehicles are imputed. The procedure involves addition and subtraction of pulses, and may

generate a number of imputed vehicles that goes below or above the true number of unregistered vehicles. The imputed vehicles are also assigned speeds, based on those of previously registered vehicles. At present, the imputations are however discarded in the estimation. The main reason for this is that their speeds are judged as undependable.

We use the following notation. The set of vehicles passing road site k (during a given time period) consists of y_k vehicles labeled $v = 1, \ldots, y_k$. For simplicity, we let the vth vehicle be represented by its label v. Hence, for the site and time period in question, the population of passing vehicles is denoted as $U_k = \{1, \ldots, v, \ldots, y_k\}$. 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[2]. The successfully observed subset of U_k is denoted r_k of size n_{r_k}. Under Strategy 0, the estimators of y_k and z_k are $\hat{y}_k^{(0)} = n_{r_k}$ and $\hat{z}_k^{(0)} = \sum_{r_k} x_v$, respectively.

2.3. Estimation with complete data

Let $t_a = \Sigma_U a_k$, where a_k is the true value of study variable a (which can be y or z) for site $k \in U$. Ideally, a_k is known for all $k \in s$. Then, from Särndal et al. (1992, Result 4.5.1.), a design-unbiased estimator of t_a is given by

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

where $\hat{t}_{\pi a i_\nu} = (N_{IIi_\nu}/n_{IIi_\nu}) \sum_{s_{IIi_\nu}} \hat{t}_{\pi a i_\nu q}$ 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 $N_{IIi_\nu} = N_{IIi}$ and $N_{i_\nu q} = N_{iq}$.) From Raj (1968, Section 6.8.2.), an approximately design-unbiased estimator of R is given by

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

If some data are missing, the true a_k's are unknown, and Equation (1)–(2) are no longer applicable. This situation is treated in Section 4.

3. Proposals for Missing Data Adjustments

In this section, we start by formulating a model for the distribution that generates the set of registered vehicles r_k for an observed road site k. The model is then used as a starting-point for the construction of estimators, which take unregistered vehicles into account.

3.1. A registration model

Our adjustment strategies both rest on the following model, which, in all essentials, coincides with the response homogeneity group model in Särndal and Swensson (1987, Equation (8.1)) or Särndal et al. (1992, Equation (15.6.6)).

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

3.1.1. Registration model

Assume that the vehicles passing road site $k \in s_{i_v q}$ during a selected day are partitioned into H_k groups $U_{kh}\,(h = 1, \ldots, H_k)$ such that, given $s_{i_v q}$

- all y_{kh} 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 only practical way of grouping a traffic flow is probably by time intervals. In our experiment (see Section 5), watch-hour is used as basis of division.

3.2. Strategy 1

Our first proposal is to adjust for missing data by use of the imputed vehicles. We do not trust their speeds, but let our Strategy 1 estimators make use only of their number. As estimator of y_k, we propose

$$\hat{y}_k^{(1)} = \sum_{h=1}^{H_k} (n_{r_{kh}} + n_{I_{kh}}) = \sum_{h=1}^{H_k} \hat{y}_{kh}^{(1)} \tag{3}$$

where $n_{r_{kh}}$ and $n_{I_{kh}}$ are the numbers of registered and imputed vehicles, respectively, in homogeneity group U_{kh}. As estimator of z_k, we propose

$$\hat{z}_k^{(1)} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{\hat{\theta}_{kh}^{(1)}} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{n_{r_{kh}}/\hat{y}_{kh}^{(1)}} = \sum_{h=1}^{H_k} (n_{r_{kh}} + n_{I_{kh}})\bar{x}_{r_{kh}} \tag{4}$$

where $\bar{x}_{r_{kh}} = \sum_{r_{kh}} x_v / n_{r_{kh}}$

The estimator $\hat{y}_k^{(1)}$ is a function of the $n_{r_{kh}}$'s, whose stochastic properties are regulated by the registration model, and of the $n_{I_{kh}}$'s, which in principle are fixed entities (since the imputation procedure is deterministic). To simplify, we will however treat the latter also as random variables.

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 $\hat{z}_k^{(1)}$ of z_k would be the census version (the special case when the ambition is to observe all members of the population, and thus missing data are the sole source of randomness) of the direct weighting estimator given in Särndal and Swensson (1987, Equation 4.10) or Särndal et al. (1992, Equation 15.6.8). Conditional on $s_{i_v a}$, and provided that the probability of an empty homogeneity group is negligible, $\hat{z}_k^{(1)}$ would then be unbiased for z_k under the registration model.

