PPS Sampling with Panel Rotation for Estimating Price Indices on Services

Sander Scholtus, Arnout van Delden[1]

Abstract

In the last decade, Statistics Netherlands (SN) has started publishing service producer price indices (SPPI) for a range of service types. The quarterly published price indices were based on a stratified simple random sampling design. We present a redesign which is based on a probability proportional to size (PPS) sampling design, using turnover as a size variable. This redesign involved 52 services across 28 economic sectors.

In this redesign, we accounted for three 'non-standard' issues. Firstly, SN does not have a sampling frame of services. Instead, we need to estimate the service indices from a PPS sample of enterprises stratified by NACE code. An enterprise may have activities in multiple services. We propose a ratio estimator for the SPPI based on a PPS sample of enterprises. Secondly, we addressed the allocation of the sample over different NACE codes, taking into account that each sector can have a different number of underlying services and a different relative importance. A Neyman allocation was used with a cost component for the number of underlying services. Finally, we wanted to design a rotating PPS panel, also accounting for births and deaths in the population, while still obtaining approximately unbiased estimators. In a simulation study, we compared different rotation strategies on the accuracy of their inclusion probabilities, and on the bias and variance of the estimators. We concluded that a Pareto sampling method gave the best results.

Key Words: PPS sampling, service indices, panel rotation

1. Introduction

Statistics Netherlands (SN) has been publishing quarterly producer price indices on services (SPPI) since 2002. To this end, a classification of services is linked to the NACE classification of enterprises by main economic activity. Separate indices are published for different types of services, as well as a total SPPI which is a weighted average of the separate SPPIs. To collect information about quarterly price mutations on services, SN conducts a panel survey for a sample of enterprises from the general business register (GBR). Initially, samples were drawn using a stratified simple random sampling design, with stratification by economic sector and size class (based on number of employees).

Two developments have prompted SN to reconsider the sampling and estimation procedure for the SPPI. Firstly, the annual turnover is now available for all enterprises in the Dutch GBR; this variable is derived from administrative data on tax declarations, combined with a monthly or quarterly census survey of the largest and most complex units (van Delden and de Wolf, 2013). This makes it possible to use probability proportional to size (PPS) sampling with turnover as a size variable. Secondly, the existing SPPI panels have suffered severely from attrition, due to a lack of structural

[1] Sander Scholtus and Arnout van Delden, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, emails: s.scholtus@cbs.nl, a.vandelden@cbs.nl.

panel maintenance. In fact, due to attrition, the existing panels are now treated *de facto* as a stratified simple random sample by sector; i.e., the additional stratification by size class is ignored during estimation. To prevent this attrition from occurring again in the future, SN wants to introduce annual panel rotation in the data collection process for the SPPI.

The switch to a PPS sampling design with panel rotation raises several methodological issues. First of all, SN has to work with a sampling frame of enterprises rather than services and it is not known at the population level in which types of services each enterprise is involved. Therefore, we need to develop an estimation strategy for price indices on service domains based on a stratified PPS sample of enterprises, where multiple service domains may belong to the same PPS stratum. Secondly, we would like to allocate the total sample size across the PPS strata in a way that optimises the accuracy of the total SPPI, while also ensuring that each service domain is covered sufficiently by the sample. Finally, no “perfect” fixed-size panel rotation method exists that exactly achieves the nominal inclusion probabilities of a general PPS sampling design for all units in all survey rounds (Grafström and Matei, 2015). Therefore, some approximate rotation procedure has to be used. We want to choose a rotation method for which the bias in the estimated price indices is negligible.

In the remainder of this paper, we explain these issues in more detail and describe the solutions we propose for the Dutch SPPI. The PPS estimator of the SPPI and its variance are given in Section 2. Sample allocation theory and its application to the Dutch SPPI are presented in Section 3. The panel rotation problem is discussed in Section 4. A conclusion follows in Section 5.

2. SPPI Estimation using a PPS Panel from a Fixed Population

2.1 Estimation and Inference at the Sector Level

As mentioned in the introduction, SPPI estimates are based on a panel survey of enterprises. The information in the GBR allows us to stratify the sample by economic sector, based on NACE codes. In some cases, the SPPI publication level coincides with a NACE sector. However, in many cases SN also publishes indices for service domains at a more detailed level. In this paper, we use the terms *sector-level SPPI* and *domain-level SPPI* to distinguish between these two situations. Throughout the paper domains are defined in such a way that they belong to a single sector (i.e., a single PPS stratum).

As a starting point, we take a sector-level SPPI that is not differentiated further into domains. Until Section 4, we consider the populations of enterprises and services to be fixed over time. For an enterprise b that belongs to NACE sector h, let $P_{hb}(t,q;0)$ denote its individual price index of services within that sector in quarter q of year t with respect to some base period 0, and let $X_{hb}(ref)$ denote its (annual) turnover in a reference year. Let N_h denote the number of enterprises in the population of sector h. A direct quarterly price index of services in sector h can be defined as:

$$I_h(t,q;0) = \frac{\sum_{b=1}^{N_h} X_{hb}(ref) P_{hb}(t,q;0)}{\sum_{b=1}^{N_h} X_{hb}(ref)} = \sum_{b=1}^{N_h} W_{hb}(ref) P_{hb}(t,q;0), \qquad (1)$$

where $W_{hb}(ref) = X_{hb}(ref)/X_h(ref)$ and $X_h(ref) = \sum_{b=1}^{N_h} X_{hb}(ref)$ denotes the total turnover in sector h in the reference year. The choice $ref = 0$ in Formula (1) yields a standard Laspeyres price index. In practice, SN uses weights from an earlier reference period ($ref < 0$) for the SPPI, which makes (1) a so-called Young index (van der Grient and de Haan, 2011; IMF, 2004). For the remainder of Sections 2 and 3, we simplify the notation by suppressing the time indices, e.g. $P_{hb} \equiv P_{hb}(t, q; 0)$ and $W_{hb} \equiv W_{hb}(ref)$.

We first review some results on how to estimate I_h from a PPS sample. Suppose a sample of n_h enterprises is taken from sector h, where the inclusion probability of enterprise b is proportional to its turnover in the reference period: $\pi_{hb} = n_h W_{hb}$. We assume here that $X_{hb} < X_h/n_h$ (or $W_{hb} < 1/n_h$) for all $b = 1, \dots, N_h$. Enterprises with $X_{hb} \geq X_h/n_h$ should be placed in a separate stratum and selected with certainty. In practice, this concerns a limited number of enterprises. We defer a discussion of this until Section 2.3.

