Dept. of Math. University of Oslo
Statistical Research Report No. 9
ISBN 82–553–1322–2 December 2001
ISSN 0806–3842

Combining two estimates applied in a survey of copyright volumes at higher educational institutions in Norway using the bootstrap

B. Storvik[1], B. Natvig[1,2], M. Aldrin[1]

ABSTRACT:

This survey is intended for estimating several types of copyright volume at different educational institutions in Norway. Using a calibration factor being equal to the ratio of the true number of machine pages taken from all machines at an institution to the corresponding estimated number, will make the estimates less biased and less variable. There are two reasonable estimates. In addition to suggesting a bootstrap procedure for selecting one of them, we propose to fit a weighted average of both. The weight is estimated through bootstraping by minimizing the mean square error of the combined estimate summed over all possible types of copyright material. We expect to assign more weight to the estimate which has less mean square error. The paper discusses ways of analyzing the bias and ways of estimating the variances of the estimates through the bootstrap. Analyzing the data by bootstraping gave significantly better performance for the combined estimate compared to using the best of the two estimates chosen by bootstrap selection. However, setting the weights equal to 0.5 gave the overall best performance. It is not straightforward to analyze theoretically the bias of the estimates. Such an analysis is given in an appendix for a fixed weighted average based on a simultanous model for the number of machine pages and the number of original pages taken by each person.

Keywords: survey sampling, bootstrap methods, adjusted estimates, combined estimates, calibration factor

1 Introduction

In this survey we want to measure the volume of copies and in particular copyright volume at higher educational institutions in Norway. The reason for doing such a survey is the extreme difficulty in measuring the copying habits of the whole population at these institutions. Copying material that is restricted by copyright is not allowed according to law, but in Norway deals can be made with the reproduction rights organization of Norway, Kopinor, to make such copies. Therefore, the higher educational institutions in Norway

[1]Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway.
[2]Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, N-0316 Oslo

pay Kopinor for copying copyright material and this payment is based on estimates from survey data. Kopinor then redistributes money received from the different institutions to the right sources according to the estimates of the survey. It should be mentioned that there are different types of copyright material to be estimated.

There are about 130 higher educational institutions in Norway. 5 of these institutions participated in the survey. The survey data from the participating institutions were used to estimate the copyright volumes at these institutions as well as at the remaining institutions.

Since different personnel groups may have different copying habits, the personnel at each institution were stratified into three groups: students, scientific staff and administrative/technical staff. From each of these three groups, persons were sampled according to a specific scheme. The total number of persons asked to participate in the present survey was 2600.

At three different sample periods during the year people were recruited into the survey. These people were asked to collect all their copies during a two-week period. Furthermore, these copies were delivered and classified into different types of material (e.g. non-copyright material, copyright material of type 1, type 2 and so on). It was important to specify two terms, one was the number of pages from the original material (originalpages) and the other was the number of pages the machine is counting (machinepages). The originalpages were defined from a code instruction note. It is too detailed to specify the whole note, but the general idea is that e.g. two pages in a book copied down to one page (A4) will be specified as two originalpages and one machinepage. We were interested in the number of originalpages of each material type from the whole population at the different institutions over a study year from summer to summer.

For each participating institution, we got preliminary estimates of the yearly total number of machinepages, as well as the yearly number of originalpages of each material type. Concerning the number of machinepages, the true number for each institution was found (almost) exactly by adding all machine volumes. Each machine has their own counting unit which registrates each copy. The ratio of the true (measured) and the preliminary estimated number of machinepages may be interpreted as an "underreporting" factor. A natural way to proceed is to multiply the preliminary estimated numbers of originalpages by this factor. Doing so will decrease the bias and variances of the proposed estimates, but at the same time a new bias effect will be introduced.

The paper will discuss ways of analyzing the bias and ways of estimating the variances of the estimates through the bootstrap. The paper is organized as follows. Section 2 presents the estimation procedure. Estimation of the proportion of persons present at the different institutions is treated in Section 3, whereas inference on institutions not participating in the survey is covered in Section 4. Numerical results and some concluding remarks are given in Section 5. In Appendix A we give bootstrap estimates for uncertainties. In Appendix B a theoretical bias discussion of the suggested estimates for a fixed weighted average is given by modeling the relationship between originalpages and machinepages. A discussion of a possible assumption in this appendix is given in Appendix C.

2 Estimation Procedure

This section will discuss methods of estimating the number of originalpages from stratified sampled data. The section is divided into 3 subsections. The first two subsections will motivate preliminary estimates and adjusted estimates by using a calibration factor. The last subsection will discuss ways of combining two adjusted estimates.

2.1 Preliminary estimates

Suppose the total number of machinepages is reported from an institution. In the following methodology capital letters will represent the population parameters and small letters represent the sampled parameters. Only section 4 will need an index for a specific institution. Therefore, for all other sections including this one, an index for a specific institution will not be in the notation. The number of persons in a sample varies with institution, personnel group, section and sample period. The numbers of persons in the sample and in the population are defined by

$$n_{lkp} = \text{The number of actual active participants in the sample in personnel group } l, \text{ at section } k \text{ and sample period } p \quad (1)$$

$$N_{lk} = \text{The number of persons in the population in personnel group } l \text{ and at section } k. \quad (2)$$

Let $G = 3$ be the number of personnel groups. Note that the number of sections that an institution is divided into is of course dependent on the institution. The number of copied originalpages and machinepages are variables which are dependent on personnel group l, section k, type of material m, sample period p, and person number s, that is

$$u_{lkmps} = \text{The number of copied originalpages in the sample in personnel group } l, \text{ at section } k, \text{ of a certain type of material } m \text{ at sample period } p \text{ and person number } s \text{ over a 14-days period} \quad (3)$$

$$v_{lkps} = \text{The number of copied machinepages in the sample of all types of material in personnel group } l, \text{ at section } k, \text{ at sample period } p \text{ and person number } s \text{ over a 14-days period.} \quad (4)$$

Obviously $v_{lkps} = 0$ implies $u_{lkmps} = 0$. Note that here we are only considering the persons included in the survey. The population parameters U_{lkmp} and V_{lkp} are defined as total yearly volume in the population of originalpages and machinepages respectively, where the indexes are as specified above. In estimating the total number of originalpages we also want to calibrate by using the information we collect on the total number of machinepages. This calibration will need an estimate of the total number of machinepages from the sampled data. Since the sampled data only consist of the machinepage volume for each person over a 14-day period, we will need a factor that blows this up to a yearly volume. This factor

being dependent on personnel group l and sample period p is denoted c_{lp}. This is a natural way to proceed since students have shorter semesters than the employed people at the institutions and some sample periods will represent a longer time of the year. The blow up factor c_{lp} is assumed only to depend on each type of person and the sample period. The following sample periods are covered:

1. Late in the semester (week 46 and 47 in year 2000).
2. Early in the semester (week 4 and 5 in year 2001).
3. In the middle of the semester (week 12 and 13 in year 2001).

