SESSION 2: FEBRUARY 10, 2022

Session 2.4 How to restrict to the necessary?

Denise Silva

Challenge

How to get the right balance between:

The required information to lead the reader through your ideas and work

vs.

The ability to summarise technical details and results

Every journal addresses a specific audience: an author must write the paper with that audience in mind

The Audience

- **Journal of Official Statistics:** “The intended readers are researchers and practitioners in academia, government, business or research organisations with an interest in survey methodology and production of official statistics”
- **Survey Methodology:** “A key source of information for survey statisticians and methodologists”
- **Statistical Journal of the IAOS** : “The journal should publish papers of wide interest to both users and producers of official statistics”
- **International Statistical Review:** “Papers must be of interest to a sufficiently broad spectrum of the members of ISI and its family of Associations, who include researchers and practitioners in academia, government, business, and industry”
- **The Survey Statistician**: As the newsletter of the IASS, it includes general information about the activities of the Association (meetings, seminars, etc.) and welcome manuscripts “that are likely to be of interest to its members”.

The Hourglass

Consider the image of an hourglass to represent the organization (**and balance**) of your manuscript

Wide at top: The introduction presents the relevance of the study, then it narrows down the scope into a specific problem/question

Narrow towards the middle: The methods section should be restricted to a manageable focus (the neck of the hourglass)

Widening out again: Results – give more room to this section

Wide at the bottom: a final thin layer broadening the focus to relate your work with what is on and proposing future developments

https://www.enago.com/academy/academic-writing-in-science-overview/

Methods

The work may be completely new or may demonstrate an improvement of an existing method

Official statistics journals assume that the writer will demonstrate technical expertise

- If you have developed an entirely new method: write it out in detail
- If you are improving previous work:
  - There is no need to repeat all the details (use references to inform about the previous work)
  - Be clear about which methods should be compared with your proposal

Methods

- Clarity is the aim
- If proofs are included in the manuscript: they should be complete and presented in the easiest possible way
- You can omit details according to expected target audience knowledge
- Do not over-explain common scientific/technical procedures
- Provide some discussion/motivation to assist the reader decoding the formulas
- “An author must provide enough detail for a reader to be able to reconstruct his/her study, but not so much that the relevant points get buried”
  (Journal of Young Investigators, 2005)

Empirical Work/Application

Focus is on the data/analysis used and the specific results

It is important that the readers can follow the analysis and understand the debate

- The novelty is the conceptual or analytical approach or the new evidence to inform the debate
- You may need to present statistical theory to support your analysis
- The empirical work can be related to testing different methods or evaluating methodological procedures in a new domain or application
- Deciding how much detail is about judgment

Results

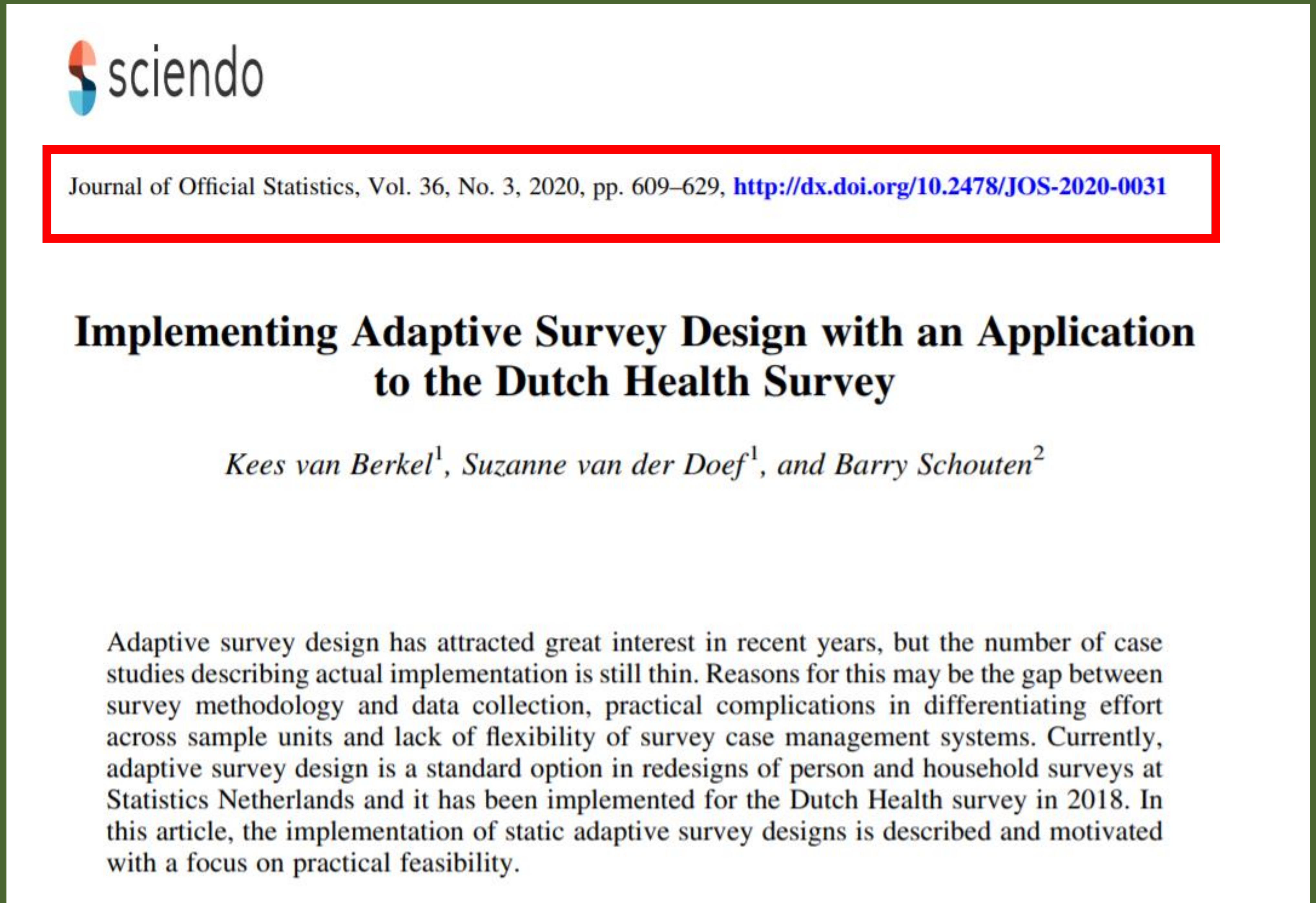
This section should get the lion’s share of your manuscript

- To show how your work fits in the context of existing literature and
- To explain “how your study adds to the body of knowledge”

- Inform the main findings
- Resist the temptation to include every result you have obtained
- Check the number of tables and figures that are allowed according to chosen journal
- Do not repeat in words everything that your tables and graphs convey (avoid redundancy)

https://spie.org/news/photonics-focus/janfeb-2020/how-to-write-a-scientific-paper?SSO=1

Methods – Example 1



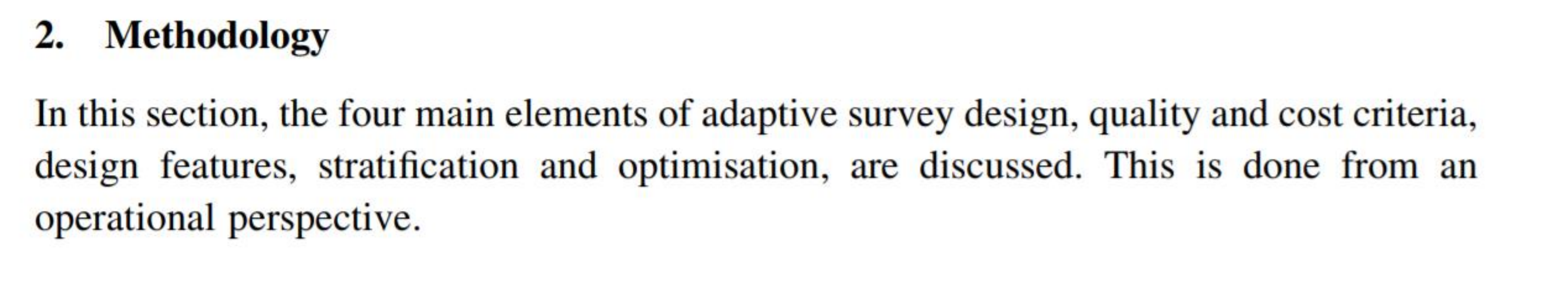
sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 609–629, http://dx.doi.org/10.2478/JOS-2020-0031

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

Kees van Berkel[1], *Suzanne van der Doef*[1], *and Barry Schouten*[2]

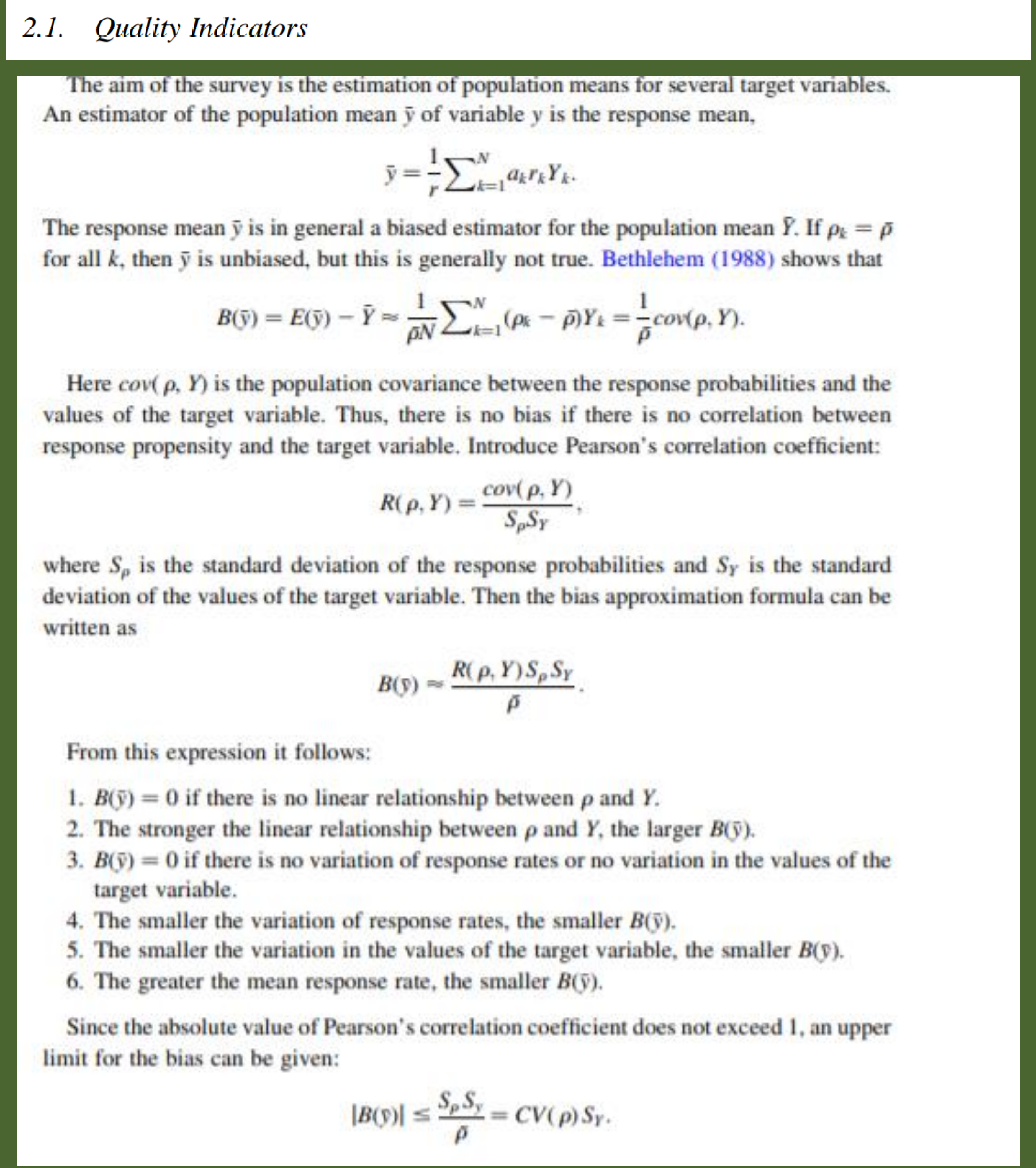
Adaptive survey design has attracted great interest in recent years, but the number of case studies describing actual implementation is still thin. Reasons for this may be the gap between survey methodology and data collection, practical complications in differentiating effort across sample units and lack of flexibility of survey case management systems. Currently, adaptive survey design is a standard option in redesigns of person and household surveys at Statistics Netherlands and it has been implemented for the Dutch Health survey in 2018. In this article, the implementation of static adaptive survey designs is described and motivated with a focus on practical feasibility.