For this PPS sample, the Horvitz-Thompson (HT) estimator of I_h has a very simple form:

$$\hat{I}_h = \sum_{i=1}^{n_h} \frac{w_{hi} p_{hi}}{\pi_{hi}} = \frac{1}{n_h} \sum_{i=1}^{n_h} p_{hi} = \bar{p}_h, \tag{2}$$

where the sampled units are indicated by $i = 1, \dots, n_h$ and lowercase characters denote sample observations. That is to say, under PPS sampling at the sector level, the *unweighted* sample mean of individual price indices is an unbiased estimator for the *weighted* price index in Formula (1); see also, e.g., Knottnerus (2011a, 2011b).

At SN, PPS samples are usually drawn by systematic sampling from a randomly ordered list (Banning et al., 2012, pp. 53–54). This corresponds to Procedure 2 of Brewer and Hanif (1983), who also listed 49 other procedures for selecting a PPS sample. Under the assumption that all $W_{hb} \ll 1$, an approximate variance formula for $\hat{I}_h$ under this sampling procedure is (Hartley and Rao, 1962; Knottnerus, 2011b):

$$\operatorname{var}(\hat{I}_h) \approx \frac{1}{n_h} \sum_{b=1}^{N_h} W_{hb}\{1 - (n_h - 1) W_{hb}\}(P_{hb} - I_h)^2. \tag{3}$$

To simplify this expression further, it is helpful to suppose that the deviations $\varepsilon_{hb} = P_{hb} - I_h$ are independently distributed with $E(\varepsilon_{hb}) = 0$ and $E(\varepsilon_{hb}^2) \propto W_{hb}^{\beta}$ for some value of β. In previous studies at SN with (S)PPI data, it was found that this model often holds with $\beta \approx 0$, which implies that the variance of the deviations does not depend on turnover. Knottnerus (2011b) derived the following approximation for $\beta = 0$:

$$\begin{aligned} \operatorname{var}(\hat{I}_h) &\approx \sigma_{ph}^2 \left\{ \frac{1}{n_h} - \frac{1}{N_h}\left(1 + CV_{wh}^2\right) \right\}, \\ \sigma_{ph}^2 &= \sum_{b=1}^{N_h} W_{hb}(P_{hb} - I_h)^2. \end{aligned} \tag{4}$$

In these expressions, it is also assumed that the sample and population sizes are large enough so that $n_h - 1$ and $N_h - 1$ may be replaced by n_h and N_h. Furthermore, σ_{ph}^2 is a measure of the variability of the individual price indices and $CV_{wh}^2 = CV_{xh}^2$ denotes the squared coefficient of variation of turnover in sector h. To apply Formula (4) in practice, we only need to estimate σ_{ph}^2, since N_h and CV_{wh}^2 can be obtained directly from the sampling frame. Knottnerus (2011b) discussed how to estimate σ_{ph}^2 from a PPS sample.

2.2 Estimation and Inference at the Domain Level

Next, we consider a sector h with $D_h \geq 2$ underlying service domains for which separate SPPIs have to be estimated. An enterprise in sector h can be active within zero, one, or more than one of these service domains. Let X_{hdb} denote the turnover of enterprise b from services within domain d, with $X_{hdb} = 0$ for enterprises that do not provide these services. For an enterprise that is active within domain d, let P_{hdb} denote its associated price index. Analogously to (1), the direct price index for domain d is defined as

$$I_{hd} = \frac{\sum_{b=1}^{N_h} X_{hdb} P_{hdb}}{\sum_{b=1}^{N_h} X_{hdb}}, \quad d = 1, \ldots, D_h, \tag{5}$$

where the product $X_{hdb}P_{hdb}$ is taken to be zero when $X_{hdb} = 0$.

As noted in the introduction, the information in the sampling frame does not allow SN to determine beforehand which enterprises are active within each domain. Therefore, we again select a PPS sample of size n_h at the sector level, with inclusion probabilities based on the sector turnover X_{hb} as before. The sampled units are then contacted by telephone to obtain their turnover specification by service domain. The panel for domain d effectively consists of all units in the total PPS sample with $x_{hdi} > 0$. Note that this implies that the effective sample size in domain d is stochastic.

To find an estimator for I_{hd}, it is convenient to rewrite Formula (5) as follows:

$$I_{hd} = \frac{\frac{1}{X_h}\sum_{b=1}^{N_h} X_{hdb} P_{hdb}}{\frac{1}{X_h}\sum_{b=1}^{N_h} X_{hdb}} = \frac{\sum_{b=1}^{N_h} \frac{X_{hb}}{X_h}\frac{X_{hdb}}{X_{hb}} P_{hdb}}{\sum_{b=1}^{N_h} \frac{X_{hb}}{X_h}\frac{X_{hdb}}{X_{hb}}} = \frac{\sum_{b=1}^{N_h} W_{hb} Y_{hdb}}{\sum_{b=1}^{N_h} W_{hb} G_{hdb}}, \tag{6}$$

where $Y_{hdb} = G_{hdb}P_{hdb}$ and $G_{hdb} = X_{hdb}/X_{hb}$ is the fraction of turnover that enterprise b derives from domain d. The numerator and denominator of (6) are both expressions of the form (1). Hence, they can both be estimated from the PPS sample using a HT estimator of the form (2). In this manner, we obtain a ratio estimator for I_{hd}:

$$\hat{I}_{hd} = \frac{\bar{y}_{hd}}{\bar{g}_{hd}} = \frac{\sum_{i=1}^{n_h} y_{hdi}}{\sum_{i=1}^{n_h} g_{hdi}} = \frac{\sum_{i=1}^{n_h} g_{hdi} p_{hdi}}{\sum_{i=1}^{n_h} g_{hdi}} = \sum_{i=1}^{n_h} \frac{1}{n_h}\frac{g_{hdi}}{\bar{g}_{hd}} p_{hdi}. \tag{7}$$

Recall that, to estimate a price index at the sector level, each sampled unit receives the same weight ($1/n_h$). To estimate a domain-level SPPI, as the last expression in (7) shows, these basic weights are adjusted by a factor $g_{hdi}/\bar{g}_{hd}$ that accounts for the fraction of turnover that a unit derives from the domain. In particular, units that are not active within the domain ($g_{hdi} = 0$) receive a zero weight.