3.3. Strategy 2

If we do not use the imputed vehicles, we have few options left for adjusting the observed flow for missing data. One remaining possibility, however, is to weight the number of registered vehicles in a suitable manner. The estimate $\hat{\theta}_{kh}^{(1)}$ is no longer an option, but another estimate of θ_{kh} is needed.

The possibility of estimating (response) probabilities from auxiliary data is only quite sparsely discussed in the literature. Some early references include Chapman (1976, Section 3.5) and Drew and Fuller (1980, 1981). Ekholm and Laaksonen (1991) model

response probabilities by logistic regression and estimate them from the fitted model. Nonparametric estimation methods are discussed for instance by Giommi (1987).

If missing data adjusted estimators of y_k and z_k include model parameters, these parameters will need to be estimated from sample data somehow. We try an easy evasion of this problem by searching for an auxiliary variable with roughly a one-to-one relationship with θ_{kh}. The ME is the variable we hope fits the description best. Thus, our second proposal for the estimator of y_k is given by:

$$\hat{y}_k^{(2)} = \sum_{h=1}^{H_k} \frac{n_{r_{kh}}}{\hat{\theta}_{kh}^{(2)}} = \sum_{h=1}^{H_k} \frac{n_{r_{kh}}}{(ME)_{kh}} = \sum_{h=1}^{H_k} \hat{y}_{kh}^{(2)} \tag{5}$$

where $(ME)_{kh}$ is the ME for homogeneity group U_{kh}. Our corresponding proposal for the estimator of z_k is given by:

$$\hat{z}_k^{(2)} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{\hat{\theta}_{kh}^{(2)}} = \sum_{h=1}^{H_k} \frac{\sum_{r_{kh}} x_v}{(ME)_{kh}} \tag{6}$$

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

3.4. Error models for $n_{I_{kh}}$ and $\hat{\theta}_{kh}^{(2)}$

Let $\hat{a}_k^{(c)}, k \in s_{i_\nu q}$, be the estimator of a_k under Strategy c ($c = 0, 1, 2$). As a means for evaluating the statistical properties of estimators based on $\hat{a}_k^{(1)}$, we formulate the following error model for $n_{I_{kh}}$ as estimator of $y_{kh} - n_{r_{kh}}$.

3.4.1. Imputation error model

Let the vector of all $n_{r_{kh}}$'s in site $k \in s_{i_\nu q}$ be denoted $\mathbf{n}_{r_k}=(n_{rk1}, \ldots, n_{rkh}, \ldots, n_{r_{kH_k}})$. Given $s_{i_\nu q}$ and $\mathbf{n}_{r_k}$, for homogeneity group $U_{kh}(h = 1, \ldots, H_k, k \in s_{i_\nu q})$,

- the number $n_{I_{kh}}$ of imputed vehicles consists of the number of unregistered vehicles times a random error $\varepsilon_{kh} : n_{I_{kh}} = (y_{kh} - n_{r_{kh}})\varepsilon_{kh}$,
- the expected value and variance of ε_{kh} are independent of $s_{i_\nu q}, n_{r_{kh}}$ and the road site k, and
- the $n_{I_{kh}}$'s are independent.

A multiplicative imputation error model makes sense since the more vehicles that are not registered, the more complicated the imputation task is, and the higher the risk of large errors arising.

In order to be able to evaluate the statistical properties of estimators based on $\hat{a}_k^{(2)}$, we formulate the following error model for $\hat{\theta}_{kh}^{(2)}$ as estimator of θ_{kh}:

3.4.2. Error model for $\hat{\theta}_{kh}^{(2)}$

For homogeneity group U_{kh} $(h = 1, \ldots, H_k, k \in s_{iq})$,

- the estimator $\hat{\theta}_{kh}^{(2)} = (ME)_{kh}$ is a function of θ_{kh} and a random error ϵ_{kh},
- the expected value and variance of ε_{kh} are independent of s_{iq}, n_{r_k} and the road site k, and
- the $\hat{\theta}_{kh}^{(2)}$'s are independent.

We restrict our attention to two simple functional relationships: the additive error model, $\hat{\theta}_{kh}^{(2)} = \theta_{kh} + \epsilon_{kh}$, and the multiplicative error model, $\hat{\theta}_{kh}^{(2)} = \theta_{kh}\epsilon_{kh}$.