2. Methodology

In this section, the four main elements of adaptive survey design, quality and cost criteria, design features, stratification and optimisation, are discussed. This is done from an operational perspective.

https://doi.org/10.2478/jos-2020-0031



2.1. Quality Indicators

The aim of the survey is the estimation of population means for several target variables. An estimator of the population mean $\bar{y}$ of variable y is the response mean,

$$\bar{y} = \frac{1}{r}\sum_{k=1}^{N} a_k r_k Y_k.$$

The response mean $\bar{y}$ is in general a biased estimator for the population mean $\bar{Y}$. If $\rho_k = \bar{\rho}$ for all k, then $\bar{y}$ is unbiased, but this is generally not true. Bethlehem (1988) shows that

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \frac{1}{\bar{\rho}N}\sum_{k=1}^{N}(\rho_k - \bar{\rho})Y_k = \frac{1}{\bar{\rho}}cov(\rho, Y).$$

Here $cov(\rho, Y)$ is the population covariance between the response probabilities and the values of the target variable. Thus, there is no bias if there is no correlation between response propensity and the target variable. Introduce Pearson's correlation coefficient:

$$R(\rho, Y) = \frac{cov(\rho, Y)}{S_\rho S_Y},$$

where S_ρ is the standard deviation of the response probabilities and S_Y is the standard deviation of the values of the target variable. Then the bias approximation formula can be written as

$$B(\bar{y}) \approx \frac{R(\rho, Y)S_\rho S_Y}{\bar{\rho}}.$$

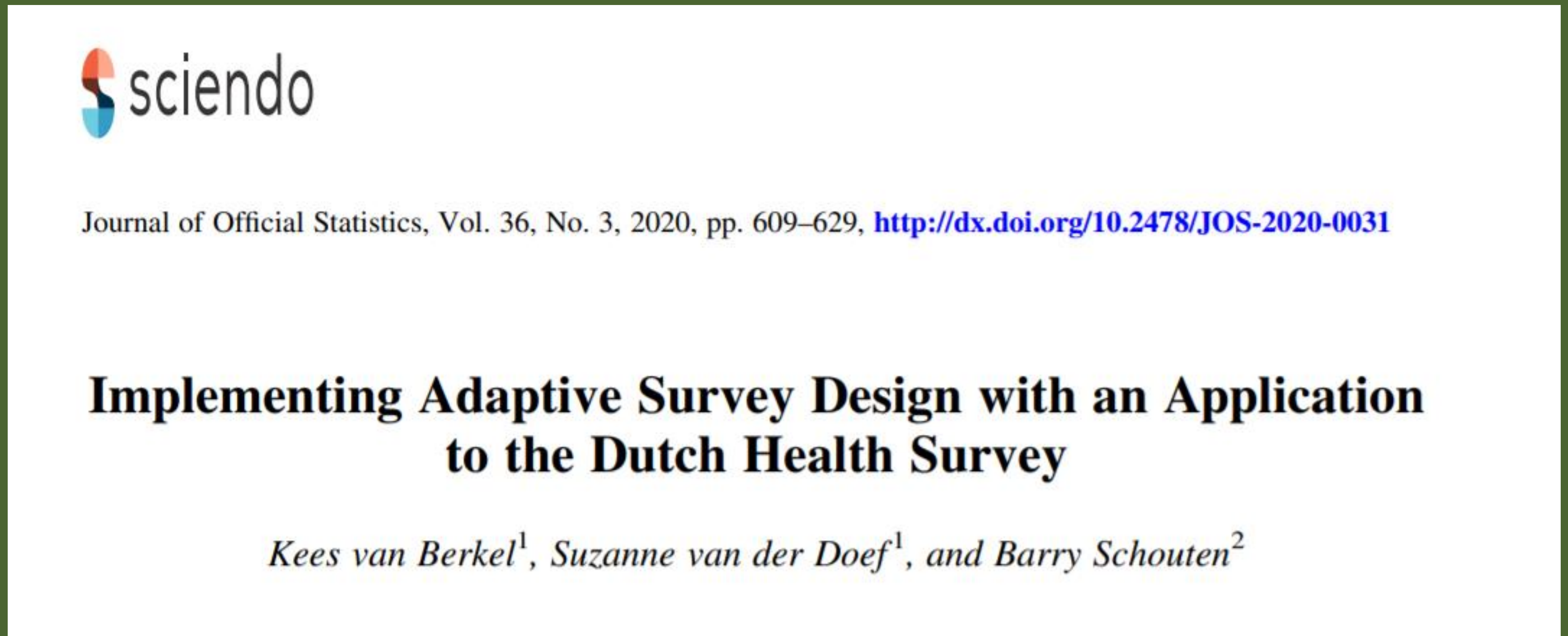
From this expression it follows:

1. $B(\bar{y}) = 0$ if there is no linear relationship between ρ and Y.
2. The stronger the linear relationship between ρ and Y, the larger $B(\bar{y})$.
3. $B(\bar{y}) = 0$ if there is no variation of response rates or no variation in the values of the target variable.
4. The smaller the variation of response rates, the smaller $B(\bar{y})$.
5. The smaller the variation in the values of the target variable, the smaller $B(\bar{y})$.
6. The greater the mean response rate, the smaller $B(\bar{y})$.

Since the absolute value of Pearson's correlation coefficient does not exceed 1, an upper limit for the bias can be given:

$$|B(\bar{y})| \leq \frac{S_\rho S_Y}{\bar{\rho}} = CV(\rho)S_Y.$$

Methods – Example 1

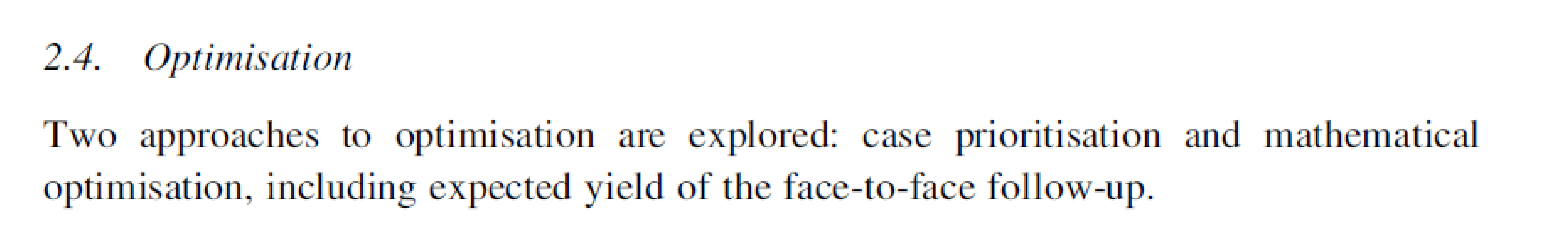


sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 609–629, http://dx.doi.org/10.2478/JOS-2020-0031

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

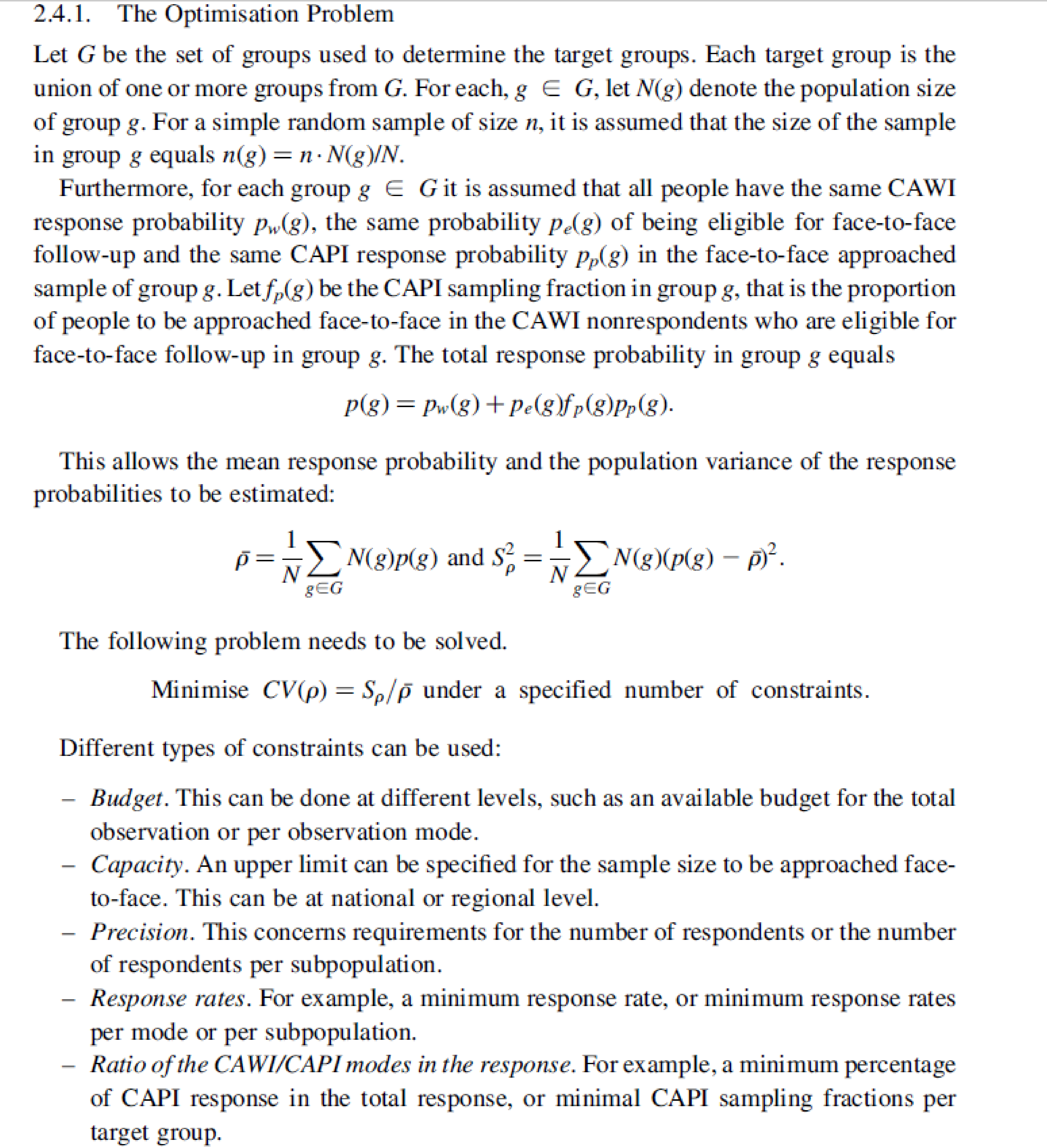
Kees van Berkel[1]*, Suzanne van der Doef*[1]*, and Barry Schouten*[2]



2.4. *Optimisation*

Two approaches to optimisation are explored: case prioritisation and mathematical optimisation, including expected yield of the face-to-face follow-up.

https://doi.org/10.2478/jos-2020-0031



2.4.1. The Optimisation Problem

Let G be the set of groups used to determine the target groups. Each target group is the union of one or more groups from G. For each, $g \in G$, let $N(g)$ denote the population size of group g. For a simple random sample of size n, it is assumed that the size of the sample in group g equals $n(g) = n \cdot N(g)/N$.

Furthermore, for each group $g \in G$ it is assumed that all people have the same CAWI response probability $p_w(g)$, the same probability $p_e(g)$ of being eligible for face-to-face follow-up and the same CAPI response probability $p_p(g)$ in the face-to-face approached sample of group g. Let $f_p(g)$ be the CAPI sampling fraction in group g, that is the proportion of people to be approached face-to-face in the CAWI nonrespondents who are eligible for face-to-face follow-up in group g. The total response probability in group g equals

$$p(g) = p_w(g) + p_e(g) f_p(g) p_p(g).$$

This allows the mean response probability and the population variance of the response probabilities to be estimated:

$$\bar{\rho} = \frac{1}{N}\sum_{g \in G} N(g)p(g) \text{ and } S_\rho^2 = \frac{1}{N}\sum_{g \in G} N(g)(p(g) - \bar{\rho})^2.$$

The following problem needs to be solved.

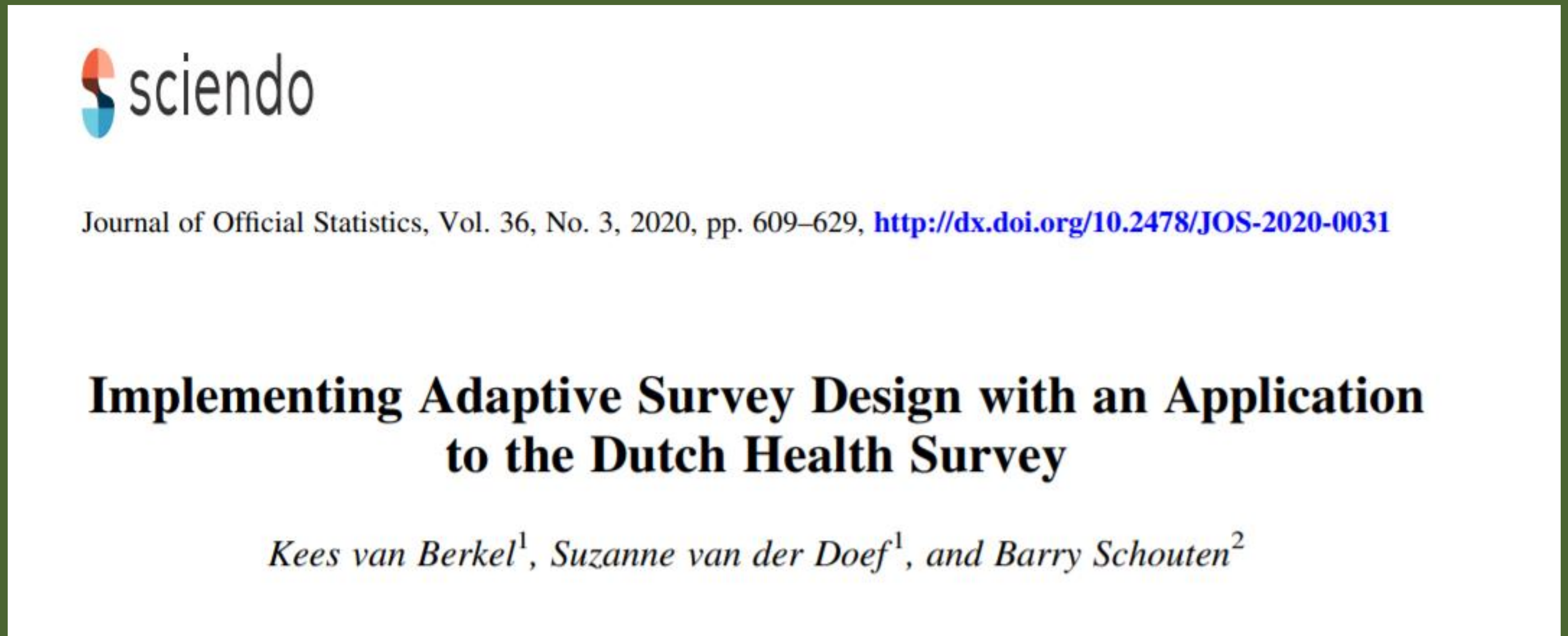
Minimise $CV(\rho) = S_\rho/\bar{\rho}$ under a specified number of constraints.