By the standard properties of a ratio estimator, $\hat{I}_{hd}$ in (7) is asymptotically unbiased for I_{hd}, the bias being of the order $O(1/n_h)$ (Cochran, 1977, p. 160). It should be noted that the denominator of (6) equals $G_{hd} = \sum_{b=1}^{N_h} X_{hdb}/X_h$, the overall share of domain d in the total turnover of sector h. A reasonable proxy for G_{hd} is usually available from the Structural Business Statistics or the National Accounts. Thus, it is not strictly necessary to estimate the denominator of (6) from the PPS panel, and we could estimate I_{hd} instead by $\hat{I}_{hd,alt} = \bar{y}_{hd}/G_{hd}$, which is unbiased. However, it is well known that, in practice, the ratio estimator $\hat{I}_{hd}$ typically leads to a better result in terms of mean squared error.

To obtain a variance formula for $\hat{I}_{hd}$, we can apply a standard Taylor linearisation argument. Using the fact that $E(\bar{g}_{hd}) = G_{hd}$, this yields:

$$\mathrm{var}(\hat{I}_{hd}) \approx \mathrm{var}\left(\frac{\bar{y}_{hd} - I_{hd}\bar{g}_{hd}}{G_{hd}}\right) = \frac{1}{G_{hd}^2}\mathrm{var}(\bar{e}_{hd}),$$

with $\bar{e}_{hd}$ the sample mean of residuals $E_{hdb} = Y_{hdb} - I_{hd}G_{hdb} = G_{hdb}(P_{hdb} - I_{hd})$. For the variance of $\bar{e}_{hd}$, approximate expressions similar to (3) and (4) could be derived.

2.3 The Take-All Stratum and the Total SPPI

As noted above, the population may contain units with large turnovers $X_{hb} \geq X_h/n_h$. These have to be placed in a separate 'take-all' stratum and selected with probability $\pi_{hb} = 1$. We denote these units by $b = N_h + 1, N_h + 2, \dots, N_h^*$ in the population and by $i = n_h + 1, n_h + 2, \dots, n_h^*$ in the sample, with $n_h^* = n_h + N_h^* - N_h$. The sample size n_h^* indirectly determines the boundary of the take-all stratum, and hence N_h, n_h and X_h.

The definition of the price indices (1) and (5) has to be extended to incorporate the take-all stratum. For the sector-level index, this yields:

$$I_h^* = \frac{\sum_{b=1}^{N_h^*} X_{hb}P_{hb}}{\sum_{b=1}^{N_h^*} X_{hb}} = \frac{X_h}{X_h^*}I_h + \frac{1}{X_h^*}\sum_{b=N_h+1}^{N_h^*} X_{hb}P_{hb},$$

with $X_h^* = \sum_{b=1}^{N_h^*} X_{hb} = X_h + \sum_{b=N_h+1}^{N_h^*} X_{hb}$ and I_h given by (1). Since the take-all stratum is completely observed and I_h can be estimated by $\hat{I}_h$ from (2), we can estimate I_h^* by

$$\hat{I}_h^* = \frac{X_h}{X_h^*}\hat{I}_h + \frac{1}{X_h^*}\sum_{i=n_h+1}^{n_h^*} x_{hi}p_{hi}. \tag{8}$$

Furthermore, the second term in (8) does not contribute to the variance of $\hat{I}_h^*$, so that $\mathrm{var}(\hat{I}_h^*) = (X_h/X_h^*)^2\,\mathrm{var}(\hat{I}_h)$, with $\mathrm{var}(\hat{I}_h)$ given by Expression (4). For the domain-level index I_{hd} and its estimator $\hat{I}_{hd}$, similar extensions I_{hd}^* and $\hat{I}_{hd}^*$ can be developed.

Finally, we mentioned in the introduction that SN also publishes a total SPPI, aggregated across economic sectors $h = 1, \dots, H$. This total SPPI is defined and estimated as

$$I_{tot}^* = \sum_{h=1}^{H} W_h^* I_h^*, \qquad \hat{I}_{tot}^* = \sum_{h=1}^{H} W_h^* \hat{I}_h^*, \tag{9}$$

with W_h^* the share of sector h in the total turnover of all sectors $h = 1, \dots, H$. In practice, these weights are based on macro-integrated data from the National Accounts. Regarding the variance of $\hat{I}_{tot}^*$, we observe that

$$\mathrm{var}(\hat{I}_{tot}^*) = \sum_{h=1}^{H}(W_h^*)^2\,\mathrm{var}(\hat{I}_h^*) \approx \sum_{h=1}^{H}\left(W_h^*\frac{X_h}{X_h^*}\right)^2 \sigma_{ph}^2\left\{\frac{1}{n_h} - \frac{1}{N_h}\left(1 + CV_{wh}^2\right)\right\}, \tag{10}$$

since an independent PPS sample is drawn from each sector h.

3. Sample Allocation

3.1 Allocation Formulas

Next, we will consider the choice of sample size n_h^* in each sector. The total sample size that is available for all SPPIs is fixed, say, at n^* units. We want to allocate this total sample size across the sectors $h = 1, \dots, H$, taking into account that these sectors differ in terms of economic importance and heterogeneity of individual price indices.

A natural starting point is to try to find the allocation $(n_1^*, \dots, n_H^*)$ with $\sum_{h=1}^{H} n_h^* = n^*$ that minimises the variance of the total SPPI, $\mathrm{var}(\hat{I}_{tot}^*)$, given by Formula (10). This is actually a difficult, non-standard optimisation problem, because several quantities that occur in the target function (n_h, N_h, X_h, σ_{ph}^2 and CV_{wh}^2) may depend indirectly on n_h^* through the delineation of the take-all stratum. In principle, we could attempt to solve this problem numerically. However, given that the variance of $\hat{I}_{tot}^*$ is not the only concern here (see below) *and* that, in practice, σ_{ph}^2 will be estimated from a small sample, it seems reasonable to simplify the problem.

To obtain a simpler problem, we note that, from the definition of the take-all stratum:

$$X_h^* = \sum_{b=1}^{N_h^*} X_{hb} \geq X_h + \frac{n_h^* - n_h}{n_h} X_h = \frac{n_h^*}{n_h} X_h.$$

Hence, $X_h / X_h^* \leq n_h / n_h^* \leq \sqrt{n_h / n_h^*}$; the last inequality holds because $0 \leq n_h \leq n_h^*$. Substituting this inequality into Formula (10), we find that (approximately)

$$\mathrm{var}(\hat{I}_{tot}^*) \leq \sum_{h=1}^{H} (W_h^*)^2 \frac{n_h}{n_h^*} \sigma_{ph}^2 \left\{ \frac{1}{n_h} - \frac{1}{N_h} \left(1 + CV_{wh}^2\right) \right\} \leq \sum_{h=1}^{H} (W_h^*)^2 \frac{\sigma_{ph}^2}{n_h^*}. \qquad (11)$$

Finally, we assume that σ_{ph}^2 does not depend on the delineation of the take-all stratum; this is in line with the assumption $\beta = 0$ that was used in the derivation of (4).