The employed people ($l = 1$ and $l = 2$) will only be present at the institution (because of holidays) 23 times 2 weeks equal to 46 weeks a year. The students ($l = 3$) have shorter semesters and are present 19 times 2 weeks equal to 38 weeks a year. We assume that sample periods 1 and 2 represent 1/4 of the year present at the institution, while sample period 3 represents 1/2 of the year. This gives for instance weight $c_{11} = 23 \cdot 1/4 = 5.75$. All weights are given in Table 1.

Table 1: The blow up factor c_{lp}.

	$l=1$	$l=2$	$l=3$	$\bar{c}_p = \frac{\sum_{l=1}^{3} c_{lp}}{\sum_{p=1}^{3}\sum_{l=1}^{3} c_{lp}}$
$p=1$	$c_{11} = 5.75$	$c_{21} = 5.75$	$c_{31} = 4.75$	1/4
$p=2$	$c_{12} = 5.75$	$c_{22} = 5.75$	$c_{32} = 4.75$	1/4
$p=3$	$c_{13} = 11.5$	$c_{23} = 11.5$	$c_{33} = 9.50$	1/2
	$\sum_{p=1}^{3} c_{1p} = 23$	$\sum_{p=1}^{3} c_{2p} = 23$	$\sum_{p=1}^{3} c_{3p} = 19$	$\sum_{p=1}^{3} \bar{c}_p = 1$

We also need a factor that blows up the mean of the sampled machinepage volumes to the whole population. However, the whole population is not present at the institution at any time, because of sickness or absence for other reasons. Therefore we will use $W_{lp} \cdot N_{lk}$ to be the population present at an institution at sample period p, where W_{lp} is the proportion of persons present at the institution, and assumed only to depend on personnel group l and on sample period p. However, W_{lp} is unknown and has to be estimated by $\widehat{W}_{lp}$. How W_{lp} will be estimated is discussed in section 3.

The preliminary estimates of total yearly volume of originalpages (for type of material m) and machinepages in the population for personnel group l and section k at a specific institution are respectively

$$\widehat{U}_{lkm}^{pre} = \sum_{p=1}^{3} \widehat{W}_{lp}\, N_{lk}\, c_{lp}\, \bar{u}_{lkmp} \tag{5}$$

$$\widehat{V}_{lk}^{pre} = \sum_{p=1}^{3} \widehat{W}_{lp}\, N_{lk}\, c_{lp}\, \bar{v}_{lkp}, \tag{6}$$

where

$$\begin{aligned} \bar{u}_{lkmp} &= \tfrac{1}{n_{lkp}} \textstyle\sum_{s=1}^{n_{lkp}} u_{lkmps} \\ \bar{v}_{lkp} &= \tfrac{1}{n_{lkp}} \textstyle\sum_{s=1}^{n_{lkp}} v_{lkps}. \end{aligned}$$

There is also another intuitive way of preliminary estimating U_{lkm}. The mean of the number of originalpages can also be estimated by first taking the mean of the ratio of the originalpage volume to the machinepage volume, only for persons taking copies in the sample periods, and then multiplying this mean ratio by the mean of the machinepage volume. The intuitive reason for this estimate is the following. Taking the mean ratio and multiplying it by the mean of the machinepage volume, only for persons taking copies in the sample periods will give us mean originalpage volume, only for persons taking copies. In order to get the total mean originalpage volume, the originalpage volume only for persons taking copies is mulitiplied by the proportion of persons taking copies. Now, the mean of the machinepage volume only for persons taking copies multiplied by the proportion of persons taking copies is the mean of the machinepage volume. Therefore, the estimate is

$$\widetilde{U}_{lkm}^{pre} = \sum_{p=1}^{3} \widehat{W}_{lp}\, N_{lk}\, c_{lp}\, \bar{v}_{lkp} \bar{r}_{lkmp}, \tag{7}$$

where $\bar{r}_{lkmp} = \text{mean}_s \left\{ \frac{u_{lkmps}}{v_{lkps}} | v_{lkps} > 0 \right\}$ is the mean of the ratios between originalpages and machinepages, only for persons taking copies. The sums of $\widehat{U}_{lkm}^{pre}$, $\widetilde{U}_{lkm}^{pre}$ and $\widehat{V}_{lk}^{pre}$ in (5), (7) and (6) over sections and personnel groups are given by

$$\widehat{U}_{m}^{pre} = \sum_{l=1}^{G} \sum_{k} \widehat{U}_{lkm}^{pre} \tag{8}$$

$$\widetilde{U}_{m}^{pre} = \sum_{l=1}^{G} \sum_{k} \widetilde{U}_{lkm}^{pre} \tag{9}$$

$$\widehat{V}^{pre} = \sum_{l=1}^{G} \sum_{k} \widehat{V}_{lk}^{pre}. \tag{10}$$

$\widehat{U}_{m}^{pre}$ and $\widetilde{U}_{m}^{pre}$ are preliminary estimates of the total yearly volume of originalpages (for type of material m), whereas $\widehat{V}^{pre}$ is a preliminary estimate of the total yearly volume of machinepages.

2.2 Adjusted estimates

Note that we have observed the true value V of the total yearly volume of machinepages. Therefore it may seem unnecessary to introduce the preliminary estimate $\widehat{V}^{pre}$. However, for several reasons we can expect that the estimates $\widehat{V}^{pre}$, $\widehat{U}_{m}^{pre}$ and $\widetilde{U}_{m}^{pre}$ are biased with the same factor. One reason for this bias is that sampled persons may systematically tend

to underreport the number of copies by for instance simply forgetting to collect some of them. Another reason is that some persons that are temporarily visiting the institution without being registered and at the same time are taking copies will not be sampled. This means that even if all persons at an institution report their copies, there would still be an error from underreported machinepages. Therefore, a theoretical quantity of the total yearly volume of machinepages, possibly underreported, is defined by

$$V' = \sum_{l,k,p} W_{lp} c_{lp} \sum_{s=1}^{N_{lk}} v_{lkps}. \tag{11}$$

It is sensible to correct for such a bias by a calibration factor. Define a calibration factor by

$$O = \frac{V}{V'}. \tag{12}$$

An obvious estimate for this factor is

$$\widehat{O} = \frac{V}{\widehat{V}^{pre}}. \tag{13}$$

This yields adjusted estimates of the total yearly volume of originalpages

$$\widehat{U}_m^{adj} = \widehat{O}\widehat{U}_m^{pre} \tag{14}$$

$$\widetilde{U}_m^{adj} = \widehat{O}\widetilde{U}_m^{pre}. \tag{15}$$

The typical effect of using the calibration factor $\widehat{O}$ is less bias and variance. However, using $\widehat{O}$ introduces a new bias, because $E\widehat{O} \neq O$. The reason for less variance is that the effects of small samples and using only 14-day periods to collect are diminished.