Different types of constraints can be used:

- *Budget*. This can be done at different levels, such as an available budget for the total observation or per observation mode.
- *Capacity*. An upper limit can be specified for the sample size to be approached face-to-face. This can be at national or regional level.
- *Precision*. This concerns requirements for the number of respondents or the number of respondents per subpopulation.
- *Response rates*. For example, a minimum response rate, or minimum response rates per mode or per subpopulation.
- *Ratio of the CAWI/CAPI modes in the response*. For example, a minimum percentage of CAPI response in the total response, or minimal CAPI sampling fractions per target group.

SESSION 2: FEBRUARY 10, 2022

Methods – Example 1

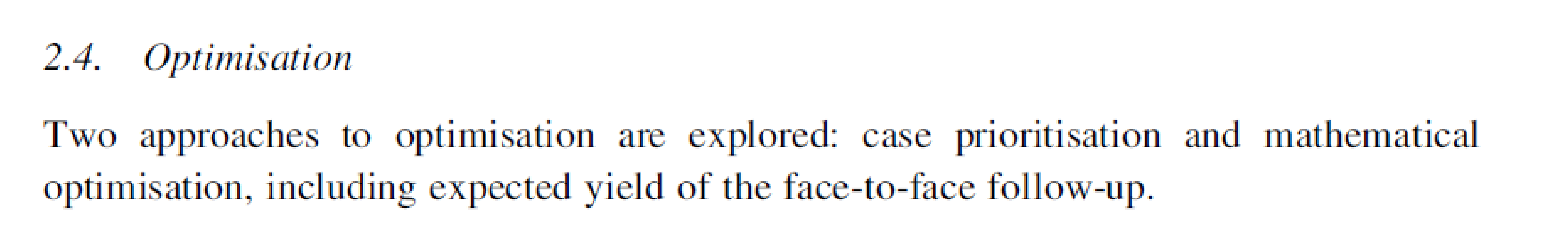


sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 609–629, http://dx.doi.org/10.2478/JOS-2020-0031

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

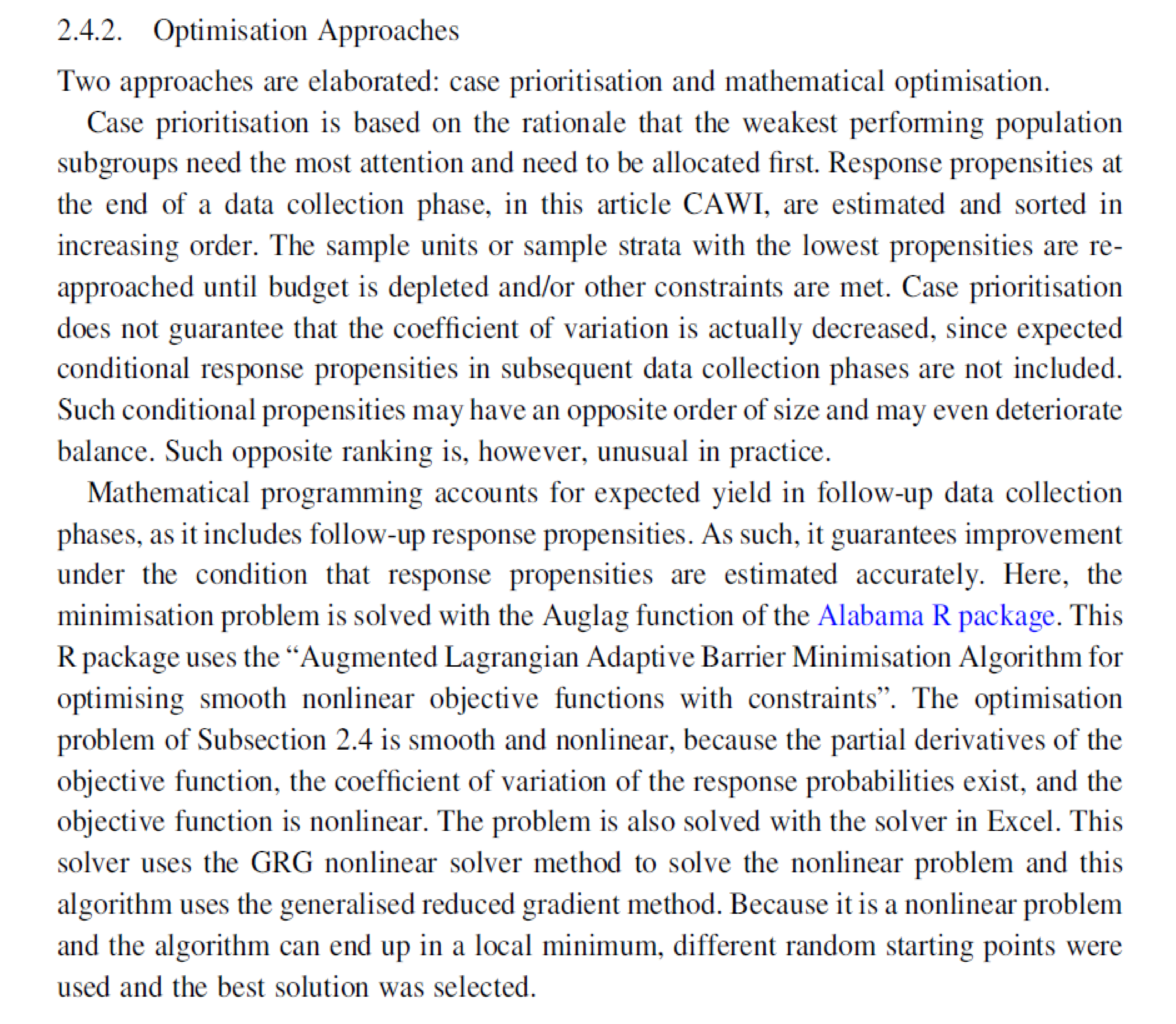
Kees van Berkel[1], Suzanne van der Doef[1], and Barry Schouten[2]



2.4. *Optimisation*

Two approaches to optimisation are explored: case prioritisation and mathematical optimisation, including expected yield of the face-to-face follow-up.

https://doi.org/10.2478/jos-2020-0031



2.4.2. Optimisation Approaches

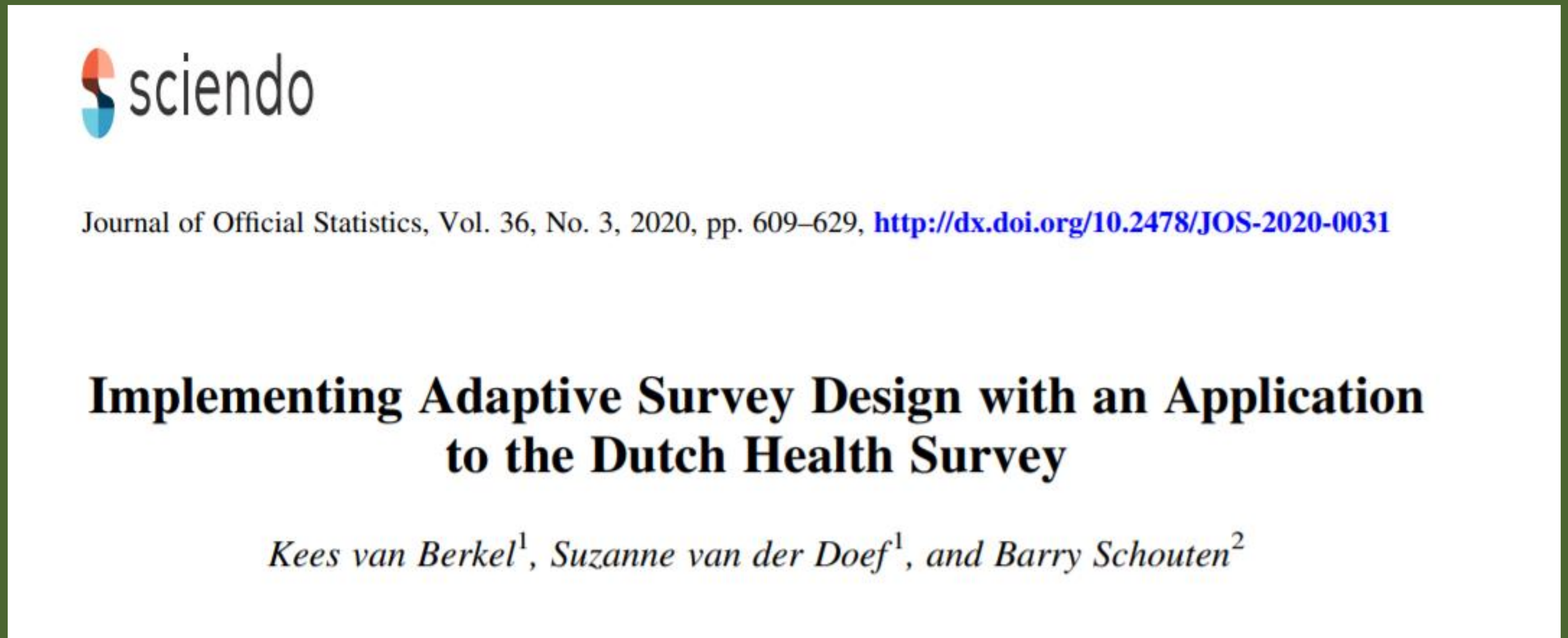
Two approaches are elaborated: case prioritisation and mathematical optimisation.

Case prioritisation is based on the rationale that the weakest performing population subgroups need the most attention and need to be allocated first. Response propensities at the end of a data collection phase, in this article CAWI, are estimated and sorted in increasing order. The sample units or sample strata with the lowest propensities are re-approached until budget is depleted and/or other constraints are met. Case prioritisation does not guarantee that the coefficient of variation is actually decreased, since expected conditional response propensities in subsequent data collection phases are not included. Such conditional propensities may have an opposite order of size and may even deteriorate balance. Such opposite ranking is, however, unusual in practice.

Mathematical programming accounts for expected yield in follow-up data collection phases, as it includes follow-up response propensities. As such, it guarantees improvement under the condition that response propensities are estimated accurately. Here, the minimisation problem is solved with the Auglag function of the Alabama R package. This R package uses the "Augmented Lagrangian Adaptive Barrier Minimisation Algorithm for optimising smooth nonlinear objective functions with constraints". The optimisation problem of Subsection 2.4 is smooth and nonlinear, because the partial derivatives of the objective function, the coefficient of variation of the response probabilities exist, and the objective function is nonlinear. The problem is also solved with the solver in Excel. This solver uses the GRG nonlinear solver method to solve the nonlinear problem and this algorithm uses the generalised reduced gradient method. Because it is a nonlinear problem and the algorithm can end up in a local minimum, different random starting points were used and the best solution was selected.

Empirical Work – Example 1

SESSION 2: FEBRUARY 10, 2022

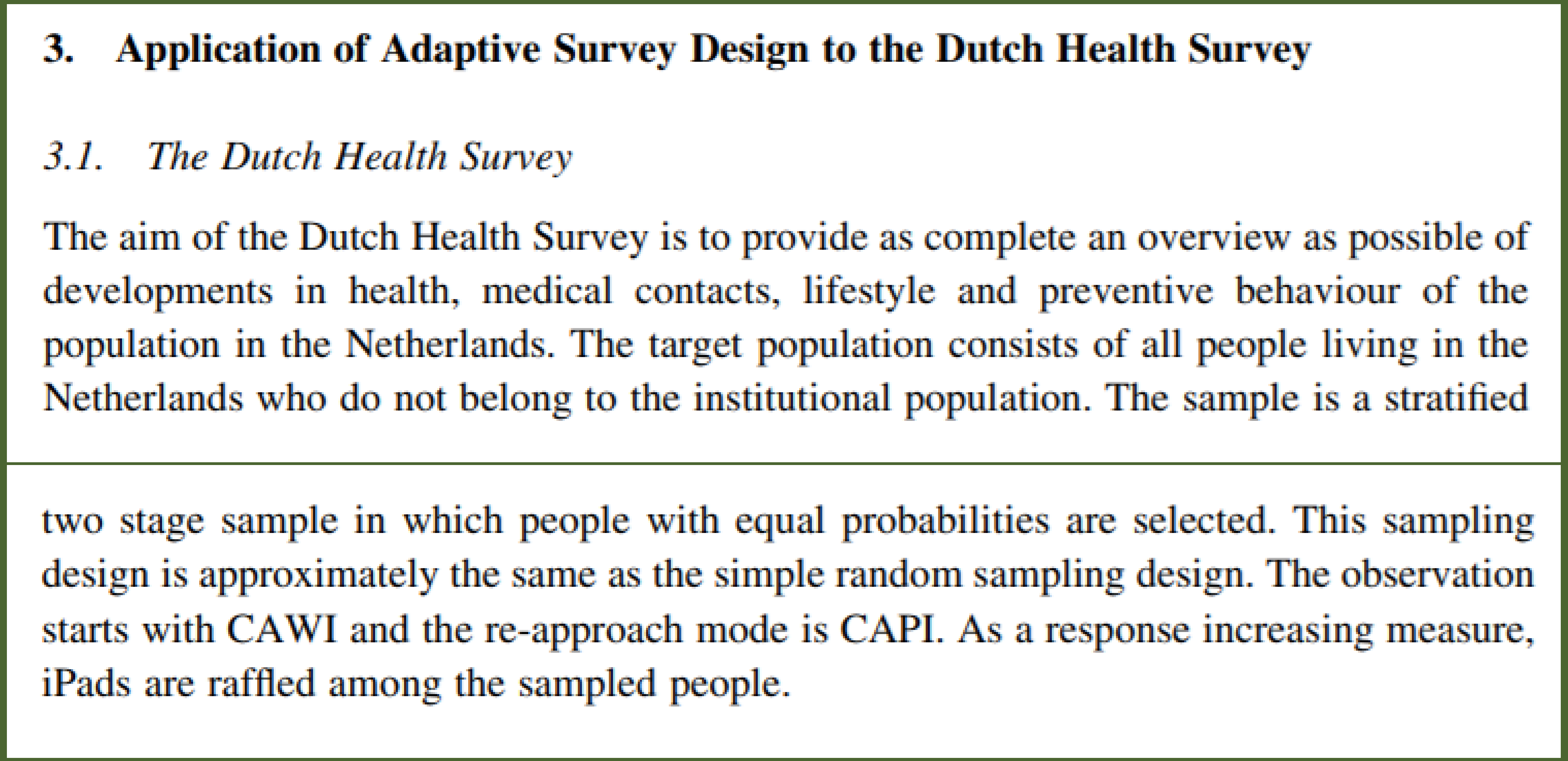


sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 609–629, http://dx.doi.org/10.2478/JOS-2020-0031