Instead of minimising $\mathrm{var}(\hat{I}_{tot}^*)$ directly, we propose to minimise upper bound (11) for all $(n_1^*, \dots, n_H^*)$ with $\sum_{h=1}^{H} n_h^* = n^*$. This problem has the same structure as the well-known Neyman allocation problem (see, e.g., Cochran, 1977, pp. 98–99). The optimal solution is therefore given by:

$$n_h^* = \frac{W_h^* \sigma_{ph}}{\sum_{g=1}^{H} W_g^* \sigma_{pg}} n^*, \quad h = 1, \dots, H. \qquad (12)$$

Thus, more sample units are allocated to sectors with larger contributions to the service economy (in terms of W_h^*) and/or more heterogeneous price indices (in terms of σ_{ph}).

Allocation (12) does not take into account that SN also wants to publish reliable domain-level SPPIs. That is to say, it may be advantageous to allocate more sample units to some sectors that consist of relatively many domains, to ensure that these domains are covered sufficiently by the sample. As SN has little information about the population at the domain level, we do not attempt to find an 'optimal' allocation strategy that directly takes the variances of the domain-level SPPIs into account. Instead, we propose an allocation strategy that incorporates information about the domains through a cost function.

Suppose that sector h consists of D_h domains and that the average number of domains in which an enterprise from this sector is active equals $\bar{A}_h$. As mentioned above, each of the n_h^* sampled units from sector h is asked to report price indices for all domains in which it is active. Thus, the costs of data collection per unit can be considered higher when units tend to be active across multiple domains (i.e., when $\bar{A}_h/D_h$ is relatively large) and lower when units tend to work in isolated domains (i.e., when $\bar{A}_h/D_h$ is small).

Let $C_h = \bar{A}_h/D_h \leq 1$. By the above reasoning, a suitable expression for the relative costs associated with a particular sample allocation may be $C_{tot} = \sum_{h=1}^{H} C_h n_h^*$. The allocation that minimises upper bound (11) for a given value of C_{tot} can be derived analogously to the Neyman allocation (Cochran, 1977, pp. 96–98):

$$n_h^* = \frac{W_h^* \sigma_{ph}/\sqrt{C_h}}{\sum_{g=1}^{H} W_g^* \sigma_{pg}/\sqrt{C_g}} n^* = \frac{W_h^* \sigma_{ph}\sqrt{D_h/\bar{A}_h}}{\sum_{g=1}^{H} W_g^* \sigma_{pg}\sqrt{D_g/\bar{A}_g}} n^*, \quad h = 1, \dots, H. \tag{13}$$

Here, we have implicitly chosen C_{tot} in such a way that the total sample size still equals n^*. Allocation (13) has the desirable properties that – all other things being equal – more units will be sampled from a sector if it contains more domains (higher D_h) and/or if enterprises tend to be active in isolated domains (lower $\bar{A}_h$). This means that, in comparison to (12), we can expect a sample with better coverage at the domain level.

To apply Formula (13), we have to estimate σ_{ph} and $\bar{A}_h$ from the existing SPPI panels. For the estimation of σ_{ph}, we cannot use the formulas in Knottnerus (2011b) because the existing panel is not a PPS sample. Recall from Section 1 that the existing panel in sector h is treated *de facto* as a simple random sample of size, say, m_h. An associated estimator for I_h is $\hat{I}_{h,SRS} = \left(\sum_{i=1}^{m_h} x_{hi} p_{hi}\right)/\left(\sum_{i=1}^{m_h} x_{hi}\right)$; cf. Knottnerus (2011a). It can be shown that an asymptotically unbiased estimator for σ_{ph}^2 from a simple random sample is given by

$$\hat{\sigma}_{ph,SRS}^2 = \frac{\sum_{i=1}^{m_h} x_{hi}\left(p_{hi} - \hat{I}_{h,SRS}\right)^2}{\sum_{i=1}^{m_h} x_{hi}} + \widehat{\mathrm{var}}\left(\hat{I}_{h,SRS}\right),$$

where $\widehat{\mathrm{var}}\left(\hat{I}_{h,SRS}\right)$ denotes an asymptotically unbiased variance estimator for $\hat{I}_{h,SRS}$. In the application to be discussed in Section 3.2, we used $\hat{\sigma}_{ph,SRS}^2$ to estimate σ_{ph}^2.

3.2 Results

We applied the theory from the previous subsection to obtain an allocation for the 28 economic sectors that are currently sampled for the Dutch SPPI. Each sector contains between one and six service domains. The total number of domains is 52. We used the GBR of the first quarter of 2013 as a population frame. The total population across all sectors contained about 235000 units. The data for 2013 of the existing SPPI panels were used to estimate σ_{ph} and $\bar{A}_h$. To obtain a robust estimate for σ_{ph}, we computed $\hat{\sigma}_{ph,SRS}^2$ for each quarter of 2013 separately and then took the median value.

Two different methods are currently used to collect price information on services from the sampled units (OECD/Eurostat, 2014): based on actual transaction prices (ATP) or based on model prices for standard products (MP). Of the 28 sectors in the Dutch SPPI, 8 are observed using ATP and 20 are observed using MP. From our data, we found that larger values of $\hat{\sigma}^2_{ph,SRS}$ – i.e., more heterogeneous price indices – tend to occur for sectors that are observed using ATP. This was in line with the expectations of the SPPI production staff.

The total sample size n^* was 1500. We considered the following allocation scenarios:

A. Use the allocation of the existing panels, inflated to a total sample size of 1500.
B. Use the allocation given by Formula (13).
C. As B, with the additional restriction that $n_h^* \geq 15/C_h$ in each sector.
D. As C, with the additional restriction that the standard error of $\hat{I}_h^*$ is at most equal to some fixed upper bound U_h.
E. As D, with W_h^* replaced by $\sqrt{W_h^*}$ in Formula (13).
F. As E, with the sample size for the sector "Advertising agencies" fixed a priori.