2.3 Combined estimates

Since we have two estimates of the total number of originalpages, it is possible to generalize them by taking their weighted mean, i.e.

$$\dot{U}_m^{com}(q) = q\widehat{U}_m^{adj} + (1-q)\widetilde{U}_m^{adj}, \tag{16}$$

where $q \in [0,1]$ is the weight. We could have thought of the weights in (16) also to be dependent on the type of material, but this would give more unstable weights due to little data for some types of material and is therefore not considered here. Define a new quantity which is (16) summed over all types of material, i.e.

$$\begin{aligned} \dot{U}^{com}(q) &= \sum_m \dot{U}_m^{com}(q) \\ &= q\widehat{U}^{adj} + (1-q)\widetilde{U}^{adj}, \end{aligned} \tag{17}$$

where $\widehat{U}^{adj} = \widehat{O}\sum_m \widehat{U}_m^{pre}$ and $\widetilde{U} = \widehat{O}\sum_m \widetilde{U}_m^{pre}$, with true value U.

We have to realize at this stage that the weights q in (16) have to be estimated from the available data. Below we will introduce two different criteria for estimating q. In the first, q is restricted to be either 0 or 1, i.e. we use either $\widehat{U}_m^{adj}$ or $\widetilde{U}_m^{adj}$. Furthermore, criterion one and two are based on minimizing the mean squared error. The two methods are

$$q_1^\star = \begin{cases} 1 & \text{if } \mathrm{Var}(\widehat{U}^{adj}) + \mathrm{bias}(\widehat{U}^{adj})^2 < \mathrm{Var}(\widetilde{U}^{adj}) + \mathrm{bias}(\widetilde{U}^{adj})^2 \\ 0 & \text{if } \mathrm{Var}(\widehat{U}^{adj}) + \mathrm{bias}(\widehat{U}^{adj})^2 \geq \mathrm{Var}(\widetilde{U}^{adj}) + \mathrm{bias}(\widetilde{U}^{adj})^2, \end{cases} \tag{18}$$

$$q_2^\star = \mathrm{argmin}_q \left\{ \mathrm{Var}(\dot{U}^{com}(q)) + \mathrm{bias}^2(\dot{U}^{com}(q)) \right\} \tag{19}$$

$$= \frac{\mathrm{Varb}(\widetilde{U}^{adj}) - \mathrm{Covb}(\widehat{U}^{adj}, \widetilde{U}^{adj})}{\mathrm{Varb}(\widehat{U}^{adj}) + \mathrm{Varb}(\widetilde{U}^{adj}) - 2\mathrm{Covb}(\widehat{U}^{adj}, \widetilde{U}^{adj})},$$

where $\mathrm{bias}(\widehat{U}^{adj}) = E(\widehat{U}^{adj}) - U$, $\mathrm{bias}(\widetilde{U}^{adj}) = E(\widetilde{U}^{adj}) - U$, $\mathrm{Varb}(\widehat{U}^{adj}) = \mathrm{Var}(\widehat{U}^{adj}) + \mathrm{bias}^2(\widehat{U}^{adj})$, $\mathrm{Varb}(\widetilde{U}^{adj}) = \mathrm{Var}(\widetilde{U}^{adj}) + \mathrm{bias}^2(\widetilde{\widetilde{U}})$ and $\mathrm{Covb}(\widehat{U}^{adj}, \widetilde{U}^{adj}) = \mathrm{Cov}(\widehat{U}^{adj}, \widetilde{U}^{adj}) + \mathrm{bias}(\widehat{U}^{adj})\mathrm{bias}(\widetilde{U}^{adj})$.

The optimal solutions of (18) and (19) depend on the unknown quantities: $\mathrm{Var}(\widehat{U}^{adj})$, $\mathrm{Var}(\widetilde{U}^{adj})$, $\mathrm{Cov}(\widehat{U}^{adj}, \widetilde{U}^{adj})$, $\mathrm{E}(\widehat{U}^{adj})$, $\mathrm{E}(\widetilde{U}^{adj})$ and U. We will estimate these quantities by the so-called bootstrap; see Efron & Tibshirani (1993) or Davison &Hinkley (1997). If we also want the variance of the estimates $\dot{U}_m^{com}(q)$ or $\dot{U}^{com}(q)$ in (16) and (17), where the weight is estimated by the bootstrap, a so-called double bootstrap has to be used, see Appendix A. The quantity $q_2^\star$, or the bootstrap estimate is not guaranteed to lie in the interval $[0, 1]$. Therefore, we define a new truncated version of the weight as

$$q_2^{trunc} = \max(0, \min(1, q_2^\star)). \tag{20}$$

Define $q_1^{trunc} = q_1^\star$. A weighted average will then give the following estimates:

$$\dot{U}_m^{com}(q_i^{trunc}) = q_i^{trunc}\widehat{U}_m^{adj} + (1 - q_i^{trunc})\widetilde{U}_m^{adj} \quad i \in \{1, 2\}. \tag{21}$$

Another advantage of using the calibration factor is that a multiplicative constant in c_{lp} will not alter the estimate $\dot{U}_m^{com}(q_i^{trunc})$ in (21). This is seen from (5) to (21).

3 Estimating the proportion of persons present at the institution

In section 2 we introduced the factor:

W_{lp} = the proportion of persons in personnel group l present at the institution in sample period p.

We will now present an unbiased estimate of W_{lp}. Note that the ones asked to participate in the survey are not the same as the actual participants, due to persons dropping out, sickness and so on. Define the following 3 non overlapping groups:

1 = Actual active participants

2 = Persons who have denied to participate and persons where no information is available due to for instance no response, unknown address or foreign language

3 = Persons (not in group 2) known to be on sabbatical leave or absent in the 14-days period due to for instance sickness

The persons in groups 1 and 2 are present at the institution in the survey period. Let n_{lp}^1, n_{lp}^2 og n_{lp}^3 denote the number of persons asked to participate in the survey from personnel group l in sample period p in the 3 different groups respectively. This means that $n_{lp}^{tot} = n_{lp}^1 + n_{lp}^2 + n_{lp}^3 =$ total number of persons that are asked to participate in the survey from personnel group l in sample period p. Note the slight inconsistence in notation since $n_{lp}^1 = \sum_k n_{lkp}$, where n_{lkp} is defined in (1). Furthermore, $100 \times (n_{lp}^1 + n_{lp}^3)/(n_{lp}^1 + n_{lp}^2 + n_{lp}^3)$ is the actual participation percentage. There will, certainly, be no copies taken by persons that come from group 3. We assume that group 2 has the same copying habits as group 1. An estimate of the proportion of persons present at the institution is then given by:

$$\widehat{W}_{lp} \;=\; (n_{lp}^1 + n_{lp}^2)/(n_{lp}^1 + n_{lp}^2 + n_{lp}^3).$$