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

Kees van Berkel[1], *Suzanne van der Doef*[1], *and Barry Schouten*[2]

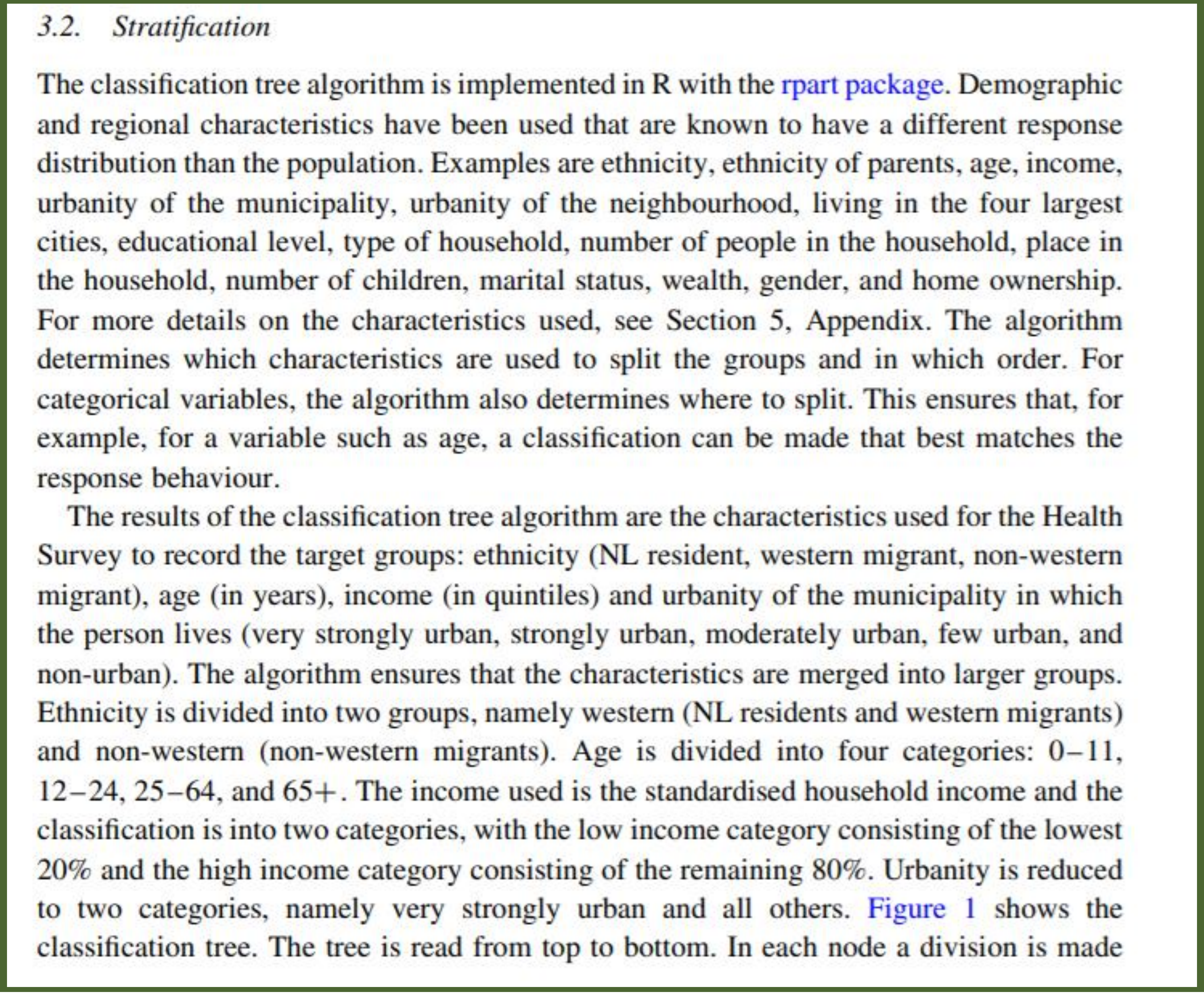


3. Application of Adaptive Survey Design to the Dutch Health Survey

3.1. The Dutch Health Survey

The aim of the Dutch Health Survey is to provide as complete an overview as possible of developments in health, medical contacts, lifestyle and preventive behaviour of the population in the Netherlands. The target population consists of all people living in the Netherlands who do not belong to the institutional population. The sample is a stratified two stage sample in which people with equal probabilities are selected. This sampling design is approximately the same as the simple random sampling design. The observation starts with CAWI and the re-approach mode is CAPI. As a response increasing measure, iPads are raffled among the sampled people.

https://doi.org/10.2478/jos-2020-0031



3.2. Stratification

The classification tree algorithm is implemented in R with the rpart package. Demographic and regional characteristics have been used that are known to have a different response distribution than the population. Examples are ethnicity, ethnicity of parents, age, income, urbanity of the municipality, urbanity of the neighbourhood, living in the four largest cities, educational level, type of household, number of people in the household, place in the household, number of children, marital status, wealth, gender, and home ownership. For more details on the characteristics used, see Section 5, Appendix. The algorithm determines which characteristics are used to split the groups and in which order. For categorical variables, the algorithm also determines where to split. This ensures that, for example, for a variable such as age, a classification can be made that best matches the response behaviour.

The results of the classification tree algorithm are the characteristics used for the Health Survey to record the target groups: ethnicity (NL resident, western migrant, non-western migrant), age (in years), income (in quintiles) and urbanity of the municipality in which the person lives (very strongly urban, strongly urban, moderately urban, few urban, and non-urban). The algorithm ensures that the characteristics are merged into larger groups. Ethnicity is divided into two groups, namely western (NL residents and western migrants) and non-western (non-western migrants). Age is divided into four categories: 0–11, 12–24, 25–64, and 65+. The income used is the standardised household income and the classification is into two categories, with the low income category consisting of the lowest 20% and the high income category consisting of the remaining 80%. Urbanity is reduced to two categories, namely very strongly urban and all others. Figure 1 shows the classification tree. The tree is read from top to bottom. In each node a division is made

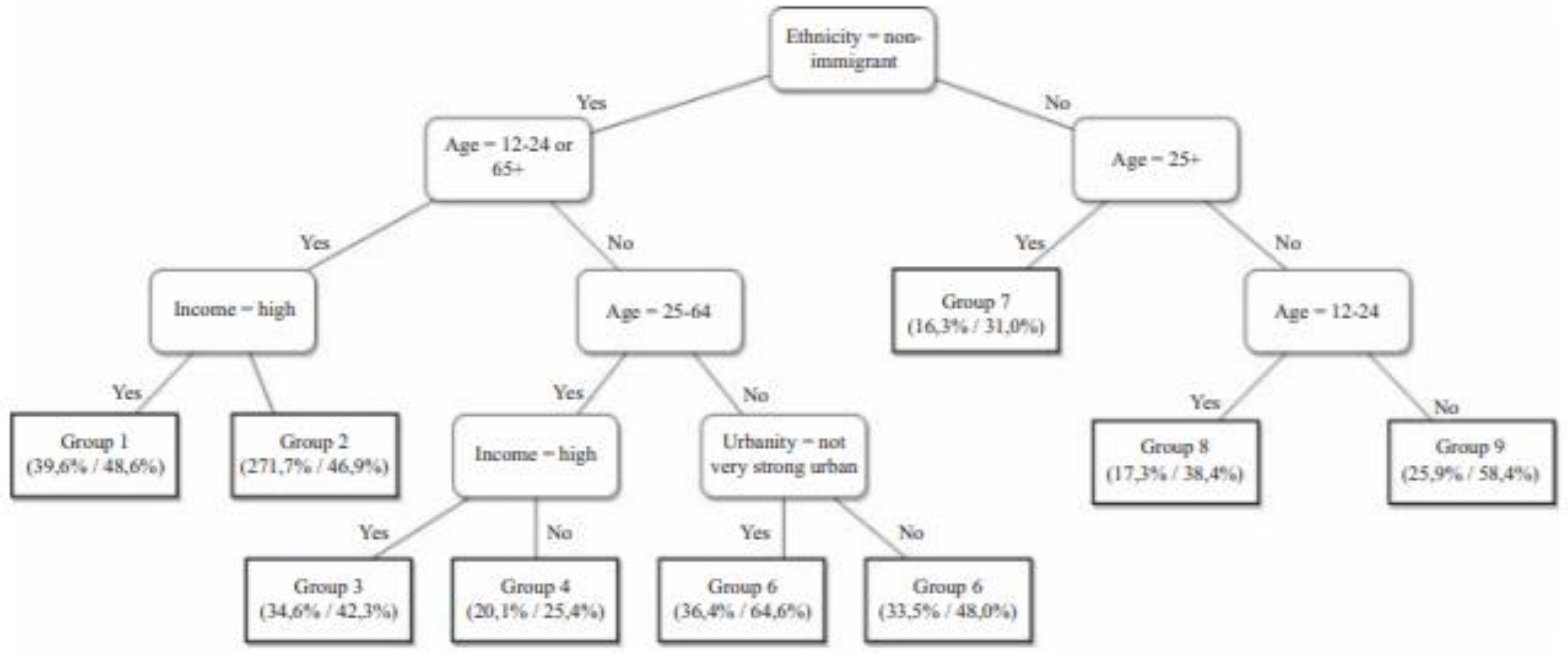
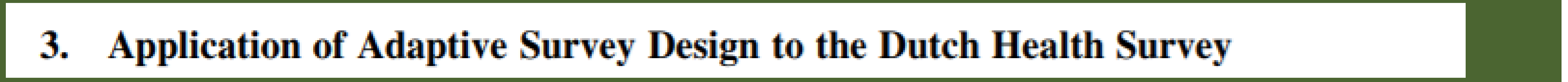
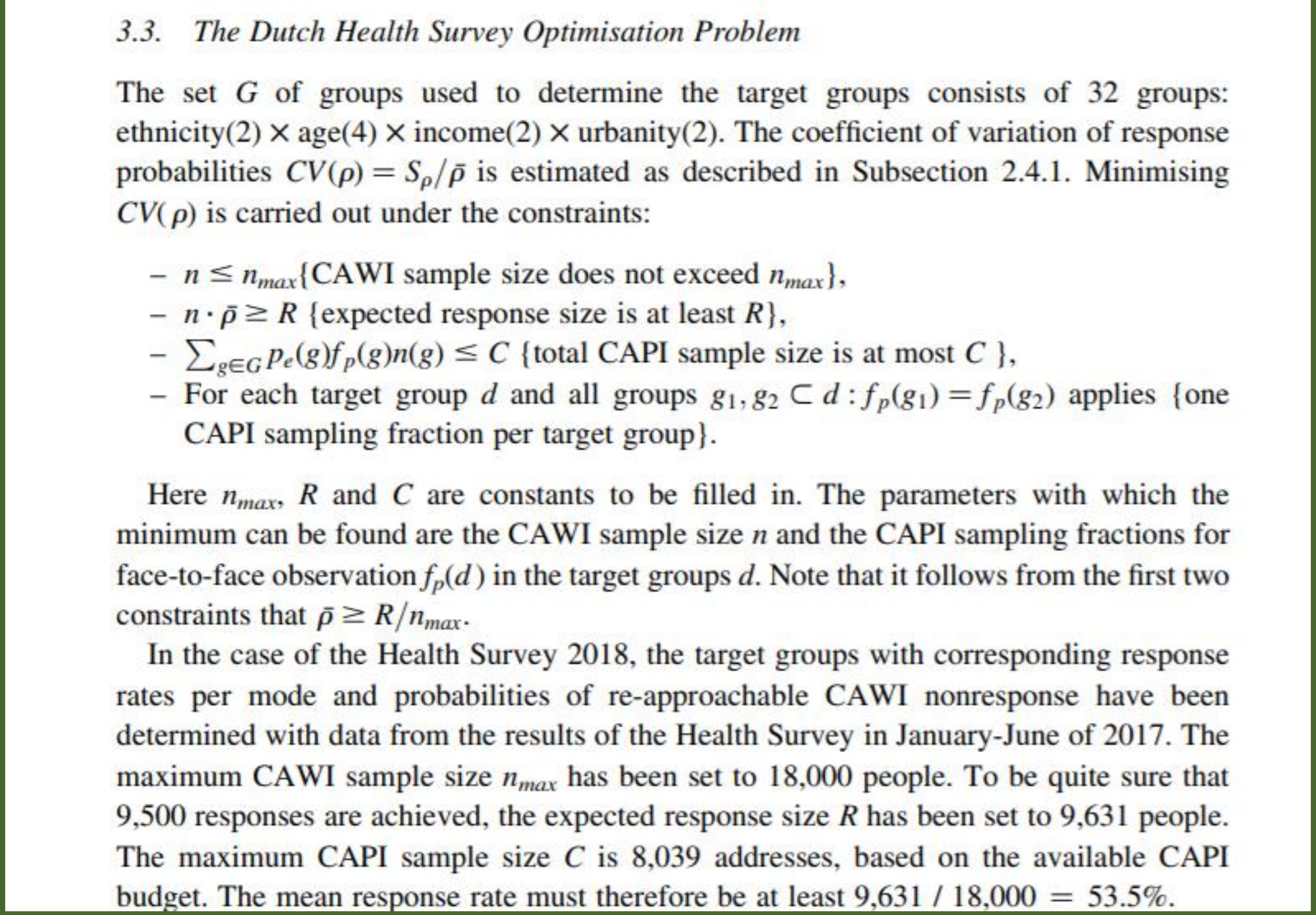


Fig. 1. Classification tree based on results of the Health Survey 2016.