Scenario A was included to compare our results with the current allocation. Scenario B is the basic allocation strategy that was proposed in Section 3.1. Under this scenario, the accuracy of the sector-level and domain-level SPPIs is not explicitly taken into account. To address this, we introduced two additional constraints. The restriction added in Scenario C ensures that the expected effective sample size in each domain equals at least 15. (This number was chosen by the SPPI production staff.) Under Scenario D, in addition an upper bound was placed on the standard error of the sector-level SPPI. In consultation with the production staff, we fixed this upper bound at 0.5 index points for sectors observed using MP and 1.0 index points for sectors observed using ATP. Scenario E maintained these restrictions but used the square root of the sector weights in Formula (13) to reduce the dependence of the allocation on economic importance. Finally, Scenario F was introduced to treat the sector "Advertising agencies" separately. This sector had by far the largest value of $\hat{\sigma}^2_{ph,SRS}$ and was therefore assigned about one-fifth of the total sample under Scenarios B–E. According to the production staff, the large value of $\hat{\sigma}^2_{ph,SRS}$ is caused by low data quality in this sector and should not be used as a basis for sample allocation. In consultation with the production staff, we fixed the sample size for this sector at $n_h^* = 80$ under Scenario F. The additional restrictions for Scenarios C–F were implemented heuristically and iteratively, by fixing any n_h^* that failed a restriction to the nearest feasible value and re-calculating (13) for the remaining sectors.

Table 1: The effect of various allocation scenarios on the accuracy of the estimated SPPI.

Scenario	*S.E.*($\hat{I}^*_{tot}$)	*S.E.*($\hat{I}^*_h$) *in index points*		
		mean (all sectors)	*mean (ATP sectors)*	*mean (MP sectors)*
A	0.238	0.538	0.939	0.377
B	0.135	0.857	0.658	0.936
C	0.137	0.511	0.709	0.432
D	0.151	0.442	0.741	0.323
E	0.154	0.420	0.685	0.313
F	0.183	0.434	0.772	0.299

By comparing the results for Scenario A and the other scenarios (Table 1), it is seen that a substantial improvement in the accuracy of the total SPPI could be achieved by re-allocating the current sample. The largest improvement in accuracy of $\hat{I}^*_{tot}$ occurred under Scenario B. However, with this scenario the gained accuracy at the total level would be offset by a large drop in the accuracy of sector-level SPPIs for sectors observed using MP. As expected, the refinements introduced by Scenarios C, D and E caused a decrease in the accuracy of the total-level SPPI – albeit a slight one – but also a substantial improvement at the sector level. On balance, we concluded that Scenario E gave the best results here. As explained above, Scenario F was introduced to deal with a particular problem with our data. The latter allocation scenario was eventually chosen.

4. Panel Rotation

4.1 SPPI Estimation for Dynamic Populations

In practice, the SPPI is calculated as a chain index rather than a direct index, to take population dynamics into account. At the population level, the chain-index formulation of the SPPI of sector h for quarter q in year t is defined recursively by:

$$\begin{aligned} I_h^*(t,q;0) &= I_h^*(t-1,u;0) \times I_h^*(t,q;t-1,u), \\ I_h^*(t,q;t-1,u) &= \frac{\sum_b X_{hb}(t-1,u;ref)P_{hb}(t,q;t-1,u)}{\sum_b X_{hb}(t-1,u;ref)}, \\ X_{hb}(t-1,u;ref) &= X_{hb}(ref) \times P_{hb}(t-1,u;ref). \end{aligned} \tag{14}$$

Here, $I_h^*(t-1,u;0)$ denotes the SPPI for the last quarter of year $t-1$ with respect to the base period, and $I_h^*(t,q;t-1,u)$ denotes the price index for the current quarter with respect to the last quarter of the previous year. Similarly, $P_{hb}(t-1,u;ref)$ and $P_{hb}(t,q;t-1,u)$ denote corresponding price indices for enterprise b. The yearly adaptation of turnover weights that occurs in the last line of (14) is known as *price updating*. See, e.g., van der Grient and de Haan (2011) for more details.

It can be shown that, if the populations of enterprises and services remain fixed between the base period and quarter q of year t, then the chain index (14) and the direct index (1) (or rather, its extension from Section 2.3 that incorporates the take-all stratum) yield identical values. In practice, these populations do change over time: new enterprises are born, others cease to exist, and some enterprises may change their activities (services) and therefore move to different domains or even sectors.

At SN, these changes in the population are handled at the yearly transitions, by letting each short-term index $I_h^*(t,q;t-1,u)$ refer only to the population of enterprises that are active within sector h in both periods. These short-term indices are then chained together using Expression (14). An additional complication is that the variable $X_{hb}(ref)$ is not defined for enterprises that were born after the original reference period. Moreover – and particularly relevant for PPS sampling –, enterprises can grow or shrink over time, which means that weights based on $X_{hb}(ref)$ may not reflect the actual importance of units in the population at later time points, particularly if long chains are used. We therefore replace $X_{hb}(t-1,u;ref)$ in Expression (14) by $X_{hb}(t-1,u;ref_t) = X_{hb}(ref_t) \times P_{hb}(t-1,u;ref_t)$, where ref_t denotes the most recent year for which turnover values are available in the GBR at the end of year $t-1$. In practice, $ref_t = t-2$.

To estimate $I_h^*(t,q;t-1,u)$, a panel of enterprises selected by PPS sampling can still be used, but now the panel needs to be updated at each yearly transition. Firstly, the panel should be made representative for the current population by removing units that are no longer active (in sector h) and by selecting new-born units and units that have moved to sector h. Secondly, the inclusion probabilities for all units should be based on the most recent available turnover information. That is to say, for the sample of size $n_h^*(t)$ that is taken in year t, it should hold that

$$\pi_{hb}(t) = \begin{cases} n_h(t)X_{hb}(ref_t)/X_h(ref_t) & \text{if } X_{hb}(ref_t) < X_h(ref_t)/n(t), \\ 1 & \text{if } X_{hb}(ref_t) \geq X_h(ref_t)/n(t). \end{cases} \tag{15}$$

Here, $X_h(ref_t)$ denotes the sum of $X_{hb}(ref_t)$ for all active units outside the take-all stratum, and $n_h(t)$ denotes the sample size outside the take-all stratum for year t. In addition to these panel updates that are necessary to avoid selection bias, SN also wants to apply panel rotation to reduce the burden on responding enterprises. We will discuss methods for achieving both objectives in the next subsection. The use of panel rotation in addition to panel updating does not affect the estimation procedure.