Suppose N_{lp}^{tot}, N_{lp}^1, N_{lp}^2 and N_{lp}^3 are population counterparts to the survey sampled n_{lp}^{tot}, n_{lp}^1, n_{lp}^2 og n_{lp}^3. It is assumed that the whole population $N_{lp}^{tot} = N_l^{tot}$ is constant during the year of observation. However, the number of persons present $N_{lp}^1 + N_{lp}^2$ will change during the year. It is then seen that

$$n_{lp}^1 + n_{lp}^2 \;\sim\; \text{Hypergeometric}(N_l^{tot}, N_{lp}^1 + N_{lp}^2, N_{lp}^3)$$

and

$$\mathrm{E}(\widehat{W}_{lp}) \;=\; W_{lp} \;=\; \frac{N_{lp}^1 + N_{lp}^2}{N_l^{tot}}, \qquad \mathrm{Var}(\widehat{W}_{lp}) \;=\; \frac{W_{lp}(1 - W_{lp})}{n_{lp}^{tot}} \frac{(N_l^{tot} - n_{lp}^{tot})}{N_l^{tot} - 1}. \tag{22}$$

4 Inference on institutions not in the survey

Based on results from the 5 institutions in the survey, inference is deducted to other institutions. Dependent on what kind of information that is gathered from these other intitutions, estimates of copyright volume is made. We will describe three different types of such information. The most desired type of information is of course the total number, V, of machinepages from the institution. Suppose each institution in the survey is denoted

by an index j. Furthermore, assume the estimate $\dot{U}^{com}_{jm}(q_j)$ in (16) is used and that the numbers of machinepages V_j for $j = 1, \ldots, 5$ are available. For this particular type of information the estimate of the number of originalpages of copyright material is:

$$V \sum_{j=1}^{5} p_j \frac{\dot{U}^{com}_{jm}(q_j)}{V_j},$$

where $\sum_{j=1}^{5} p_j = 1$ and the p_js indicate how much the different institutions in the survey should be weighted for the particular institution in mind. Some of the p_js, $j = 1, \ldots, 5$ will typically be equal to 0. How to decide on these weights was the concern of the representatives from the higher educational institutions and Kopinor. It is also possible to make estimates from the section, but then it has to be decided which section from the survey is relevant for comparison. This is difficult and the option is therefore not considered here.

If the number of machinepages at the institution in mind is not available, we will have to use the number of persons at the institution. There are several options here. First, assume N is the total number of persons at the institution. Correspondingly, let N_j be the total number of persons at the jth institution in the survey, $j = 1, \ldots, 5$. In this case the natural estimate is

$$N \sum_{j=1}^{5} p_j \frac{\dot{U}^{com}_{jm}(q_j)}{N_j}.$$

The last case we will consider is when the institution in mind provides N^l, the total number of persons of personnel type l. Let correspondingly N_{jl} be the total number of persons of personnel type l at the jth institution, $j = 1, \ldots, 5$ respectively. Accordingly, the natural estimate is

$$\sum_{j=1}^{5} p_j \sum_{l=1}^{G} N^l \frac{\dot{U}^{com}_{jlm}(q_j)}{N_{jl}},$$

where $\dot{U}^{com}_{jlm}(q_j) = \sum_k \dot{U}^{com}_{jlkm}(q_j)$, with an obvious definition of $\dot{U}^{com}_{jlkm}(q_j)$.

5 Numerical results and concluding remarks

We will give a summary of some of the numerical results in this section. The numerical results include three of the five institutions, because two institutions at the present time have not reported the total number of machinepages. From the simulations presented in Table 2 it is seen by only comparing rmse or rmse/estimate in this table that $\widetilde{U}^{adj}$ gives

Table 2: Total yearly copy volume of originalpages from the three institutions. The results are given as estimate, bias, standard deviation (sd), root mean squared error (rmse), scaled standard deviation and scaled root mean squared error. The numbers are in 10000 copies. The two different methods for estimating q in $\dot{U}^{com}(q)$ are given. Estimating q requires double bootstrap (5000 bootstrap samples in the inner loop and 1000 in the outer loop). The other method uses 1000 bootstrap samples (same as the outer loop above).

	estimate	bias	sd	rmse	sd/estimate	rmse/estimate
Institution 1						
$\widehat{U}^{adj}$	216.56	2.47	28.77	28.88	0.133	0.133
$\widetilde{U}^{adj}$	241.92	0.62	26.09	26.1	0.108	0.108
$\dot{U}^{com}(0.5)$	229.24	1.55	25.04	25.09	0.109	0.109
$\dot{U}^{com}(\widehat{q}_1^{trunc})$	241.92	-8.25	32.89	33.91	0.136	0.140
$\dot{U}^{com}(\widehat{q}_2^{trunc})$	234.03	-2.62	26.69	26.82	0.114	0.115
Institution 3						
$\widehat{U}^{adj}$	43.81	-0.20	5.15	5.15	0.117	0.118
$\widetilde{U}^{adj}$	36.30	0.26	3.42	3.43	0.094	0.094
$\dot{U}^{com}(0.5)$	40.06	0.03	3.93	3.93	0.098	0.098
$\dot{U}^{com}(\widehat{q}_1^{trunc})$	36.30	0.32	3.60	3.62	0.099	0.100
$\dot{U}^{com}(\widehat{q}_2^{trunc})$	36.36	0.87	4.00	4.09	0.11	0.113
Institution 4						
$\widehat{U}^{adj}$	6.95	0.04	1.03	1.03	0.148	0.148
$\widetilde{U}^{adj}$	5.95	0.06	0.86	0.86	0.144	0.144
$\dot{U}^{com}(0.5)$	6.45	0.05	0.90	0.90	0.139	0.140
$\dot{U}^{com}(\widehat{q}_1^{trunc})$	5.95	0.08	0.91	0.91	0.153	0.154
$\dot{U}^{com}(\widehat{q}_2^{trunc})$	6.00	0.10	0.93	0.94	0.155	0.156

Table 3: The calibration factor $\widehat{O}$ from three institutions. sd is the standard deviation and rmse is the root mean squared error or square root of square bias plus the variance. The bias, sd and rmse estimates are evaluated from 5000 bootstrap samples.

Institution	estimate	bias	sd	rmse
1	2.188	0.030	0.263	0.265
3	1.765	0.034	0.245	0.247
4	2.439	0.031	0.278	0.279

better performance than $\widehat{U}^{adj}$ for the tree institutions. However, data have to be used to find out which of the two estimates that gives the best overall performance. This means that data should be used to choose the method. Using the estimates $\dot{U}^{com}(\widehat{q}_1^{trunc})$ or $\dot{U}^{com}(\widehat{q}_2^{trunc})$ by first finding the bootstrap estimates of q_1^{trunc} or q_2^{trunc} respectively will include extra variability which is acounted for in Table 2 by using a double bootstrap. It is seen that choosing an estimated weighted mean (method 2) gives better overall performance than an estimated weight that is 0 or 1 (method 1) since it is definitely the best for institution 1. The conclusion from this survey is that $\dot{U}^{com}(q_2^{trunc})$ gives the best performance when q is estimated. However, an average between the two competing estimates ($\dot{U}^{com}(0.5)$) will give better overall performance for the three institutions.