Empirical Work / Results – Example 1



3. Application of Adaptive Survey Design to the Dutch Health Survey



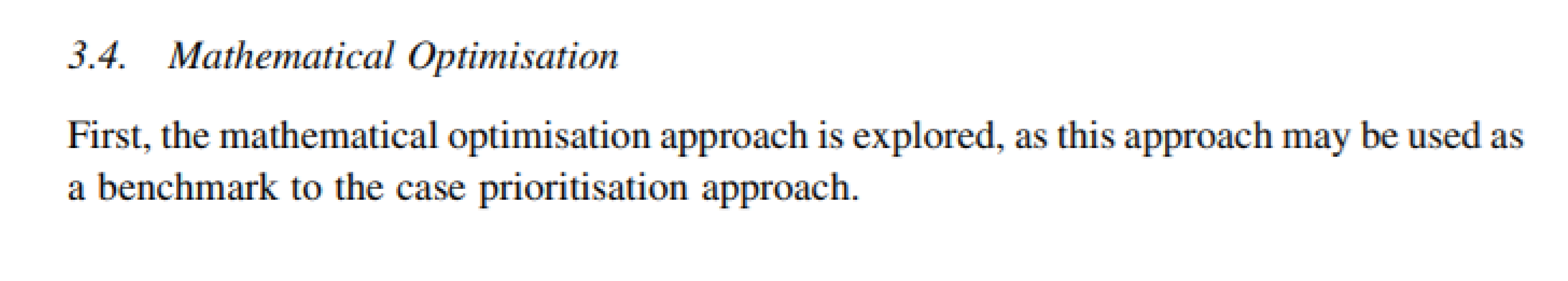
3.3. The Dutch Health Survey Optimisation Problem

The set G of groups used to determine the target groups consists of 32 groups: ethnicity(2) × age(4) × income(2) × urbanity(2). The coefficient of variation of response probabilities $CV(\rho) = S_\rho/\bar{\rho}$ is estimated as described in Subsection 2.4.1. Minimising $CV(\rho)$ is carried out under the constraints:

- $n \leq n_{max}$ {CAWI sample size does not exceed n_{max}},
- $n \cdot \bar{\rho} \geq R$ {expected response size is at least R},
- $\sum_{g \in G} p_e(g) f_p(g) n(g) \leq C$ {total CAPI sample size is at most C },
- For each target group d and all groups $g_1, g_2 \subset d : f_p(g_1) = f_p(g_2)$ applies {one CAPI sampling fraction per target group}.

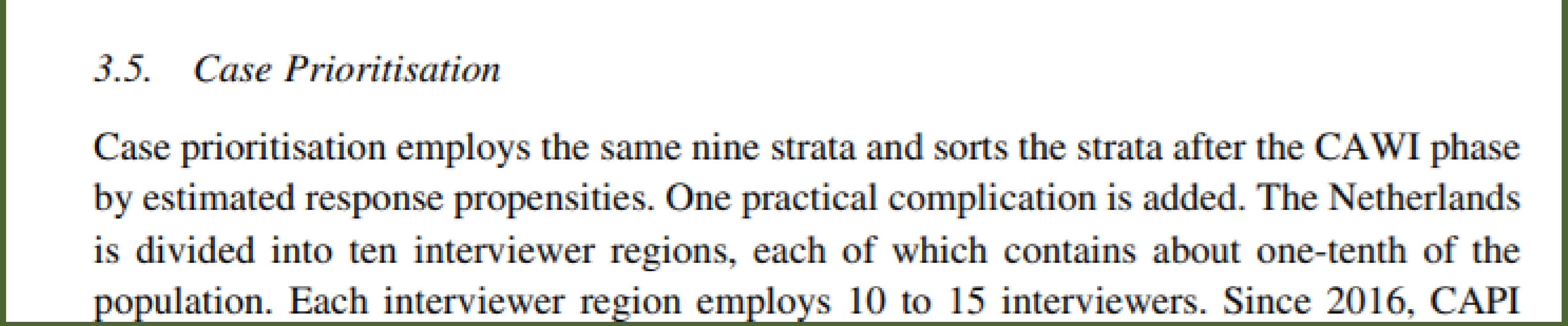
Here n_{max}, R and C are constants to be filled in. The parameters with which the minimum can be found are the CAWI sample size n and the CAPI sampling fractions for face-to-face observation $f_p(d)$ in the target groups d. Note that it follows from the first two constraints that $\bar{\rho} \geq R/n_{max}$.

In the case of the Health Survey 2018, the target groups with corresponding response rates per mode and probabilities of re-approachable CAWI nonresponse have been determined with data from the results of the Health Survey in January-June of 2017. The maximum CAWI sample size n_{max} has been set to 18,000 people. To be quite sure that 9,500 responses are achieved, the expected response size R has been set to 9,631 people. The maximum CAPI sample size C is 8,039 addresses, based on the available CAPI budget. The mean response rate must therefore be at least 9,631 / 18,000 = 53.5%.



3.4. Mathematical Optimisation

First, the mathematical optimisation approach is explored, as this approach may be used as a benchmark to the case prioritisation approach.



3.5. Case Prioritisation

Case prioritisation employs the same nine strata and sorts the strata after the CAWI phase by estimated response propensities. One practical complication is added. The Netherlands is divided into ten interviewer regions, each of which contains about one-tenth of the population. Each interviewer region employs 10 to 15 interviewers. Since 2016, CAPI

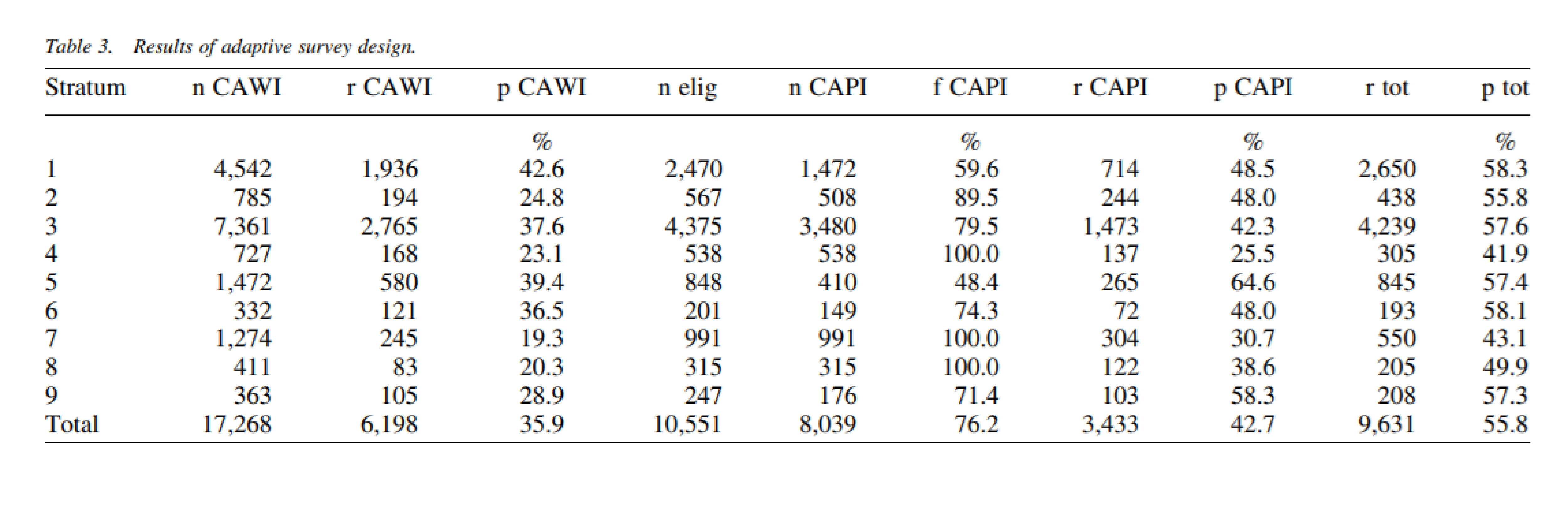
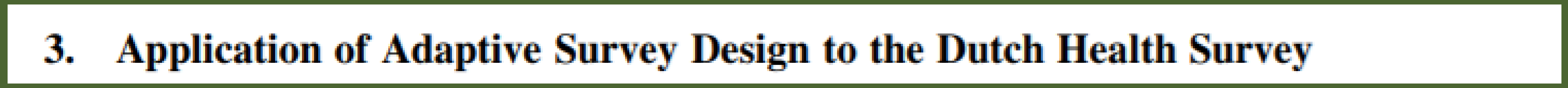


Table 3. Results of adaptive survey design.

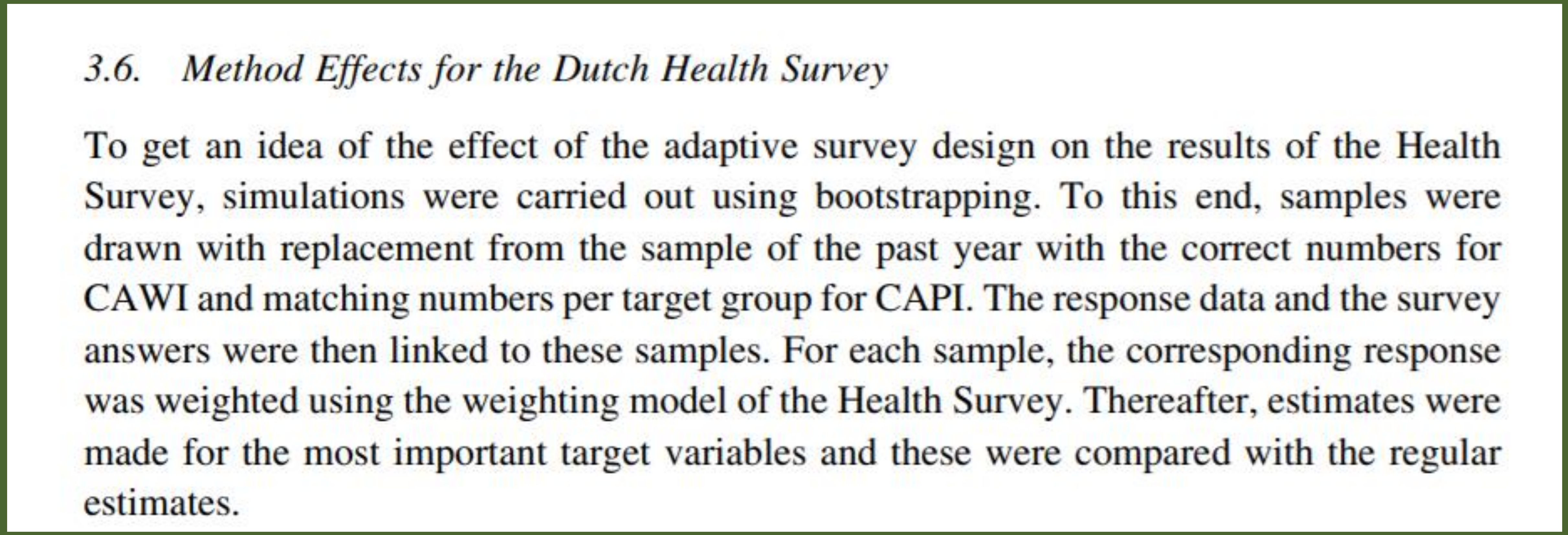
Stratum	n CAWI	r CAWI	p CAWI	n elig	n CAPI	f CAPI	r CAPI	p CAPI	r tot	p tot
			%			%		%		%
1	4,542	1,936	42.6	2,470	1,472	59.6	714	48.5	2,650	58.3
2	785	194	24.8	567	508	89.5	244	48.0	438	55.8
3	7,361	2,765	37.6	4,375	3,480	79.5	1,473	42.3	4,239	57.6
4	727	168	23.1	538	538	100.0	137	25.5	305	41.9
5	1,472	580	39.4	848	410	48.4	265	64.6	845	57.4
6	332	121	36.5	201	149	74.3	72	48.0	193	58.1
7	1,274	245	19.3	991	991	100.0	304	30.7	550	43.1
8	411	83	20.3	315	315	100.0	122	38.6	205	49.9
9	363	105	28.9	247	176	71.4	103	58.3	208	57.3
Total	17,268	6,198	35.9	10,551	8,039	76.2	3,433	42.7	9,631	55.8