4. Estimation with Missing Data

Let the estimator of t_a obtained by replacing a by $\hat{a}^{(c)}$ in Equation (1) be denoted $\hat{t}_{\hat{a}^{(c)}}$. We are interested in the statistical properties of the missing data estimator

$$\hat{R}^{(c)} = \frac{\hat{t}_{\hat{y}^{(c)}}}{\hat{t}_{\hat{z}^{(c)}}} \tag{7}$$

of R. The joint probability distribution (conditional on $s_{i_\nu q}$) of $\hat{a}_k^{(c)}$ is here called Model ξ. (Under Strategy 0, Model ξ corresponds to the registration model, under Strategy 1 to the joint registration and imputation error model, and under Strategy 2 to the joint registration model and error model for $\hat{\theta}_{kh}^{(2)}$.)

Let E_p and V_p denote expectation and variance with respect to the sampling design p described in Section 2.1. For nonlinear estimators, such as the ratio of two estimated population totals, it is the practice to use the variance of a linearized statistic as an approximation to the exact variance. Let AV_p denote such an approximative variance, again with respect to p. (For details on the linearization technique, see Särndal et al. (1992, Section 5.5).) Correspondingly, expectations and variances are indicated by subscript ξ if taken with respect to model ξ, and by $p\xi$ if taken with respect jointly to the sampling design p and model ξ.

Jointly under the sampling design p and model ξ, the estimator $\hat{R}^{(c)}$ has the approximate expected value

$$E_{p\xi}(\hat{R}^{(c)}) \approx \frac{E_{p\xi}(\hat{t}_{\hat{y}^{(c)}})}{E_{p\xi}(\hat{t}_{\hat{z}^{(c)}})} = \frac{\sum_U E_\xi(\hat{y}_k^{(c)}|s_{iq})}{\sum_U E_\xi(\hat{z}_k^{(c)}|s_{iq})} \tag{8}$$

In general, 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 $\hat{R}$) is unknown. Consider however the favorable case when $\hat{y}_k^{(c)}$ and $\hat{z}_k^{(c)}$ are unbiased for y_k and z_k, respectively. Then, $\hat{R}^{(c)}$ is approximately unbiased for R. Also, as shown in Isaksson (2003, Appendix B), the variance increase due to using $\hat{R}^{(c)}$ instead of $\hat{R}$ as estimator of R is given by

$$AV_{p\xi}(\hat{R}^{(c)}) - AV_p(\hat{R}) = \frac{1}{t_z^2}\frac{1}{m_I}\sum_{i=1}^{N_I}\frac{1}{p_i}\frac{N_{IIi}}{n_{IIi}}\sum_{U_{IIi}}\frac{N_{iq}}{n_{iq}}\sum_{U_{iq}} V_\xi(\hat{y}_k^{(c)} - R^{(c)}\hat{z}_k^{(c)}|s_{iq}) \tag{9}$$

Assume that the registration model holds. Then, under Strategy 1, $\hat{y}_k^{(c)}$ and $\hat{z}_k^{(c)}$ are unbiased for y_k and z_k, respectively, if ε_{kh} has expectation one. The estimators of y_k and z_k under Strategy 2 are theoretically complicated, being ratios of random variables. If $\hat{\theta}_{kh}^{(2)}$ is unbiased for θ_{kh}, by first-order Taylor approximation, the estimators $\hat{y}_k^{(2)}$ and $\hat{z}_k^{(2)}$ are however approximately unbiased for y_k and z_k, respectively.

Under all strategies, the explicit expression for $V_\xi(\hat{y}_k^{(c)} - R\hat{z}_k^{(c)}|s_{iq})$ is quite complicated. For details, see, Isaksson (2003, Section 4.4).

5. Empirical Study

In order to evaluate the error models presented in Section 3.4, and investigate the empirical behaviors of $\hat{a}_k^{(c)}$, in summer 2001, we conducted an experiment. Data were collected for five road sites in the city of Linköping, Sweden. The sites were chosen to represent different types of traffic environments. However, to simplify, the study was limited to two-way, two-lane streets with a speed limit of 50 kilometers per hour.

5.1. Data collection

At each chosen site, data were collected during 24 successive hours by use of two pairs of pneumatic tubes and three traffic analyzers. One pair of tubes, connected to traffic analyzer M_0, was used for simultaneous observation of vehicles on both street lanes. A second pair of tubes was installed in parallel with the first, only with a slight lateral displacement. By use of valves, these tubes were plugged at the center line marking of the street: a procedure which enables separate measurement of each lane. The tube ends on each side of the valves were connected to the traffic analyzers M_1 and M_2, respectively.