Table 3 shows that the calibration factor $\widehat{O}$ is positively biased, but not much compared to the variance. How the calibration factor affects the bias and variance of the different competing estimates will now be discussed. First, the calibration factor $\widehat{O}$ is used as a multiplicative factor through $\widehat{U}^{adj} = \widehat{O}\widehat{U}^{pre}$ and $\widetilde{U}^{adj} = \widehat{O}\widetilde{U}^{pre}$. If it was possible to use O instead of $\widehat{O}$ as a multiplicative factor, both $\widehat{U}^{adj}$ and $\widetilde{U}^{adj}$ would have been unbiased, see Appendix B. If we assume that $\widehat{O}$ is independent of $\widehat{U}^{pre}$ and $\widetilde{U}^{pre}$, the biases of $\widehat{U}^{adj}$ and $\widetilde{U}^{adj}$ would also have been positive. However, the bootstrap bias of $\widehat{U}^{adj}$ and $\widetilde{U}^{adj}$ are not positive for all institutions, because there is an effect of dependence between $\widehat{O}$ and $\widehat{U}^{pre}$ and $\widetilde{U}^{pre}$ respectively. The estimates $\widehat{U}^{adj}$ and $\widetilde{U}^{adj}$ can be viewed as ratio estimates, where the numerator and denominator are dependent. In general, whether such ratio estimates give smaller or larger variance compared to the theoretical estimates $O\widehat{U}^{pre}$ and $O\widetilde{U}^{pre}$ is dependent on the problem.

A serious problem is the fact that the size of the calibration factor is about 2 (se Table 3), which means that there is a huge amount of underreporting when persons participate and hand in their copies. The underreporting does not necessarily explain the whole picture. As explained before, other reasons could be persons temporarily visiting the institution and that not registered persons take copies which will not be sampled. This is a serious drawback of this kind of survey. Another serious drawback of the survey is that persons not participating in the survey can have different copying habits than the ones participating. Table 4 shows the actual participation percentage, which is quite good compared to the survey reported in Zhang et al. (1999) which had about 50 percent participation.

Table 4: The actual participation percentage from the three institutions weighted by $\bar{c}_p$ for the different sample periods.

Institution	$l = 1$	$l = 2$	$l = 3$
1	55	62	69
3	65	81	66
4	88	86	73

Acknowledgments. The paper was supported by the The Research Council of Norway through the project "Knowledge, Data and Decision - Modern Statistics in Action".

A Bootstrap estimates for uncertainties

In the survey reported in Zhang et al. (1999), the variance estimates were based on rude approximations. It is, however, possible to use simulation tools to estimate the variance. One well known method is bootstraping where the data are resampled in a certain way. In this survey we will have to resample for each type of personnel group, in every section and from each of the three sample periods the data are sampled, since we must condition on all that is known. The way resampling is done, is by drawing persons from the sampled data *uniformly with replacement* in such a way that the number of new pseudodata is equal to the number of persons in the original sampled data. This procedure is repeated until we have B sets of pseudodata. The estimate in mind is calculated by using the pseudodata in the same way as the original data were used to produce the estimate. This means we will get B estimates for the estimates we are interested in. If B is large enough (usually $B = 1000$ is sufficient), the B estimates can be used to approximate the distribution of the estimates that used the original data set. This means we can construct the standard deviation in the usual way from the B estimates. In practice this will add an extra loop when we write the code for estimation.

Note that we will have to be careful by using the ordinary bootstrap as described above. This is the case since the population is finite and when the survey data are drawn from this population the data will be dependent. This dependence is not present if the survey data are drawn with replacement from the original data set. Let N be the number of persons in the population and n be the number of persons in the survey. If $p = n/N$ is small, the dependence can be neglected and we can use the method described above. If $p > 0.1$ (rule of thumb), the dependence should not be ignored. How can such dependence be brought into the sampling scheme? One option is to make a pseudopopulation of size N and then draw without replacement a sample from the pseudopopulation.
Algorithm 1.

1. Draw uniformly with replacement from the original sample until we have N data points. This is the pseudopopulation.

2. Draw without replacement from the pseudopopulation in 1. until we have n data

points. This is the pseudosample.

3. Repeat 1. and 2. B times.

This method is the so-called "superpopulation bootstrap", see Davison &Hinkley (1997). The total number of persons over all sample periods is equal to the total number of persons in the original sample, i.e. $\sum_{p=1}^{3} n_{lkp} = n_{lk}$, where n_{lkp} is defined in (1). The estimates we are interested in use u_{lkmps}, v_{lkps} and $\widehat{W}_{lp}$. In each sample period p, the persons were drawn uniformly with replacement among all persons in the sample in this period until we reached the population number of persons in the sample period. Since each sample period is assumed to have the same population, this results in a pseudopopulation of size N_{lk}. We then draw without replacement from this pseudopopulation until we have n_{lkp} data points of originalpages and machinepages. We believe that the proportion of persons present in a specific personnel group, section and in a specific sample period is independent of the copying habits in the same personnel group, section and sample period. Therefore, independently, the same procedure is followed, but now the focus is on the number of persons in the three nonoverlapping groups of participation (see section 3). The sampling procedure described above is continued until we have B such samples of the type $\{u^{\star}_{lkmps}, v^{\star}_{lkps}, \widehat{W}^{\star}_{lp}\}$. Based on these bootstrap samples it is possible to estimate $\{\widehat{U}^{adj,\star}(b), \widetilde{U}^{adj,\star}(b)\}_{b=1}^{B}$. The bootstrap variance, covariance, bias and expectation estimates are given by

$$\begin{aligned}
\widehat{\mathrm{Var}}_{\star}(\widehat{U}^{adj}) &= \frac{1}{B}\sum_{b=1}^{B}(\widehat{U}^{adj,\star}(b) - \widehat{U}^{adj,\star})^2 \\
\widehat{\mathrm{Cov}}_{\star}(\widehat{U}^{adj}, \widetilde{U}^{adj}) &= \frac{1}{B}\sum_{b=1}^{B}(\widehat{U}^{adj,\star}(b) - \widehat{U}^{adj,\star})(\widetilde{U}^{adj,\star}(b) - \widetilde{U}^{adj,\star}) \\
\widehat{\mathrm{bias}}_{\star}(\widehat{U}^{adj}) &= \widehat{U}^{adj,\star} - \widehat{U}^{adj} \\
\widehat{\mathrm{E}}_{\star}(\widehat{U}^{adj}) &= \widehat{U}^{adj,\star},
\end{aligned}$$

where $\widehat{U}^{adj,\star} = \frac{1}{B}\sum_{b=1}^{B}\widehat{U}^{adj,\star}(b)$. The bootstrap quantities for $\widetilde{U}^{adj}$ are evaluated in the same way as for $\widehat{U}^{adj}$. Suppose we want to estimate (by bootstrap) the quantity $q_2^{\star}$ in (19). It then follows that