https://doi.org/10.2478/jos-2020-0031

Empirical Work / Results – Example 1



3. Application of Adaptive Survey Design to the Dutch Health Survey

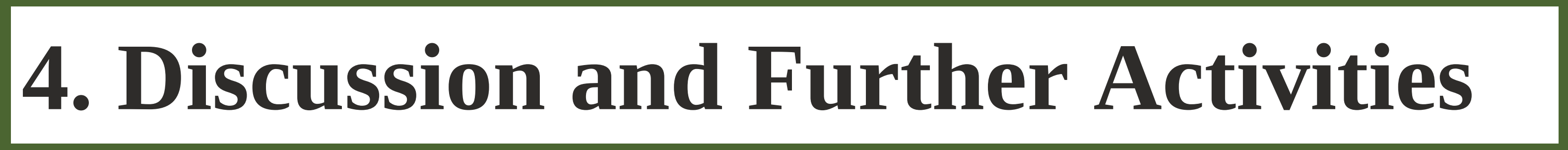


3.6. Method Effects for the Dutch Health Survey

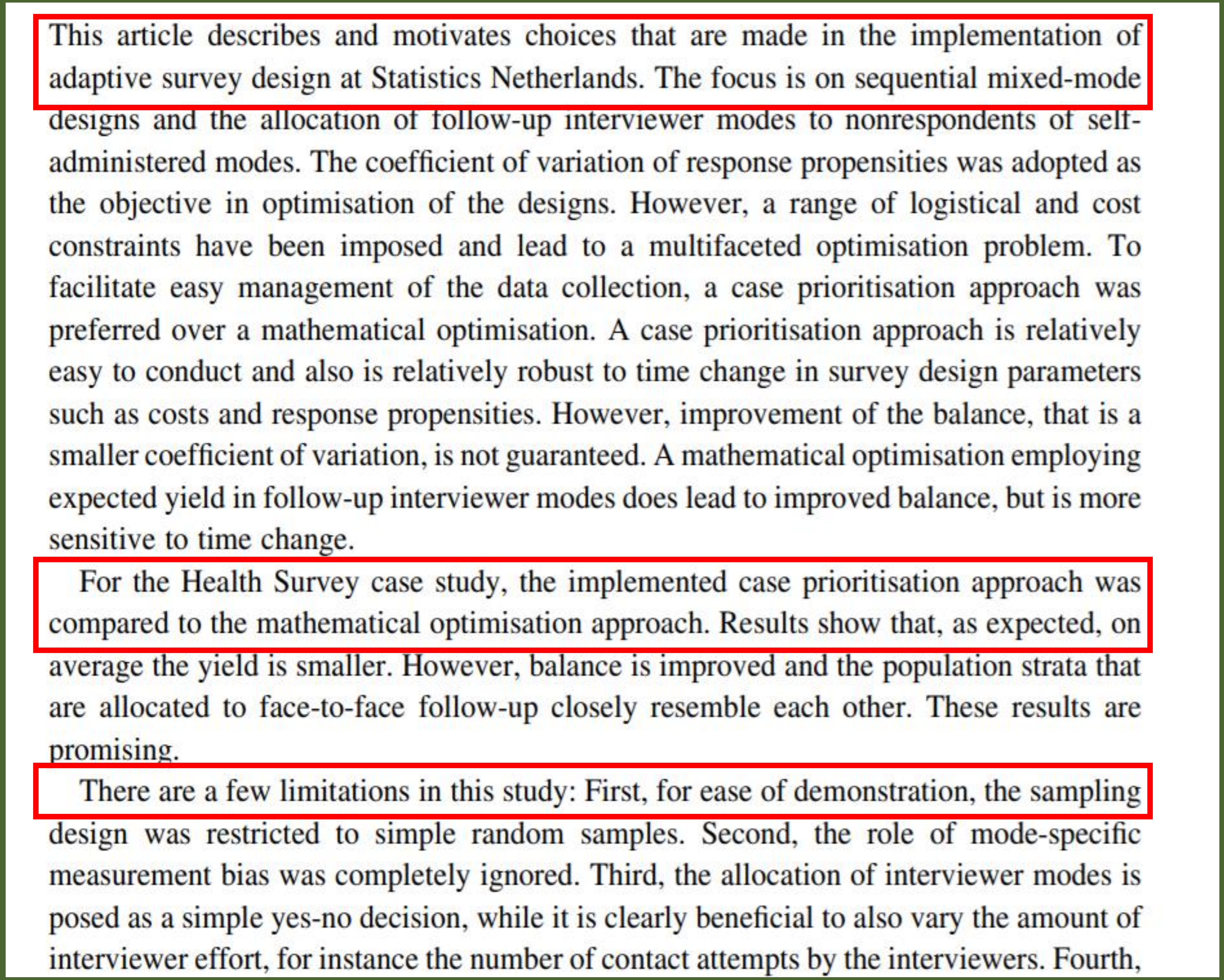
To get an idea of the effect of the adaptive survey design on the results of the Health Survey, simulations were carried out using bootstrapping. To this end, samples were drawn with replacement from the sample of the past year with the correct numbers for CAWI and matching numbers per target group for CAPI. The response data and the survey answers were then linked to these samples. For each sample, the corresponding response was weighted using the weighting model of the Health Survey. Thereafter, estimates were made for the most important target variables and these were compared with the regular estimates.

Section 3.6 presents results of simulations and compare them with regular estimates

https://doi.org/10.2478/jos-2020-0031



4. Discussion and Further Activities

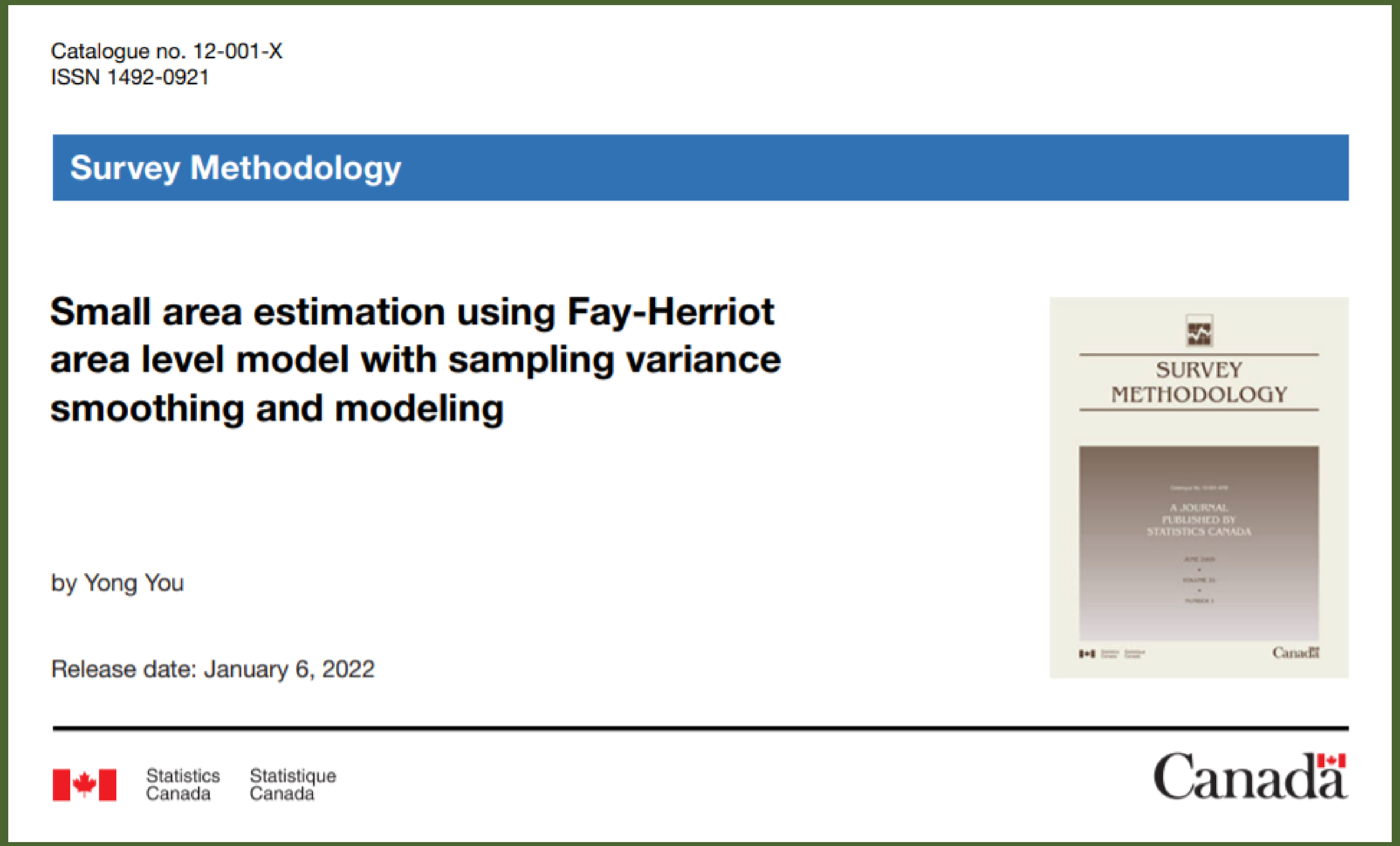


This article describes and motivates choices that are made in the implementation of adaptive survey design at Statistics Netherlands. The focus is on sequential mixed-mode designs and the allocation of follow-up interviewer modes to nonrespondents of self-administered modes. The coefficient of variation of response propensities was adopted as the objective in optimisation of the designs. However, a range of logistical and cost constraints have been imposed and lead to a multifaceted optimisation problem. To facilitate easy management of the data collection, a case prioritisation approach was preferred over a mathematical optimisation. A case prioritisation approach is relatively easy to conduct and also is relatively robust to time change in survey design parameters such as costs and response propensities. However, improvement of the balance, that is a smaller coefficient of variation, is not guaranteed. A mathematical optimisation employing expected yield in follow-up interviewer modes does lead to improved balance, but is more sensitive to time change.

For the Health Survey case study, the implemented case prioritisation approach was compared to the mathematical optimisation approach. Results show that, as expected, on average the yield is smaller. However, balance is improved and the population strata that are allocated to face-to-face follow-up closely resemble each other. These results are promising.

There are a few limitations in this study: First, for ease of demonstration, the sampling design was restricted to simple random samples. Second, the role of mode-specific measurement bias was completely ignored. Third, the allocation of interviewer modes is posed as a simple yes-no decision, while it is clearly beneficial to also vary the amount of interviewer effort, for instance the number of contact attempts by the interviewers. Fourth,

Methods – Example 2



Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

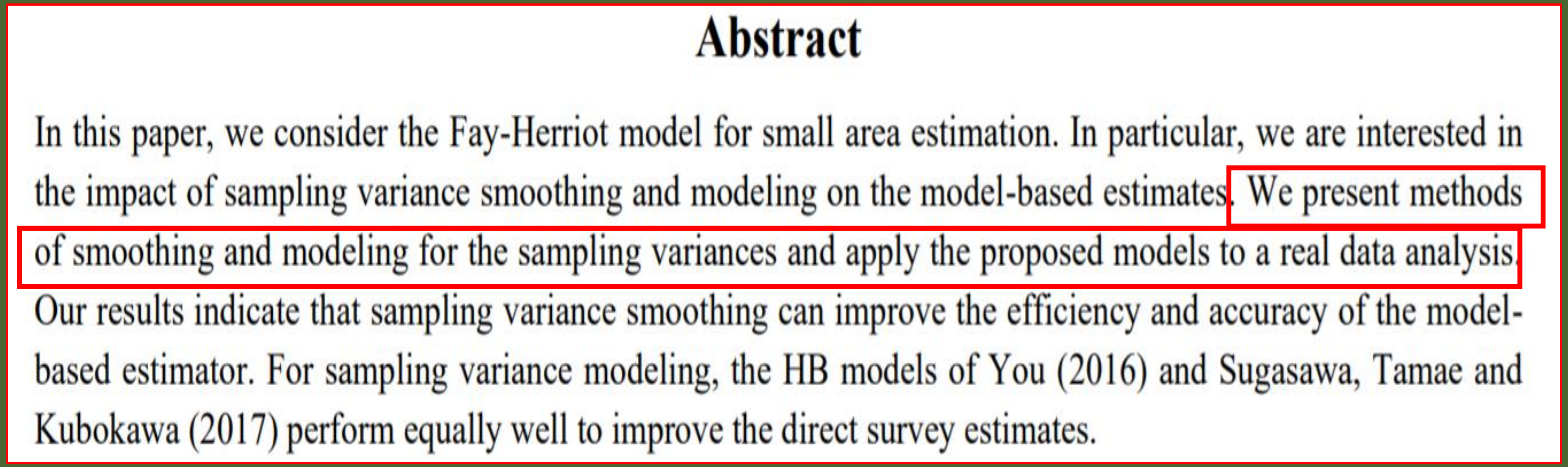
Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling

by Yong You

Release date: January 6, 2022

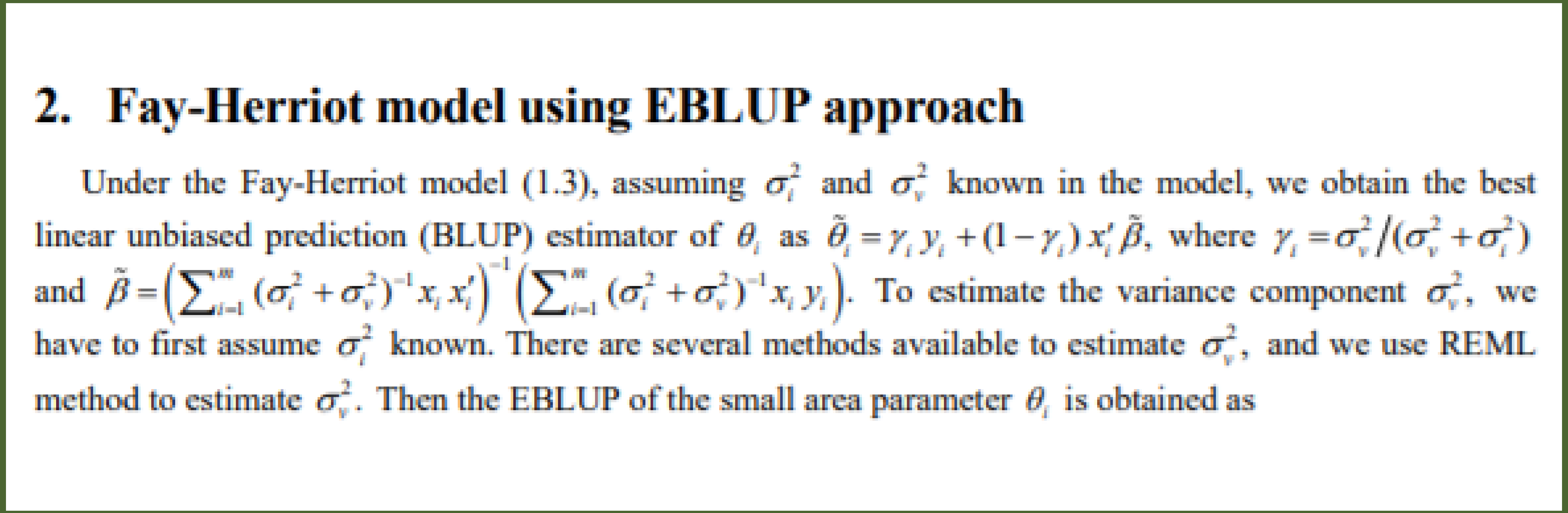
Statistics Canada Statistique Canada

Canadä



Abstract

In this paper, we consider the Fay-Herriot model for small area estimation. In particular, we are interested in the impact of sampling variance smoothing and modeling on the model-based estimates. We present methods of smoothing and modeling for the sampling variances and apply the proposed models to a real data analysis. Our results indicate that sampling variance smoothing can improve the efficiency and accuracy of the model-based estimator. For sampling variance modeling, the HB models of You (2016) and Sugasawa, Tamae and Kubokawa (2017) perform equally well to improve the direct survey estimates.