Based on a PPS sample of active enterprises in sector h with inclusion probabilities $\pi_{hb}(t)$ given by (15), the short-term sector-level index $I_h^*(t, q; t-1, u)$ can be estimated analogously to Expression (8). For the domain-level SPPIs and total SPPI, similar definitions as chain indices can be given and the associated PPS-based estimators follow analogously. It should be noted that no straightforward expressions exist for the variances of these estimated chain indices. Therefore, we used the direct indices as an approximation to derive a sample allocation in Section 3.

4.2 Methods for Panel Rotation

In the context of a PPS sample, it is obvious that panel rotation cannot be applied to the take-all stratum. For units outside the take-all stratum, panel rotation is applied to each sector independently. For simplicity, we will consider one sector and suppress the index h in the notation for the remainder of this section. Denote the panel for years $t-1$ and t without the take-all stratum by $S(t-1)$ and $S(t)$.

To update the PPS panel at the yearly transition between $t-1$ and t, we begin by determining the new take-all stratum for year t. We then remove any units from $S(t-1)$ that are no longer active (within this sector). Next, a PPS sample is drawn from the subpopulation of new-born units and added to the current panel. The inclusion probability of new-born unit b for this step is given by $n(t-1)X_b(ref_t)/X(ref_{t-1})$; i.e., we mimic $\pi_b(t-1)$ from Expression (15) but use $X_b(ref_t)$, as $X_b(ref_{t-1})$ is undefined for new-born units. This step ensures that new-born units and continuing units have the same point of departure for panel rotation, which is important to avoid selection bias in the long run. Denote the updated panel without the take-all stratum by $\check{S}(t-1)$. Next, the updated panel is rotated. The amount of panel rotation is controlled by the so-called rotation fraction. We define a rotation fraction of λ $(0 \leq \lambda \leq 1)$ to mean that $\lambda \times 100\%$ of the units in $\check{S}(t-1)$ are not part of $S(t)$.

Various methods for rotating PPS samples have been proposed. We will compare four of them here. The first three are variations of so-called Permanent Random Number (PRN) methods. PRNs are often used for the general problem of sample coordination (Ohlsson, 1995; Lindblom, 2014), of which panel rotation is a special case. The PRN principle entails that each unit in the population is assigned an independent random number R_b from the uniform distribution on $[0,1)$ which does not change over time. New units are assigned a PRN at birth. Samples are drawn by selecting units based on their PRNs.

1. Poisson sampling with PRNs

Probably the simplest PRN method that is suitable for PPS sampling is obtained by considering the PPS sample as a Poisson sample. In this case, the selection mechanism is independent across units. To select an initial Poisson sample in year 0 from a population of units with PRNs $R_1, \ldots, R_N$ and inclusion probabilities $\pi_1(0), \ldots, \pi_N(0)$, one chooses a starting point $a(0) \in [0,1)$ and selects all units with $a(0) < R_b \leq a(0) + \pi_b(0)$ (Ohlsson, 1995). It is easy to see that this indeed yields a sample with the desired first-order inclusion probabilities. When $a(0) + \pi_b(0) > 1$, the selection interval is truncated at 1 and carried on from 0; see Figure 1.

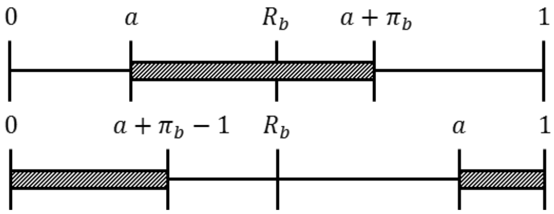

Figure 1: An illustration of Poisson sampling using PRNs. Unit b is selected in the sample if, and only if, its PRN R_b lies in the shaded part of the interval $[0,1]$.

To apply panel rotation to a Poisson sample that is drawn this way, one simply moves the starting point $a(t-1)$ to a new position $a(t) \geq a(t-1)$. This causes some units to be removed from the sample, because their PRNs do not belong to their new selection intervals, while other units may enter the sample. For a given population and sample, it is straightforward to work out the minimal adjustment to $a(t-1)$ that is required to obtain a desired rotation fraction λ. This approach also works for method 2 and 3 below.

An advantage of Poisson sampling for panel rotation is that it produces PPS samples that exactly respect the nominal first-order inclusion probabilities of all units in the population at all times. Of the four methods that will be considered in this paper, only Poisson sampling has this property. On the other hand, because this method selects units independently, the sample size is random. Thus, the panel size is not known beforehand and may vary considerably from year to year, which is an important practical drawback. [In addition, the randomness of the sample size might increase the variance, but this can be alleviated by using a ratio estimator such as (7).] Alternative, fixed-size sampling methods are therefore of interest. Here, we consider two simple PRN-based methods. Recently, Grafström and Matei (2015) have suggested an alternative coordination method based on conditional Poisson sampling (i.e., conditional on the sample size); we did not include this more complicated approach in the present study.

2. Sequential Poisson sampling with PRNs

We rewrite the PRN of unit b for year t as

$$r_b(t) = \left(R_b - a(t)\right) \bmod 1 = \begin{cases} R_b - a(t) & \text{if } a(t) \leq R_b \\ R_b - a(t) + 1 & \text{if } R_b < a(t) \end{cases} \tag{16}$$

and define $\rho_b(t) = r_b(t)/\pi_b(t)$. For Poisson sampling, it can be shown that unit b is selected in the PPS panel for year t precisely when $\rho_b(t) \leq 1$. Ohlsson (1995, 1998) proposed to obtain an approximate PPS sample of fixed size $n(t)$ by sorting the population in ascending order of $\rho_b(t)$ and selecting the first $n(t)$ units. This method is known as sequential Poisson sampling. The sample size is now fixed, but the actual inclusion probability of unit b may not be exactly equal to $\pi_b(t)$. Some bias may therefore be incurred in the HT estimator that uses the nominal inclusion probabilities.

3. Pareto sampling with PRNs

Pareto sampling, due to Rosén (1997), adjusts the above PRN transformation $\rho_b(t)$ to:

$$\tilde{\rho}_b(t) = \frac{r_b(t)/[1 - r_b(t)]}{\pi_b(t)/[1 - \pi_b(t)]}.$$

Rosén (2000) argued that this transformation should improve the approximation to the nominal inclusion probabilities in practice, in comparison to sequential Poisson sampling. In a simulation study, Ohlsson (2000) verified this for small sample sizes ($n \leq 4$). Aires and Rosén (2005) showed by simulation that, for populations of sizes $N \leq 200$, the Pareto sampling method works well under a variety of conditions. Nonetheless, the Dutch SPPI involves populations that are much larger and the turnover distribution in these populations is more skewed than was considered by Aires and Rosén (2005).