The plugging method has been developed at the SNRA as a means of improving data quality. The registration task of 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 facing the valves. If a valve, for instance, becomes filled with rainwater, or squeezed by a vehicle wheel, it may quit working.

5.2. Analysis

In our analyses, for each site, the data set produced by M_0 represents the output from a regular measurement. The data set produced jointly by M_1 and M_2, on the other hand, represents the 'truth.'

If the data collection had turned out perfectly, the data sets from M_1 and M_2 would have been complete. Some missing data, and hence some imputed vehicles, did however arise also in the valve measurements. In certain cases, imputations in the M_1 or M_2 data could be matched with vehicles properly registered by M_0. These situations were most likely to occur when passing vehicles straddled 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 replaced by the registered vehicles. Since this matching was not always possible, we perform our analyses with remaining imputations retained as well as removed (in other words, we work with two sets of "true" values).

5.2.1. The multiplicative imputation error model

When the observed ϵ_{kh}'s are plotted against the number of missing vehicles, for some sites there is a tendency of the error variance to decrease as the number of missing vehicles

increases (which contradicts the model) – an example is given in Figure 1. Due to 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 numbers of registered vehicles, 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 ANOVA model:

$$\hat{\varepsilon}_{kh} = \alpha + \beta_k + e_{kh} \begin{cases} k = 1, 2, \ldots, b \\ h = 1, 2, \ldots, c \end{cases} \tag{10}$$

where 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 normally and independently distributed (NID) with mean zero and variance σ_β^2, the e_{kh}'s $NID(0, \sigma_e^2)$, and that β_k and e_{kh} are independent. This random effects model (see, for instance, Montgomery (1997, Sections 3–7)) 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.

When we test the hypothesis $H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$, our conclusions differ for different treatments of the imputations in the valve measurements. If the imputations are removed, the null hypothesis is not rejected at the .05 level of significance. If, on the other hand, the imputations are retained, the null hypothesis is rejected. Hence, we do not get a clear indication as to whether there is a variability between sites or not. We are further interested in the mean $\mu_{\hat{\varepsilon}} = \alpha$ of $\hat{\varepsilon}_{kh}$. When we test $H_0 : \mu_{\hat{\varepsilon}} = 1$ versus $H_1 : \mu_{\hat{\varepsilon}} \neq 1$, if the imputations are removed, the null hypothesis is not rejected at the .05 level of significance. If, on the other hand, the imputations are retained, the hypothesis is rejected. Thus, it remains an open question whether the number of imputed vehicles is conditionally unbiased for the true number of missing vehicles or not.

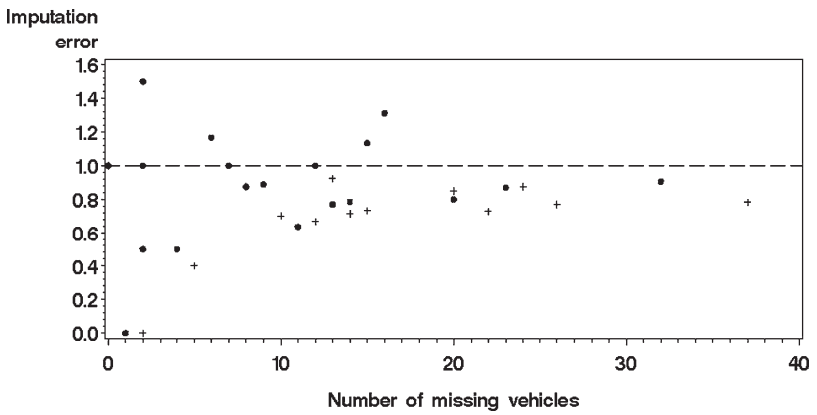

Fig. 1. Estimated imputation errors versus "true" numbers of missing vehicles for Site 4. One data point corresponds to one hour. Dots indicate that "true" values with imputations removed are used; plus signs indicate that "true" values with imputations retained are used

5.2.2. The error model for $\hat{\theta}_{kh}^{(2)}$

Under both the additive and the multiplicative error model for $\hat{\theta}_{kh}^{(2)}$, the observed ε_{kh}'s seem independent of the "true" probabilities as well as of the numbers of registered vehicles.