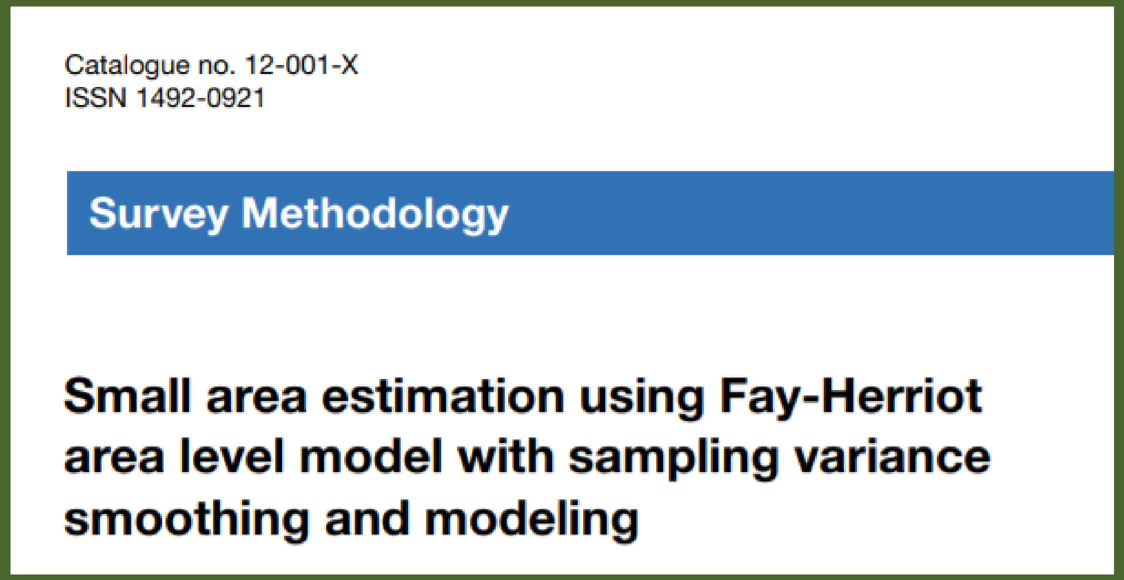


2. Fay-Herriot model using EBLUP approach

Under the Fay-Herriot model (1.3), assuming σ_i^2 and σ_v^2 known in the model, we obtain the best linear unbiased prediction (BLUP) estimator of θ_i as $\tilde{\theta}_i = \gamma_i y_i + (1-\gamma_i) x_i' \tilde{\beta}$, where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$ and $\tilde{\beta} = \left(\sum_{i=1}^m (\sigma_i^2 + \sigma_v^2)^{-1} x_i x_i' \right)^{-1} \left(\sum_{i=1}^m (\sigma_i^2 + \sigma_v^2)^{-1} x_i y_i \right)$. To estimate the variance component σ_v^2, we have to first assume σ_i^2 known. There are several methods available to estimate σ_v^2, and we use REML method to estimate σ_v^2. Then the EBLUP of the small area parameter θ_i is obtained as

https://doi.org/10.2478/jos-2020-0031

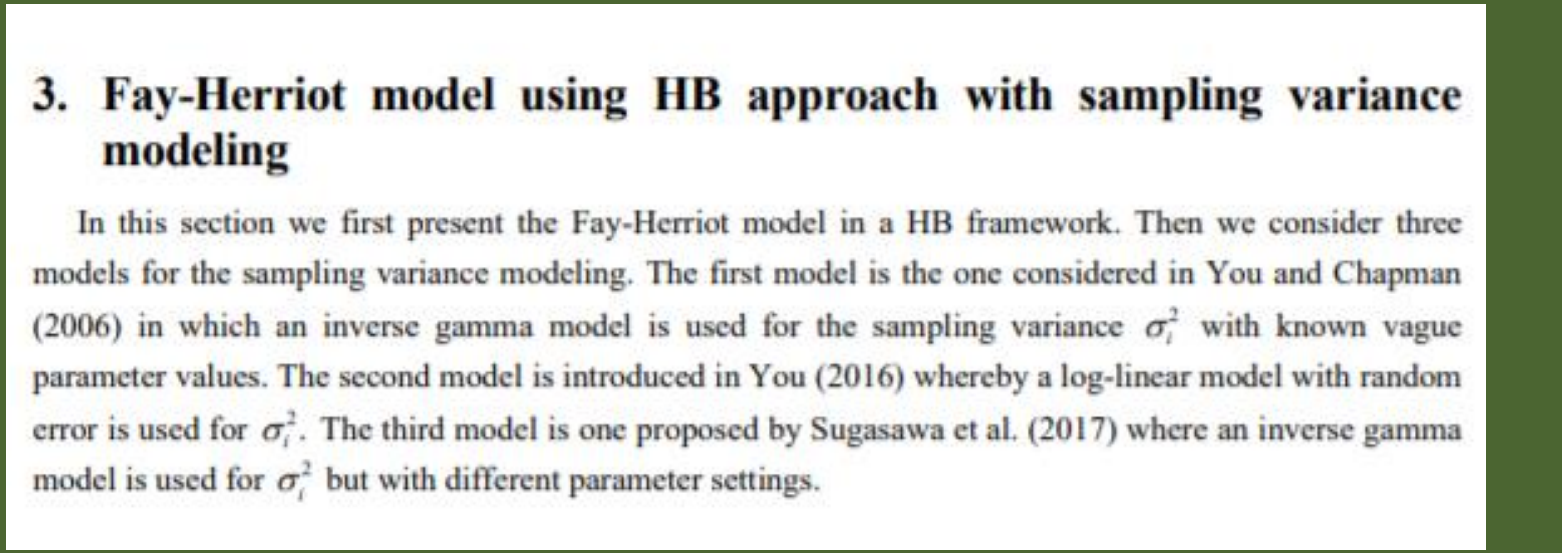
Methods- Example 2



Catalogue no. 12-001-X
ISSN 1492-0921

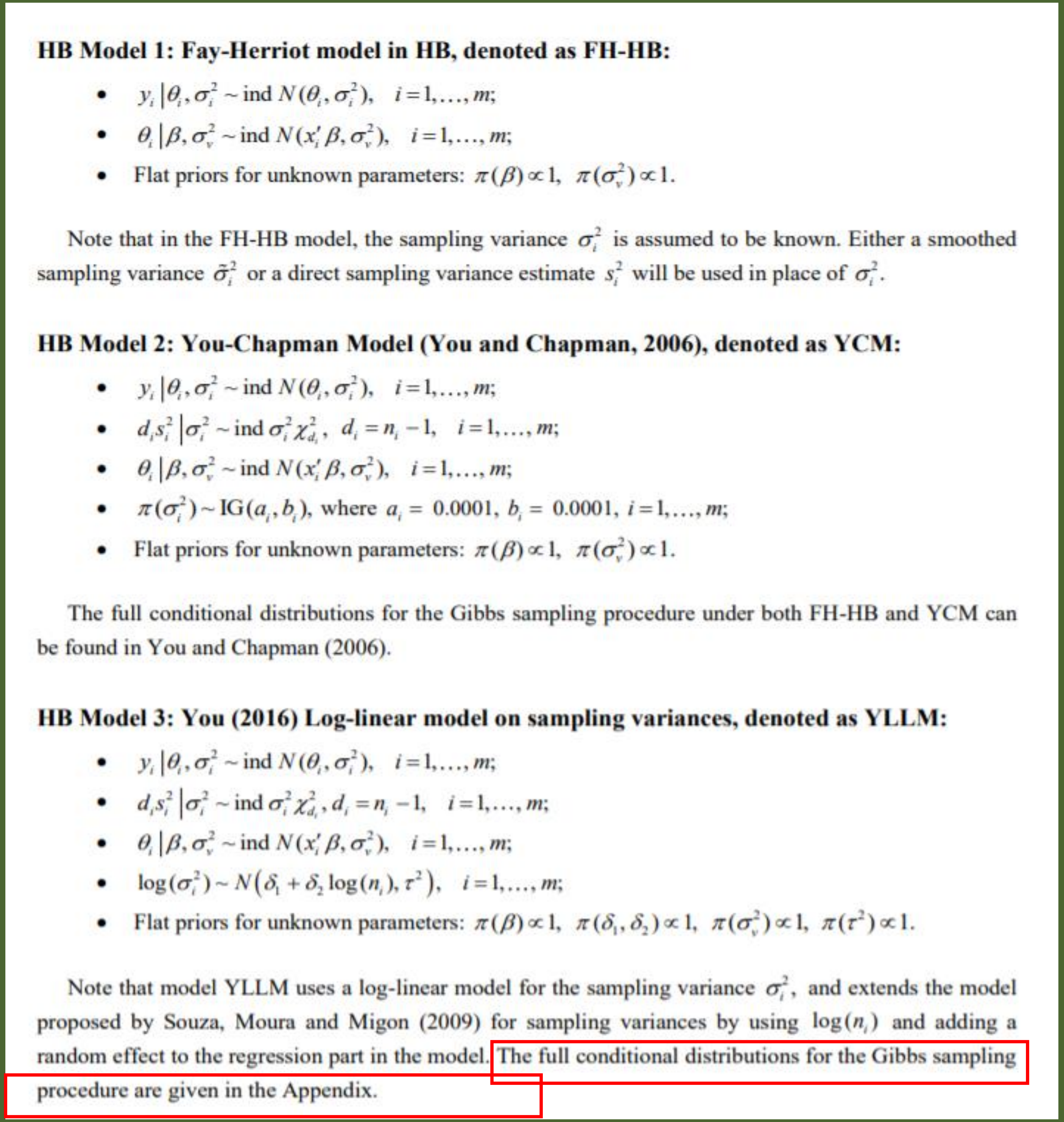
Survey Methodology

Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling



3. Fay-Herriot model using HB approach with sampling variance modeling

In this section we first present the Fay-Herriot model in a HB framework. Then we consider three models for the sampling variance modeling. The first model is the one considered in You and Chapman (2006) in which an inverse gamma model is used for the sampling variance σ_i^2 with known vague parameter values. The second model is introduced in You (2016) whereby a log-linear model with random error is used for σ_i^2. The third model is one proposed by Sugasawa et al. (2017) where an inverse gamma model is used for σ_i^2 but with different parameter settings.



HB Model 1: Fay-Herriot model in HB, denoted as FH-HB:

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1, \; \pi(\sigma_v^2) \propto 1.$

Note that in the FH-HB model, the sampling variance σ_i^2 is assumed to be known. Either a smoothed sampling variance $\tilde{\sigma}_i^2$ or a direct sampling variance estimate s_i^2 will be used in place of σ_i^2.

HB Model 2: You-Chapman Model (You and Chapman, 2006), denoted as YCM:

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi^2_{d_i}, \;\; d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$, where $a_i = 0.0001$, $b_i = 0.0001$, $i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1, \; \pi(\sigma_v^2) \propto 1.$

The full conditional distributions for the Gibbs sampling procedure under both FH-HB and YCM can be found in You and Chapman (2006).

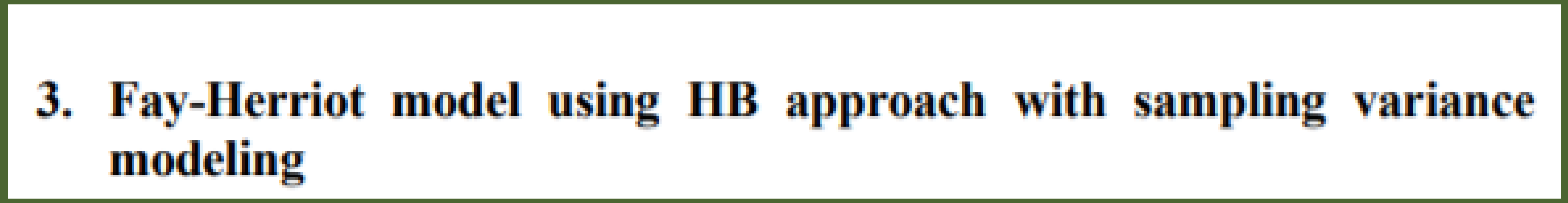
HB Model 3: You (2016) Log-linear model on sampling variances, denoted as YLLM:

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi^2_{d_i}, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\log(\sigma_i^2) \sim N\left(\delta_1 + \delta_2 \log(n_i), \tau^2\right), \quad i = 1, \ldots, m;$
- Flat priors for unknown parameters: $\pi(\beta) \propto 1, \; \pi(\delta_1, \delta_2) \propto 1, \; \pi(\sigma_v^2) \propto 1, \; \pi(\tau^2) \propto 1.$