$$\widehat{q}_2^{\star} = \frac{\widehat{\mathrm{Varb}}_{\star}(\widetilde{U}^{adj}) - \widehat{\mathrm{Covb}}_{\star}(\widehat{U}^{adj}, \widetilde{U}^{adj})}{\widehat{\mathrm{Varb}}_{\star}(\widehat{U}^{adj}) + \widehat{\mathrm{Varb}}_{\star}(\widetilde{U}^{adj}) - 2\widehat{\mathrm{Covb}}_{\star}(\widehat{U}^{adj}, \widetilde{U}^{adj})},$$

where $\widehat{\mathrm{Varb}}_{\star}(\widehat{U}^{adj}) = \widehat{\mathrm{Var}}_{\star}(\widehat{U}^{adj}) + \widehat{\mathrm{bias}}^2_{\star}(\widehat{U}^{adj})$, $\widehat{\mathrm{Varb}}_{\star}(\widetilde{U}^{adj}) = \widehat{\mathrm{Var}}_{\star}(\widetilde{U}^{adj}) + \widehat{\mathrm{bias}}^2_{\star}(\widetilde{U}^{adj})$ and $\widehat{\mathrm{Covb}}_{\star}(\widehat{U}^{adj}, \widetilde{U}^{adj}) = \widehat{\mathrm{Cov}}_{\star}(\widehat{U}^{adj}, \widetilde{U}^{adj}) + \widehat{\mathrm{bias}}_{\star}(\widehat{U}^{adj})\widehat{\mathrm{bias}}_{\star}(\widetilde{U}^{adj})$. The estimate using the bootstrap estimate of q will then be $\dot{U}^{com}(\widehat{q}_2^{trunc})$. To be able to estimate the variance and bias of $\dot{U}^{com}(\widehat{q}_2^{trunc})$ a so called double bootstrap has to be done. Each bootstrap sample is treated like the original data set and the above procedure is followed. In the end a double bootstrap sample will be available and variance and bias can be evaluated. The other method is evaluated in the same way.

B Theoretical bias discussion of the estimates for a fixed weighted average

In this appendix we will consider the estimate defined in (16), with q fixed. This estimate uses both information on originalpages and machinepages (estimated number and total number). The properties of such an estimate is dependent on the relationship between originalpages and machinepages. In this section we will try to model such a relationship. It is helpful to first define a few quantities:

$$\alpha_{lkmp} = \frac{U_{lkmp}}{U_m} \tag{23}$$

$$b_{lkp} = \frac{V_{lkp}}{V} \tag{24}$$

$$a_{lkmp} = \alpha_{lkmp}/b_{lkp}, \tag{25}$$

where

$$V_{lkp} = O\, W_{lp} c_{lp} \sum_{s=1}^{N_{lk}} v_{lkps} \tag{26}$$

$$U_{lkmp} = O\, W_{lp} c_{lp} \sum_{s=1}^{N_{lk}} u_{lkmps} \tag{27}$$

are the total yearly volumes of machinepages and originalpages of one type of material m for personnel group l, section k and sample period p respectively (assuming the same underreporting factor O for the machinepages and the originalpages. The quantity α_{lkmp} in (23) can be looked upon as the ratio of the number of copied originalpages in one personnel group, in one section, of one type of material and at one sample period of the total number of copied originalpages of this type of material at the institution. The quantity b_{lkp} in (24) is the proportion of the number of copied machinepages of all types of material in one personnel group, at a specific section and at one sample period of the total number of copied machinepages of all types of material at the institution. Note that the quantity a_{lkmp} is the ratio of the two mentioned proportions of the number of copied originalpages and machinepages at a specific institution in one personnel group, at a specific section, of one type of material and in one sample period. From (23), (24) and (25) it follows that

$$\sum_{p=1}^{3} \sum_{l=1}^{G} \sum_{k} a_{lkmp} V_{lkp} = V \tag{28}$$

$$\frac{U_{lkmp}}{V_{lkp}} = a_{lkmp} \frac{U_m}{V}. \tag{29}$$

Further, define:

$$E[u_{lkmps}] = \frac{1}{N_{lk}} \sum_{s'=1}^{N_{lk}} u_{lkmps'} \tag{30}$$

$$E[v_{lkps}] = \frac{1}{N_{lk}} \sum_{s'=1}^{N_{lk}} v_{lkps'} \tag{31}$$

$$\epsilon_{lkmps} = \frac{\frac{u_{lkmps}}{v_{lkps}}}{\frac{U_{lkmp}}{V_{lkp}}} = \frac{\frac{u_{lkmps}}{U_{lkmp}}}{\frac{v_{lkps}}{V_{lkp}}} = \frac{\frac{u_{lkmps}}{Eu_{lkmps}}}{\frac{v_{lkps}}{Ev_{lkps}}}, \quad s = 1, \ldots, N_{lk}, \tag{32}$$

where $v_{lkps} > 0$, and

$$\eta_{lkps} = \frac{O\, W_{lp}\, N_{lk} c_{lp} v_{lkps}}{V_{lkp}} = \frac{v_{lkps}}{Ev_{lkps}}, \quad s = 1, \ldots, N_{lk}, \tag{33}$$

where O is defined in (12). ϵ_{lkmps} is a non-negative variable that can be looked upon as a ratio of two proportions. The numerator in the ratio is the proportion of the number of copied originalpages of one type of material to the number of copied machinepages of all types of material of one specific person in one personnel group, in one section and in one sample period. The denominator in the ratio is the proportion of the total number of copied originalpages of one type of material of the total number of machinepages of all types of material, over all persons in one personnel group, in one section and in one sample period. Therefore, this ratio says something about skewness/difference in copying types of pages for different individuals. η_{lkps} is the ratio of the yearly machinepage volume of one specific person in one personnel group, in one section and at one sample period to the total yearly machinepage volume over all individuals in the same personnel group, section and sample period. From (29) and (32) we get

$$\frac{u_{lkmps}}{v_{lkps}} = a_{lkmp} \frac{U_m}{V} \epsilon_{lkmps}, \quad s = 1, \ldots, N_{lk}, \tag{34}$$

where $v_{lkps} > 0$. Note that assuming $a_{lkmp} = a_{lkp}$ is unreasonable, see Appendix C. Note also that (26), (27), (32), (33) and (34) are only based on definitions. It follows from (33) and (32) that

$$\begin{array}{lcccl} E[\eta_{lkps}] & = & \frac{1}{N_{lk}} \sum_{s'=1}^{N_{lk}} \eta_{lkps'} & = & 1 \\ E[\eta_{lkps}\epsilon_{lkmps}] & = & \frac{1}{N_{lk}} \sum_{s'=1}^{N_{lk}} \eta_{lkps'}\epsilon_{lkmps'} & = & 1, \end{array} \tag{35}$$

where $v_{lkps} > 0$. We now need the following conditions:

Conditions 1

i) ϵ_{lkmps} (where $v_{lkps} > 0$) and η_{lkps} are independent for $s = 1, \ldots, N_{lk}$

ii) ϵ_{lkmps} (where $v_{lkps} > 0$) and η_{lkps} are identically distributed for $s = 1, \ldots, N_{lk}$

Condition i) means that the ratio of the number of copied originalpages of one type of material to the number of copied machinepages of all types of materials is independent of the number of copied machinepages of all types of material for each individual in one personnel group, in one section and in one sample period, which is reasonable.