Aires (1999) derived a recursive method for computing the realised inclusion probabilities under Pareto sampling. Using this method, it is – in theory – possible to make adjustments to obtain the exact nominal inclusion probabilities with Pareto sampling. However, this approach is too computationally demanding for our application. (The median population size of the SPPI sectors is about 3300; the median sample size is about 40.)

4. Circular systematic PPS sampling

The final method that we consider is not a PRN method, but a relatively simple extension of the systematic PPS sampling method that is already used at SN for fixed panels and cross-sectional PPS samples. This method was proposed by Knottnerus and Enthoven (2012). The idea is to draw systematic PPS samples by cycling through a randomly ordered list; when the bottom of the list is reached, we start again at the top. Panel rotation is then applied in two steps. Denote the size of the current panel (after panel updating but before rotation) by $\breve{n}(t)$. First, a simple random subsample of $\max\{\lambda\breve{n}(t), \breve{n}(t) - n(t)\}$ units is removed from the panel. Then, an additional PPS sample is drawn from the units that are currently not in the panel, so that the total new panel size becomes $n(t)$. The first step uses the fact that a simple random subsample of a PPS sample is again a PPS sample. Like methods 2 and 3, this method yields fixed-size samples with inclusion probabilities that may differ from the nominal ones.

With any sampling method where the panel is being updated and/or rotated, there might be some units that alternate from year to year between being (just) inside the take-all stratum and outside it with an inclusion probability close to 1.0. These units might then be rotated in and out of the sample quite often, which is undesirable in practice. To alleviate this problem, Knottnerus and Enthoven (2012) suggested to define a slightly larger take-all stratum by including all units with $\max\{\pi_b(t), \pi_b(t-1), \pi_b(t-2)\} \geq 1$ and $\pi_b(t) \geq \gamma$ for some value $0 < \gamma < 1$. We followed this suggestion with $\gamma = 0.8$.

4.3 Simulation Study

To compare the usefulness of the above panel rotation methods for the Dutch SPPI, we conducted a simulation study. To this end, we created a synthetic population with price indices for ten years, based on real SPPI panel data and real sampling frames for the sector "Road transport of goods". This sector consists of five domains. Population dynamics and turnover distributions were modelled on the sampling frames for 2013 and 2014. The quarterly price mutations for the synthetic population were modelled on the real SPPI panel data of 2013 and 2014. We imputed price mutations in such a way that the sector-level and domain-level SPPIs for the synthetic population approximated the published quarterly indices for "Road transport of goods" in the period 2005–2014. In what follows, we therefore describe the results as if the study refers to this ten-year period. The population size varied from 9,815 units in "2005" to 10,908 units in "2014".

Panel surveys were simulated using each of the four rotation methods from Section 4.2, by drawing a PPS sample from the synthetic population of "2005" and then performing yearly panel updates and rotation until "2014". Different PRNs were assigned in each simulation round, but within each round the same PRNs were used for the three PRN-based methods. For each simulated panel, the sector and domain SPPIs were also estimated using the theory of Section 4.1.

The panel rotation methods were evaluated based on the difference between the actual and nominal inclusion probabilities and the accuracy of the estimated indices. Using R simulation rounds, the actual inclusion probability of unit b in year t can be estimated by $\hat{\pi}_{bR}(t) = R^{-1} \sum_{r=1}^{R} \iota\{b \in S_r^*(t)\}$, with $S_r^*(t)$ the panel for year t in round r and $\iota\{.\}$ an indicator function. Under the hypothesis that a rotation method produces PPS samples with the nominal inclusion probabilities $\pi_b(t)$, $\iota\{b \in S_r^*(t)\}$ is distributed as a Bernoulli variable with mean $\pi_b(t)$ and variance $\pi_b(t)[1 - \pi_b(t)]$. Hence, assuming that $R\pi_b(t) \gg 1$ and $R[1 - \pi_b(t)] \gg 1$, the residual

$$T_{bR}(t) = \sqrt{R}\,\frac{\hat{\pi}_{bR}(t) - \pi_b(t)}{\sqrt{\pi_b(t)[1 - \pi_b(t)]}} \tag{17}$$

should follow a distribution that is approximately standard normal. Any large deviations from the standard normal distribution would indicate that some units are selected much more or less often than they should be according to their nominal inclusion probabilities. To evaluate the effect of such deviations on estimates, we also computed the empirical bias and root mean squared error (RMSE) of the estimated sector and domain SPPIs.

The simulations were repeated with different parameter settings. Here, we present the results for $R = 20{,}000$ samples of size $n^*(t) = 71$ in all years and a rotation fraction $\lambda = 0.1$; this corresponds to the actual situation in production for the sector "Road transport of goods". The results for other settings led to similar conclusions.

Figure 2 displays the histograms of $T_{bR}(2006)$ (the second year of the simulation). Given the assumptions stated above Formula (17), we only included units with $0.01 \leq \pi_b(t) \leq 0.99$ here. For comparison, the standard normal density is plotted as a red dashed line. As expected, the Poisson sampling method achieved the nominal inclusion probabilities; for this method, the distribution of $T_{bR}(t)$ was similar for all years. The other methods all produced reasonable PPS samples for 2005 (not shown here; the distributions were similar to the upper-left panel in Figure 2), but from 2006 onwards some differences occurred. The Pareto sampling method approximated the nominal inclusion probabilities very well. Sequential Poisson sampling performed reasonably well, but not as well as Pareto sampling. With the circular systematic PPS method, some very large residuals occurred. The results for the other years (2007–2014) were similar.

Table 2: Bias and RMSE (in index points) of the estimated sector SPPI (distribution over 40 quarters), with $n^*(t) = 71$ and $\lambda = 0.1$, based on $R = 20{,}000$ simulations.