Both the additive and the multiplicative error model state that the variance of the errors is independent of the site. To investigate this, we use the same ANOVA model as in Equation 10 – only with $\hat{\varepsilon}_{kh}$ replaced by $\hat{\varepsilon}_{kh}$. Again, the aim is to test the hypothesis $H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$. No matter how the imputations in the valve measurements are treated, or if the model is additive or multiplicative, the null hypothesis is rejected at the .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 $\hat{\theta}_{kh}^{(2)}$. This objection to the model requires further investigation.

We also tested if $\hat{\theta}_{kh}^{(2)}$ is unbiased for θ_{kh}. At the .05 level of significance for the additive error model, the hypothesis of $\mu_{\hat{\varepsilon}} = 0$ (versus $\mu_{\hat{\varepsilon}} \neq 0$) is not rejected. Also, for the multiplicative model, the hypothesis of $\mu_{\hat{\varepsilon}} = 1$ (versus $\mu_{\hat{\varepsilon}} \neq 1$) is not rejected. These results stand no matter how the imputations in the valve measurements are treated.

5.2.3. Empirical behavior of estimators of y_k, z_k and u_k

Our limited data do not allow us to study the long-term performances of the estimators of flow and travel time, but can give some indication of them. In Table 1, estimates of y_k, z_k and $u_k = y_k/z_k$ are standardized by the "true" figures (the averages are taken over the five studied sites). The estimates obtained when we standardize by the "true" values with imputations *removed* are presented in the column 'Without imputations'; the estimates obtained when we standardize by the "true" values with imputations *retained* are presented in the column 'With imputations.' The estimates of z_k are only standardized by the "true" values with imputations removed, since we do not trust the travel times of imputed vehicles.

As expected, the Strategy 0 estimates of y_k and z_k all fall below one, whereas the missing data adjusted estimates under Strategy 1 and Strategy 2 look quite well. Depending on what entity is used for standardization (whether the imputations in the valve measurements are included or not), for both strategies, the average estimates land slightly below or above one. In all, it is far from obvious which adjustment strategy (1 or 2) ought to be recommended.

The ratio u_k is the counterpart on element-level to R. All standardized estimates of u_k in Table 1 are very close to one. Formally, we cannot use these estimates to evaluate the performances of present or proposed estimators of R. We take the result, however, as a small hint that missing data adjustments are not a necessity when estimating R.

Table 1. Average standardized estimates of y_k, z_k and u_k

	Without imputations			With imputations		
	Strategy			Strategy		
Parameter	0	1	2	0	1	2
y_k	0.94601	1.01196	1.01530	0.93012	0.99295	0.99631
z_k	0.94039	1.00821	1.01142	–	–	–
u_k	1.01308	1.00146	1.00199	0.99720	0.98576	0.98629

6. Simulation Study

The experiment accounted for in Section 5 was restricted to a small number of road sites, and hence gave far from conclusive results. Since we lacked the necessary resources to perform a larger experiment, we tried to make the most of our data by using them for a simulation study. From the "true" vehicle data (imputations removed), for road sites Nos. 1, 2, 4, and 5 (Site 3 was excluded due to a large amount of missing data in its valve measurements), stratified Bernoulli sampling (STBE) was used to select vehicles. The stratification was by watch-hour, and the sampling procedure repeated independently and with replacement $D = 1{,}000$ times. The set of vehicles obtained in a given repetition was meant to represent a successfully observed subset of all vehicles passing the site. Within watch-hour h and a road site k, three probabilites of selection θ_{kh} were applied: the proportion p_{kh} of registered vehicles in the original experiment; $.9p_{kh}$ and $.8p_{kh}$. In addition, at Site 2 (the site with the highest traffic flow: nearly 15,000 vehicles during the observed 24-hour period), we tried a worst-case scenario: the registration probabilites $.8p_{kh}$ during off-peak hours (defined as hours with a "true" flow of less than 1,000 vehicles), otherwise $.5p_{kh}$.