The sampling scheme is as follows: persons are drawn independently from the whole population within personnel group l. From (35) and conditions 1, the errors η_{lkps} and ϵ_{lkmps} (where $v_{lkps} > 0$) have mean 1 for $s = 1, \ldots, N_{lk}$.

We now consider the properties of the general estimate in (16), with q fixed. First, define

$$\bar{u}_{lkmp}(q) \quad = \quad q\bar{u}_{lkmp} + (1-q)\bar{v}_{lkp}\bar{r}_{lkmp}. \tag{36}$$

From (36) and (34)

$$\begin{aligned} & E(\bar{u}_{lkmp}(q)|\{v_{lkps}\}) \qquad (37) \\ &= qE(\bar{u}_{lkmp}|\{v_{lkps} > 0\}) + (1-q)\bar{v}_{lkp}E(\bar{r}_{lkmp}|\{v_{lkps} > 0\}) \\ &= q\bar{v}_{lkp}a_{lkmp}\frac{U_m}{V} + (1-q)\bar{v}_{lkp}a_{lkmp}\frac{U_m}{V} \\ &= \bar{v}_{lkp}a_{lkmp}\frac{U_m}{V}. \end{aligned}$$

Note that this conditional mean is independent of q. It is assumed that $\widehat{W}_{lp}$ and $\bar{u}_{lkmp}$ or $\bar{v}_{lkp}$ are independent.

It is seen from section 3 that $E(\widehat{W}_{lp}) = W_{lp}$. Suppose $\ddot{U}_m^{com}(q) = qO\widehat{U}_m^{pre} + (1-q)O\widetilde{U}_m^{pre}$. From (5),(7)-(9), (12), (36) and (37) it follows that:

$$\begin{aligned} & E(\ddot{U}_m^{com}(q)|\{v_{lkps}\}) \qquad (38) \\ &= V\frac{\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k E\widehat{W}_{lp}\, N_{lk}c_{lp}\, E(\bar{u}_{lkmp}(q)|\{v_{lkps}\})}{V'} \\ &= U_m\frac{\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k W_{lp}\, N_{lk}c_{lp}\, a_{lkmp}\, \bar{v}_{lkp}}{V'}. \end{aligned}$$

Since, with an obvious definition of $\widehat{V}_{lkp}^{pre}$, from (6) and (33), $OE\widehat{V}_{lkp}^{pre} = V_{lkp}$, the expectation of $E(\ddot{U}_m^{com}(q))$ is

$$\begin{aligned} & E[\ddot{U}_m^{com}(q)] \\ &= EE(\ddot{U}_m^{com}(q)|\{v_{lkps}\}) \\ &= U_m\frac{\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k W_{lp}\, N_{lk}c_{lp}\, a_{lkmp}\, E\bar{v}_{lkp}}{V'} \\ &= U_m\frac{\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k a_{lkmp}\, V_{lkp}}{OV'} \\ &= U_m\frac{V}{OV'} \\ &= U_m, \end{aligned}$$

having used (28) and (12). The difference between the estimates $\dot{U}^{com}(q)$ and $\ddot{U}^{com}(q)$ is that the last estimate is not based on an estimate of the calibration factor O while the former is. This introduces a bias to $\dot{U}^{com}(q)$, but not necessarily a higher variance because the estimated calibration factor $\widehat{O}$ defined by (13) is dependent on $\widehat{U}_m^{pre}$ and $\widetilde{U}_m^{pre}$ as is shown by the following argument. The denominator of $\widehat{O}$, which is the preliminary estimate of the yearly volume of machinepages, is defined by (6) and (10). The estimates $\widehat{U}_m^{pre}$ and $\widetilde{U}_m^{pre}$ are preliminary estimates of the yearly volume of originalpages and are defined by (5), (8) and (7), (9) respectively. The same persons are included in estimates of both yearly volumes of machinepages and originalpages. Hence, since the number of machinepages and originalpages are dependent for each person in the sample period also the corresponding yearly volumes will be dependent.

In order to discuss bias properties of the combined estimate $\dot{U}_m^{com}(q)$, a Taylor approximation of the calibration factor will be needed. A first-order Taylor approximation of the estimator $\widehat{O}$ (neglecting the remainder terms) yields

$$\begin{aligned}
\widehat{O} &= V\frac{1}{\widehat{V}^{pre}} \\
&\approx V\left\{\frac{1}{V'} - \frac{1}{(V')^2}(\widehat{V}^{pre} - V')\right\} \\
&= O - \frac{O}{V'}(\widehat{V}^{pre} - V'). \qquad (39)
\end{aligned}$$

From (37) it is now possible to investigate the expectation of the estimator $\dot{U}_m^{com}(q)$

$$\begin{aligned}
E[\dot{U}_m^{com}(q)] &= E[\widehat{O}(q\widehat{U}_m^{pre} + (1-q)\widetilde{U}_m^{pre})] \\
&= \sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k} W_{lp}\, N_{lk}\, c_{lp}\, E[\bar{u}_{lkmp}(q)\widehat{O}] \\
&= \sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k} W_{lp}\, N_{lk}\, c_{lp}\, E[\widehat{O}E[\bar{u}_{lkmp}(q)|\{v_{lkps}\}]] \\
&= \sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k} W_{lp}\, N_{lk}\, c_{lp}\frac{1}{n_{lkp}}\sum_{s=1}^{n_{lkp}} E[\widehat{O}\frac{a_{lkmp}U_m v_{lkps}}{V}] \\
&= \frac{U_m}{V}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k} W_{lp}\, N_{lk}\, c_{lp}\, a_{lkmp}\frac{1}{n_{lkp}}\sum_{s=1}^{n_{lkp}} E[v_{lkps}E[\widehat{O}|v_{lkps}]]. \qquad (40)
\end{aligned}$$

In order to find an approximation to $E[\dot{U}_m^{com}(q)]$, $E[\widehat{O}|v_{lkps}]$ will have to be evaluated using (39), (10), (6), (22), (31), (11), (26) and (12)

$$\begin{aligned} E[\widehat{O}|v_{lkps}] &\approx O - \frac{O}{V'}(E[\sum_{p'=1}^{3}\sum_{l'=1}^{3}\sum_{k'}\widehat{W}_{l'p'}\,N_{l'k'}\,c_{l'p'}\,\frac{1}{n_{l'k'p'}}\sum_{s'=1}^{n_{l'k'p'}}v_{l'k'p's'}|v_{lkps}] - V') \\ &= O - \frac{O}{V'}(V' - \frac{V_{lkp}}{On_{lkp}} + \frac{1}{n_{lkp}}W_{lp}N_{lk}c_{lp}v_{lkps} - V') \\ &= O(1 + \frac{1}{V}\frac{V_{lkp}}{n_{lkp}} - \frac{1}{V'n_{lkp}}W_{lp}\,N_{lk}\,c_{lp}v_{lkps}). \end{aligned} \tag{41}$$