		PRN-Poisson	*PRN-Sequential*	*PRN-Pareto*	*circular PPS*
bias	*min.*	–0.01	–0.01	–0.01	–1.00
	median	0.00	0.00	–0.00	–0.00
	max.	0.07	0.03	0.06	0.17
RMSE	*min.*	0.20	0.20	0.20	0.22
	median	0.46	0.46	0.46	0.49
	max.	0.69	0.69	0.69	1.26

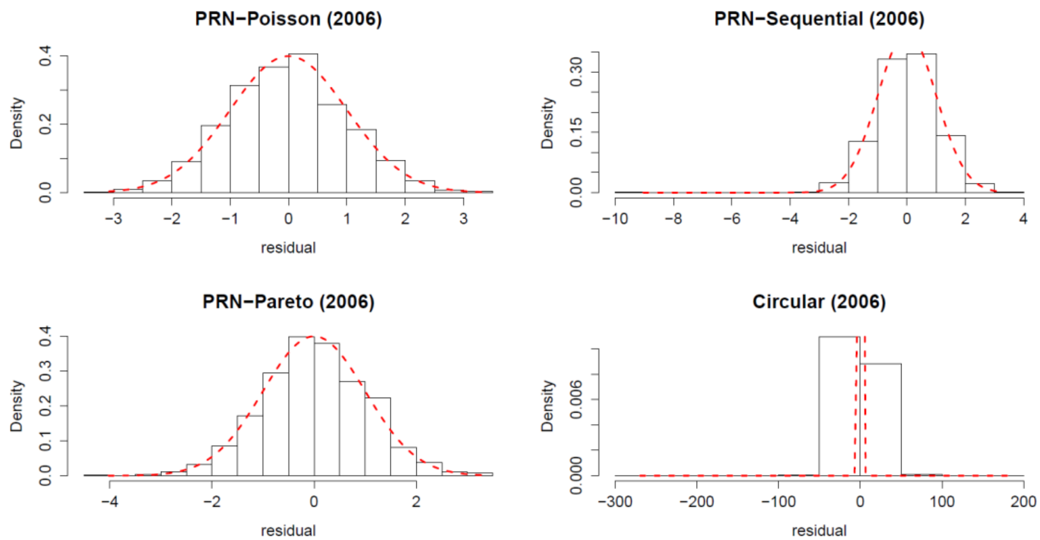

Figure 2: Histograms of the residuals $T_{bR}(t)$ for $t = 2006$ (second year) for each rotation method, with $n^*(t) = 71$ and $\lambda = 0.1$, based on $R = 20{,}000$ simulations.

The empirical bias and RMSE of the estimated indices were computed for each quarter in 2005–2014 and summarised across the 40 quarters in this period (Table 2). It is seen that the accuracy of the estimated indices was similar for all three PRN methods. In particular, the random sample size of the Poisson method did not increase the variance. The circular systematic PPS method yielded estimates that were slightly less accurate. This was also the only method for which a significant bias occasionally occurred. For the domain-level SPPIs, similar results were found (not shown here).

An explanation for the relatively poor performance of the circular systematic PPS method may lie in the use of simple random sampling to remove units from the panel. As noted above, the resulting subsample of the panel for year $t - 1$ is again a PPS sample, but importantly, it is a PPS sample with inclusion probabilities proportional to turnover in year ref_{t-1}, not ref_t. Therefore, this method will yield incorrect inclusion probabilities for units with large growths or declines in turnover. For the three PRN methods, the selection of the panel for year t is based only on turnover values in year ref_t. These methods are therefore more robust to growing and shrinking units.

Based on the results of this simulation study, we propose to use Pareto sampling to obtain panel rotation for the Dutch SPPI. This method achieves the same accuracy and approximately unbiased estimation as Poisson sampling, but with fixed sample sizes.

5. Conclusion

In this paper, we have described the recent redesign of the sampling and estimation strategy for the Dutch producer price indices on services. The new panels of enterprises are based on a stratified PPS sample design, with annual turnover as a size variable. A Neyman-like allocation is used to distribute the total available sample size across different economic sectors, and a ratio estimator is used to estimate price indices on service domains at a more detailed level than the stratification by sector. Finally, a panel rotation strategy is based on Pareto sampling with Permanent Random Numbers.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. We thank Jan de Haan for his useful comments.

References

Aires, N. (1999). Algorithms to Find Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs. *Methodology and Computing in Applied Probability* **1**, 457–469.

Aires, N. and B. Rosén (2005). On Inclusion Probabilities and Relative Estimator Bias for Pareto πps Sampling. *Journal of Statistical Planning and Inference* **128**, 543–567.

Banning, R., A. Camstra and P. Knottnerus (2012). Sampling Design and Estimation Methods. Report (Methods Series No. 201207), Statistics Netherlands, The Hague and Heerlen.

Brewer, K.R.W. and M. Hanif (1983). *Sampling with Unequal Probabilities*. Springer-Verlag, New York.

Cochran, W.G. (1977). *Sampling Techniques* (3rd Ed.). John Wiley & Sons, New York.

Grafström, A. and A. Matei (2015). Coordination of Conditional Poisson Samples. *Journal of Official Statistics* **31**, 649–672.

Hartley, H.O. and J.N.K. Rao (1962). Sampling with Unequal Probabilities and Without Replacement. *Annals of Mathematical Statistics* **33**, 350–374.

IMF (2004), *Producer Price Index Manual, Theory and Practice*. International Monetary Fund, Washington, DC.

Knottnerus, P. (2011a). Business Panels. Report (Methods Series No. 201119), Statistics Netherlands, The Hague.

Knottnerus, P. (2011b). On the Efficiency of Randomized Probability Proportional to Size Sampling. *Survey Methodology* **37**, 95–102.

Knottnerus, P. and L. Enthoven (2012). Panel Updating for the Business Cycle Survey. Internal memo (in Dutch), Statistics Netherlands, The Hague.

Lindblom, A. (2014). Theme: Sample Co-ordination. In: *Memobust Handbook on Methodology of Modern Business Statistics*, Eurostat, available online.

OECD/Eurostat (2014), *Methodological Guide for Developing Producer Price Indices for Services: Second Edition*. OECD Publishing.

Ohlsson, E. (1995). Coordination of Samples using Permanent Random Numbers. In: Cox, Binder, Chinnappa, Christianson, Colledge, and Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, pp. 153–169.

Ohlsson, E. (1998). Sequential Poisson Sampling. *Journal of Official Statistics* **14**, 149–162.

Ohlsson, E. (2000). Coordination of PPS Samples over Time. In: *The 2nd International Conference on Establishment Surveys*, pp. 255–264.

Rosén, B. (1997). On Sampling with Probability Proportional to Size. *Journal of Statistical Planning and Inference* **62**, 159–191.

Rosén, B. (2000). On Inclusion Probabilities for Order πps Sampling. *Journal of Statistical Planning and Inference* **90**, 117–143.

van der Grient, H. and J. de Haan (2011). Index Numbers. Report (Methods Series No. 201118), Statistics Netherlands, The Hague.

van Delden, A. and P.-P. de Wolf (2013). A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data. Paper presented at the conference New Techniques and Technologies for Statistics 2013, Brussels.