Let $\hat{y}_{hkd}^{(c)}$ denote the estimated flow for watch-hour h, road site k and repetition d under Strategy c, and let $\hat{y}_{kd}^{(c)} = \sum_{h=1}^{24} \hat{y}_{hkd}^{(c)}$. The travel time estimates $\hat{z}_{hkd}^{(c)}$ and $\hat{z}_{kd}^{(c)}$ are defined correspondingly. For each site and choice of registration probability, and for Strategy $c = 0$ and 2, we estimate u_k by

$$\hat{u}_k^{(c)} = \frac{1}{D} \sum_{d=1}^{D} \hat{u}_{kd}^{(c)} \tag{11}$$

where $\hat{u}_{kd}^{(c)} = \hat{y}_{kd}^{(c)} / \hat{z}_{kd}^{(c)}$. An approximation value of the variance of $\hat{u}_k^{(c)}$ is calculated as

$$V\left(\hat{u}_k^{(c)}\right) = \frac{1}{D} \sum_{d=1}^{D} \left(\hat{u}_{kd}^{(c)} - \hat{u}_k^{(c)}\right)^2 \tag{12}$$

Strategy 1 is omitted from the study since we have no way of knowing the number of vehicles that would have been imputed in replacement for those missing each hour in each data set. In our Strategy 2 estimates of u_k, we have to use a rough value on the ME: the ME obtained if each missing vehicle generates exactly as many pulses as it has wheels.

The estimates are presented in Table 2. We see that all estimates of u_k come very close to the true value, and that the variability in the estimates is strikingly small (even in our worst-case scenario).

7. Summary

We have put forward two possible strategies for missing data adjustments in the speed survey. In our investigation of the resulting estimators' theoretical properties, we make use of several models. Most of the model assumptions seem to agree reasonably well with our experimental data. Also, the proposed estimators seem to produce less biased estimates of t_y and t_z than today's unadjusted estimators. This is however not necessarily true for R: the present unadjusted estimator of average speed may be surprisingly resistant to bias due to missing data. We thus conclude that when the aim is to estimate total vehicle mileage or total travel time, Strategy 1 or Strategy 2 should be used to adjust for missing data,

Table 2. Estimates of u_k under Strategy 0 and Strategy 2, by site, based on 1,000 STBE samples. $*.8p_{kh}$ during off-peak hours; else $.5p_{kh}$

Site no.	θ_{kh}	u_k	$\hat{u}_k^{(0)}$	$V(\hat{u}_k^{(0)})$	$\hat{u}_k^{(2)}$	$V(\hat{u}_k^{(2)})$
1	p_{kh}	52.44	52.44	.00011	52.44	.00012
1	$.9p_{kh}$	52.44	52.44	.00162	52.44	.00163
1	$.8p_{kh}$	52.44	52.44	.00334	52.44	.00330
2	p_{kh}	52.18	52.26	.00038	52.18	.00038
2	$.9p_{kh}$	52.18	52.26	.00115	52.18	.00113
2	$.8p_{kh}$	52.18	52.26	.00207	52.18	.00198
2	$*$	52.18	52.68	.00333	52.18	.00478
4	p_{kh}	45.37	45.37	.00008	45.37	.00008
4	$.9p_{kh}$	45.37	45.37	.00053	45.37	.00052
4	$.8p_{kh}$	45.37	45.37	.00115	45.37	.00112
5	p_{kh}	54.31	54.32	.00013	54.31	.00013
5	$.9p_{kh}$	54.31	54.32	.00063	54.31	.00061
5	$.8p_{kh}$	54.31	54.32	.00130	54.31	.00126

whereas when the aim is to estimate the average speed, missing data adjustments seem unnecessary (Strategy 0 is just as good for estimating average speed as Strategies 1 or 2).

8. References

Chapman, D.W. (1976). A Survey of Nonresponse Imputation Procedures. Proceedings of the American Statistical Association, Social Statistics Section, 245–251.

Drew, J.H. and Fuller, W.A. (1980). Modeling Nonresponse in Surveys with Callbacks. Proceedings of the American Statistical Association, Section on Survey Research Methods, 639–642.

Drew, J.H. and Fuller, W.A. (1981). Nonresponse in Complex Multiphase Surveys. Proceedings of the American Statistical Association, Section on Survey Research Methods, 623–628.

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. Journal of Official Statistics, 7, 325–337.

Giommi, A. (1987). Nonparametric Methods for Estimating Individual Response Probabilities. Survey Methodology, 13, 127–134.

Isaksson A. (2003) Survey Models for a Vehicle Speed Survey, Doctoral Thesis. Linköping Studies in Statistics No.2, Linköping University, Sweden.

Montgomery, D.C. (1997). Design and Analysis of Experiments. 4th ed. New York: Wiley.

Raj, D. (1968). Sampling Theory. New York: McGraw-Hill.

Särndal, C.-E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review, 55, 279–294.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer.

Received November 2002
Revised June 2004