Inserting (41) into (40) gives

$$\begin{aligned} &E[\dot{U}_m^{com}(q)] \\ &\approx U_m\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}W_{lp}\,N_{lk}\,c_{lp}\,a_{lkmp}\,\frac{1}{n_{lkp}}\sum_{s=1}^{n_{lkp}}E[\frac{v_{lkps}}{V}O\{1 \\ &\quad + \frac{1}{V}\frac{V_{lkp}}{n_{lkp}} - \frac{1}{V'}\,\frac{1}{n_{lkp}}W_{lp}\,N_{lk}\,c_{lp}v_{lkps}\}]. \end{aligned} \tag{42}$$

In the following we will need the expectation of $\frac{1}{n_{lkp}}\sum_{s=1}^{n_{lkp}}v_{lkps}^2$. Let the indicator function I_{lkps} be 1 or 0 depending on whether s is in the sample or not. We then have

$$\begin{aligned} E\left[\frac{1}{n_{lkp}}\sum_{s=1}^{n_{lkp}}v_{lkps}^2\right] &= E\left(\frac{1}{n_{lkp}}\sum_{s=1}^{N_{lk}}v_{lkps}^2 I_{lkps}\right) \\ &= \frac{1}{n_{lkp}}\sum_{s=1}^{N_{lk}}v_{lkps}^2\frac{n_{lkp}}{N_{lk}} \\ &= \frac{1}{N_{lk}}\sum_{s=1}^{N_{lk}}v_{lkps}^2. \end{aligned} \tag{43}$$

Let $V_{lkps} = OW_{lp}N_{lk}c_{lp}v_{lkps}$. Furthermore, using (42), (31), (26), (28), (43), (25), (23), (24) and (12) the following first order approximation is arrived at

$$\begin{aligned} &E[\dot{U}_m^{com}(q)] \\ &\approx U_m + \frac{U_m}{V^2}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}\frac{a_{lkmp}V_{lkp}^2}{n_{lkp}} - \frac{U_m}{V'VO}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}\frac{a_{lkmp}}{n_{lkp}}\frac{1}{N_{lk}}\sum_{s=1}^{N_{lk}}V_{lkps}^2 \\ &= U_m + \frac{1}{V}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}\frac{U_{lkmp}V_{lkp}}{n_{lkp}} - \frac{1}{V}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}\frac{U_{lkmp}}{n_{lkp}V_{lkp}}\frac{1}{N_{lk}}\sum_{s=1}^{N_{lk}}V_{lkps}^2. \end{aligned}$$

A summary bias of all types of material is given by

$$E[\dot{U}^{com}(q)] - U \approx \frac{1}{V}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}\frac{U_{lkp}V_{lkp}}{n_{lkp}} - \frac{1}{V}\sum_{p=1}^{3}\sum_{l=1}^{3}\sum_{k}\frac{U_{lkp}}{n_{lkp}V_{lkp}}\frac{1}{N_{lk}}\sum_{s=1}^{N_{lk}}V_{lkps}^2. \tag{44}$$

The estimator $E[\dot{U}^{com}(q)]$ has an approximate bias which is dependent on the sample size. A bias estimator is constructed by inserting the sample versions of the population parameters in (44). For instance, the sample version of $\frac{1}{N_{lk}}\sum_{s=1}^{N_{lk}} V_{lkps}^2$ is $\frac{1}{n_{lkp}}\sum_{s=1}^{n_{lkp}} \widehat{V}_{lkps}^2$, where $\widehat{V}_{lkps} = O\widehat{W}_{lp}N_{lk}c_{lp}v_{lkps}$. The approximate bias will become smaller with increasing samples. It is found by replacing n_{lkp} by N_{lk} in (44).

C The assumption $a_{lkmp} = a_{lkp}$

This appendix is trying to answer some questions regarding the assumption $a_{lkmp} = a_{lkp}$ in (29), i.e. independence of the type of material m. (29) implies that

$$U_{lkp} \stackrel{def}{=} \sum_m U_{lkmp} = \sum_m a_{lkmp}U_m\frac{V_{lkp}}{V}. \tag{45}$$

Introduce d_{lkp} by

$$\frac{U_{lkp}}{U} \stackrel{def}{=} d_{lkp}\frac{V_{lkp}}{V}. \tag{46}$$

With an obvious definition of V_{lkmp}, see (26), we now define an alternative model by

$$\frac{U_{lkmp}}{U_{lkp}} = e_{lkmp}\frac{V_{lkmp}}{V_{lkp}}. \tag{47}$$

Let $U_{lkp}/V_{lkp} = g_{lkp}$ and $U_{lkmp}/V_{lkmp} = g_{lkmp}$. Assuming $e_{lkmp} = 1$ is obviously equivalent to $g_{lkmp} = g_{lkp}$. However, the latter assumption means that the ratio of originalpages to machinepages is independent of type of material. There is, however, little reason to assume that this proportion is influenced by the type of material, which would mean that several originalpages of some type of material is more difficult to copy to one machinepage than others. An exception is newspapers of large format. We hence assume in the rest of this appendix that $e_{lkmp} = 1$. Equation (46) gives

$$\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k d_{lkp}V_{lkp} = V. \tag{48}$$

Using $e_{lkmp} = 1$, (47) and (46) gives

$$\begin{aligned}\frac{U_{lkmp}}{U_m} &= \frac{U_{lkp}V_{lkmp}/V_{lkp}}{\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k U_{lkp}V_{lkmp}/V_{lkp}} \\ &= \frac{V_{lkmp}d_{lkp}V/V_{lkp}}{\sum_{p=1}^{3}\sum_{l=1}^{G}\sum_k V_{lkmp}d_{lkp}}\frac{V_{lkp}}{V}.\end{aligned} \tag{49}$$

This will be the same as letting (see (29))

$$a_{lkmp} = \frac{V_{lkmp} d_{lkp} V / V_{lkp}}{\sum_{p=1}^{3} \sum_{l=1}^{G} \sum_{k} V_{lkmp} d_{lkp}}. \tag{50}$$

Immediately we see that m enters into a_{lkmp} through V_{lkmp}. If $V_{lkmp}/V_{lkp} = f_m$, i.e. independent of personnel type l, section k and sample period p, (48) gives

$$a_{lkmp} = d_{lkp}. \tag{51}$$

The assumption of $V_{lkmp}/V_{lkp} = f_m$ is, however, unreasonable. The conclusion must be that a_{lkmp} is dependent of type of material.

References

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics.

Zhang, L. C., & Solheim, L. & Roll-Hansen, D. (1999). *Photocopying in higher education*. Department of Social Statistics, Statistics Norway, P.O.B. 8131 Dep., N-0033 Oslo.