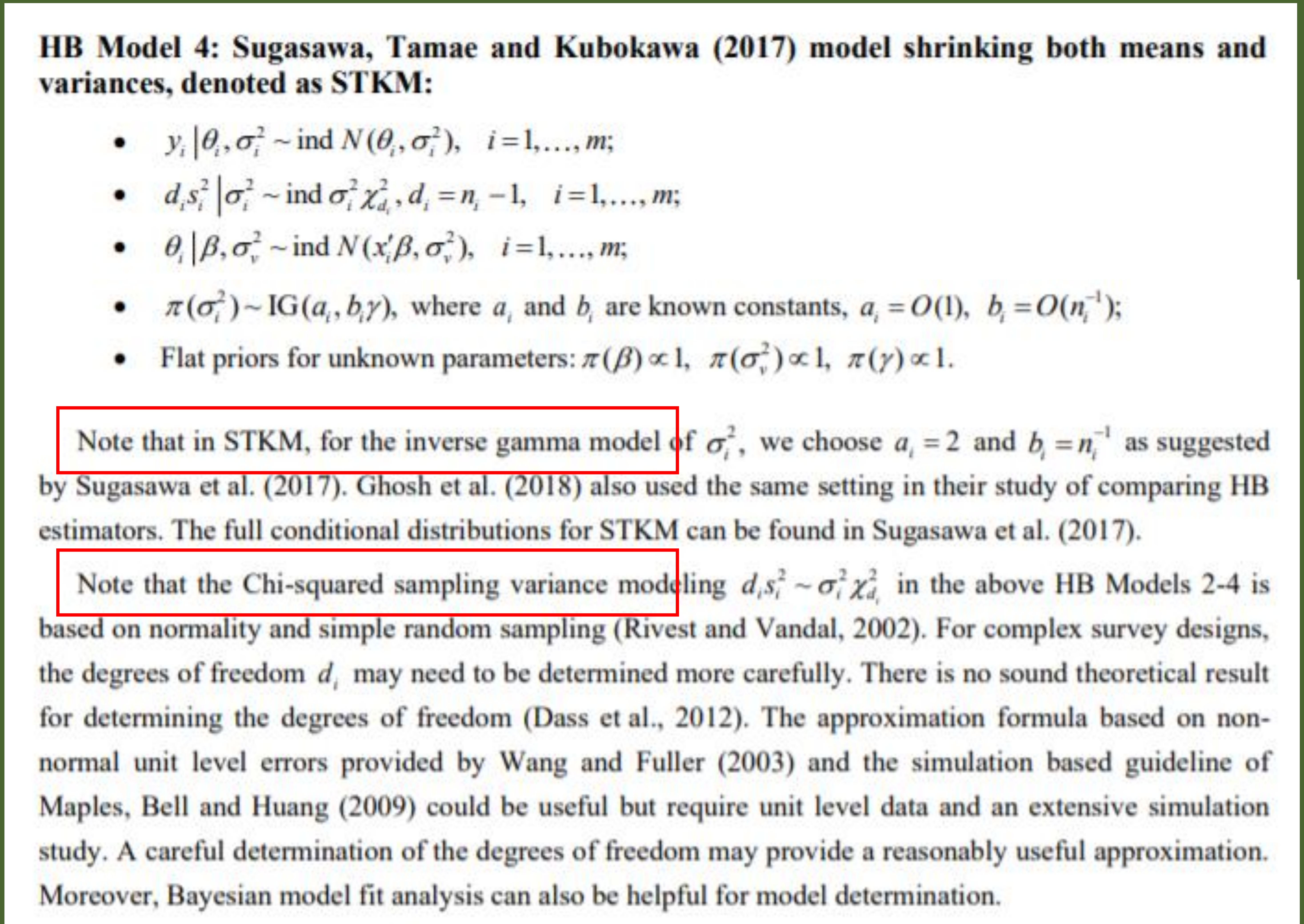
Note that model YLLM uses a log-linear model for the sampling variance σ_i^2, and extends the model proposed by Souza, Moura and Migon (2009) for sampling variances by using $\log(n_i)$ and adding a random effect to the regression part in the model. The full conditional distributions for the Gibbs sampling procedure are given in the Appendix.

https://doi.org/10.2478/jos-2020-0031

Methods- Example 2



3. **Fay-Herriot model using HB approach with sampling variance modeling**

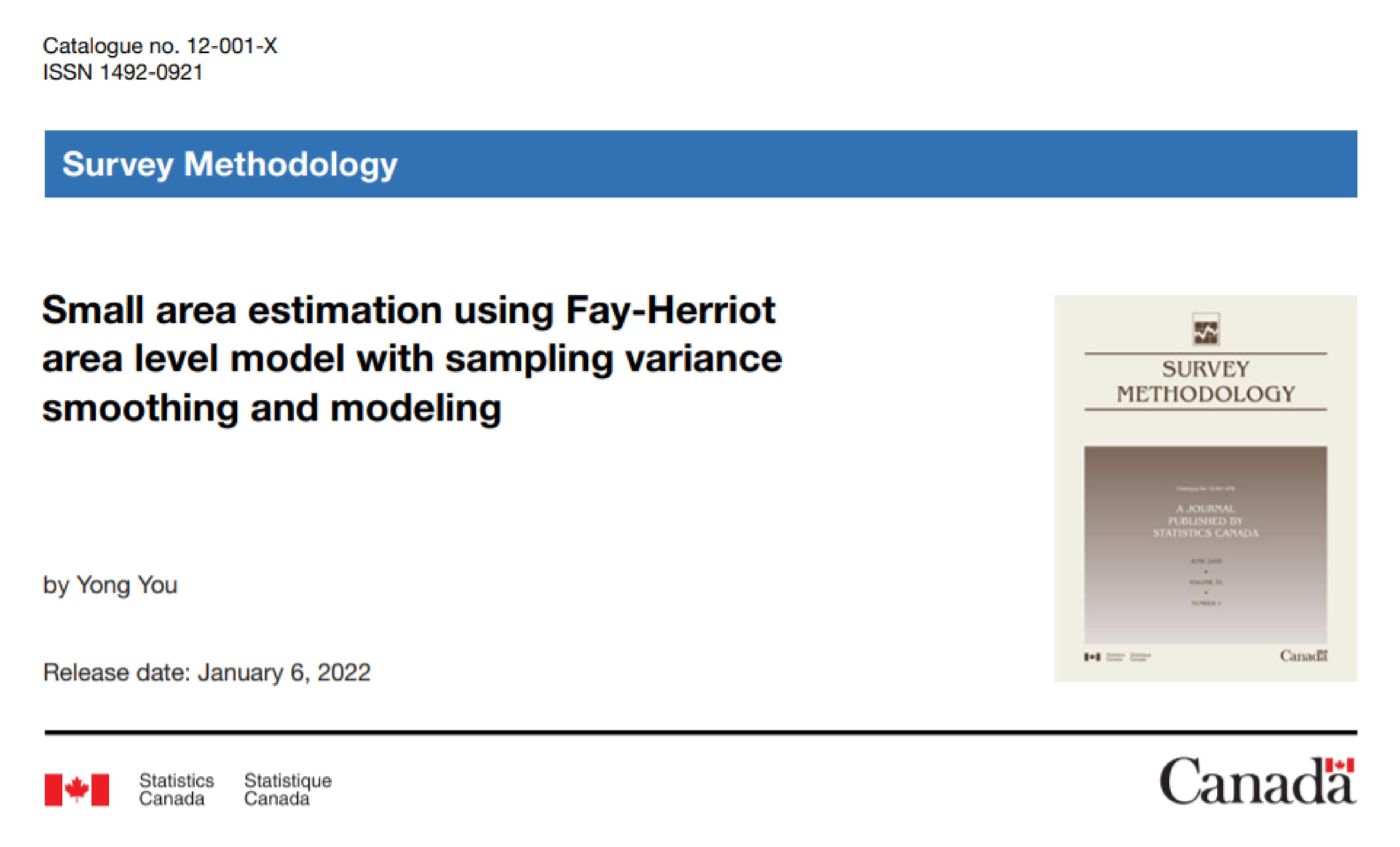


HB Model 4: Sugasawa, Tamae and Kubokawa (2017) model shrinking both means and variances, denoted as STKM:

- $y_i \,|\, \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), \quad i = 1, \ldots, m;$
- $d_i s_i^2 \,|\, \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi^2_{d_i}, d_i = n_i - 1, \quad i = 1, \ldots, m;$
- $\theta_i \,|\, \beta, \sigma_v^2 \sim \text{ind } N(x_i'\beta, \sigma_v^2), \quad i = 1, \ldots, m;$
- $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i\gamma)$, where a_i and b_i are known constants, $a_i = O(1)$, $b_i = O(n_i^{-1})$;
- Flat priors for unknown parameters: $\pi(\beta) \propto 1$, $\pi(\sigma_v^2) \propto 1$, $\pi(\gamma) \propto 1$.

Note that in STKM, for the inverse gamma model of σ_i^2, we choose $a_i = 2$ and $b_i = n_i^{-1}$ as suggested by Sugasawa et al. (2017). Ghosh et al. (2018) also used the same setting in their study of comparing HB estimators. The full conditional distributions for STKM can be found in Sugasawa et al. (2017).

Note that the Chi-squared sampling variance modeling $d_i s_i^2 \sim \sigma_i^2 \chi^2_{d_i}$ in the above HB Models 2-4 is based on normality and simple random sampling (Rivest and Vandal, 2002). For complex survey designs, the degrees of freedom d_i may need to be determined more carefully. There is no sound theoretical result for determining the degrees of freedom (Dass et al., 2012). The approximation formula based on non-normal unit level errors provided by Wang and Fuller (2003) and the simulation based guideline of Maples, Bell and Huang (2009) could be useful but require unit level data and an extensive simulation study. A careful determination of the degrees of freedom may provide a reasonably useful approximation. Moreover, Bayesian model fit analysis can also be helpful for model determination.



Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling

by Yong You

Release date: January 6, 2022

Statistics Canada Statistique Canada

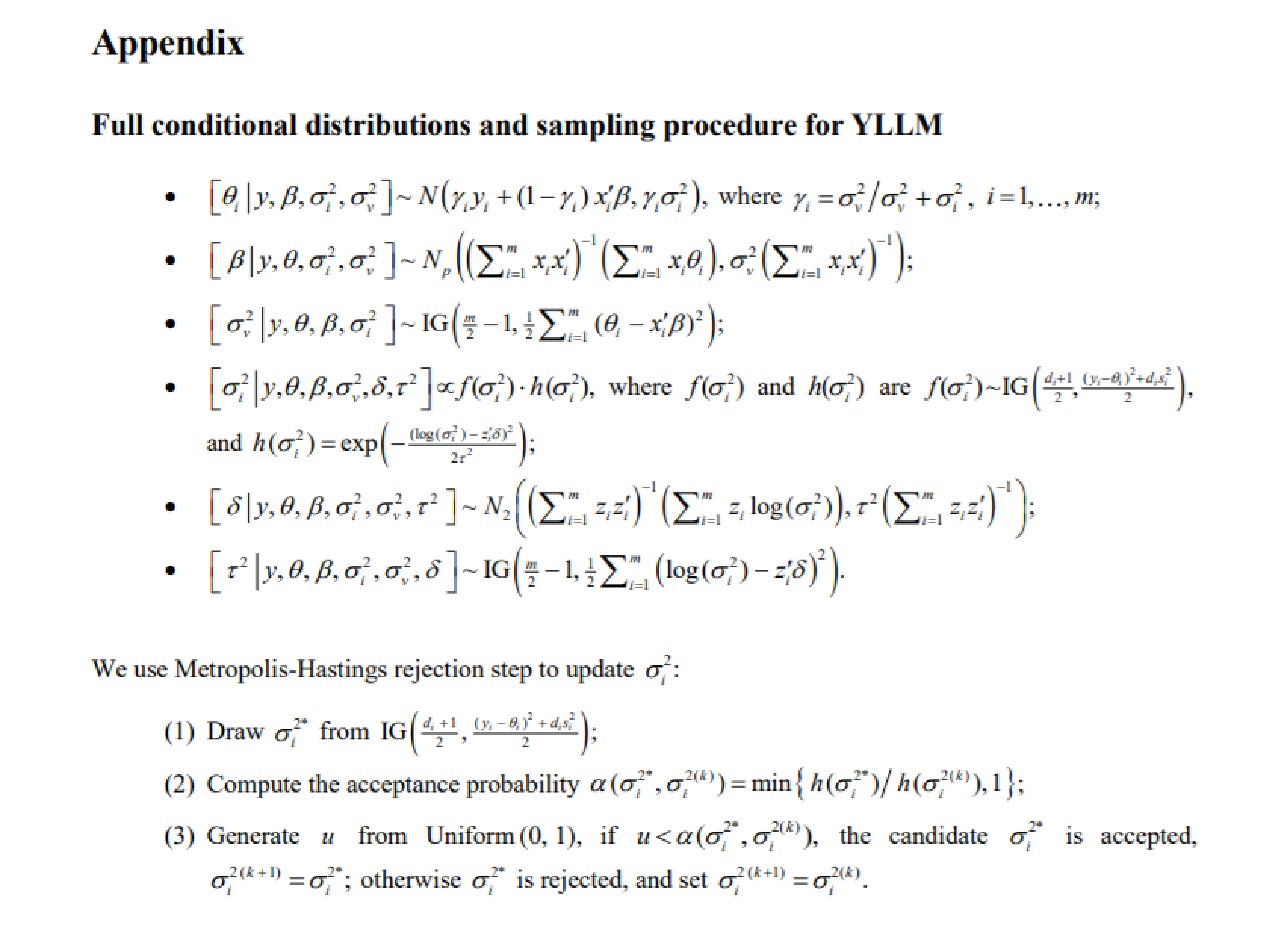
Canada

https://doi.org/10.2478/jos-2020-0031

Appendix - Example 2

Survey Methodology

Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling



Appendix

Full conditional distributions and sampling procedure for YLLM

- $\left[\theta_i \,|\, y, \beta, \sigma_i^2, \sigma_v^2\right] \sim N\left(\gamma_i y_i + (1-\gamma_i)\, x_i'\beta, \gamma_i \sigma_i^2\right)$, where $\gamma_i = \sigma_v^2 / \sigma_v^2 + \sigma_i^2$, $i = 1, \ldots, m$;
- $\left[\,\beta \,|\, y, \theta, \sigma_i^2, \sigma_v^2\,\right] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\left(\sum_{i=1}^m x_i \theta_i\right), \sigma_v^2\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right)$;
- $\left[\,\sigma_v^2 \,|\, y, \theta, \beta, \sigma_i^2\,\right] \sim \mathrm{IG}\left(\frac{m}{2} - 1, \frac{1}{2}\sum_{i=1}^m (\theta_i - x_i'\beta)^2\right)$;
- $\left[\sigma_i^2 \,|\, y, \theta, \beta, \sigma_v^2, \delta, \tau^2\right] \propto f(\sigma_i^2) \cdot h(\sigma_i^2)$, where $f(\sigma_i^2)$ and $h(\sigma_i^2)$ are $f(\sigma_i^2) \sim \mathrm{IG}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, and $h(\sigma_i^2) = \exp\left(-\frac{(\log(\sigma_i^2) - z_i'\delta)^2}{2\tau^2}\right)$;
- $\left[\,\delta \,|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \tau^2\,\right] \sim N_2\left(\left(\sum_{i=1}^m z_i z_i'\right)^{-1}\left(\sum_{i=1}^m z_i \log(\sigma_i^2)\right), \tau^2\left(\sum_{i=1}^m z_i z_i'\right)^{-1}\right)$;
- $\left[\,\tau^2 \,|\, y, \theta, \beta, \sigma_i^2, \sigma_v^2, \delta\,\right] \sim \mathrm{IG}\left(\frac{m}{2} - 1, \frac{1}{2}\sum_{i=1}^m \left(\log(\sigma_i^2) - z_i'\delta\right)^2\right)$.

We use Metropolis-Hastings rejection step to update σ_i^2:

(1) Draw σ_i^{2*} from $\mathrm{IG}\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$;

(2) Compute the acceptance probability $\alpha(\sigma_i^{2*}, \sigma_i^{2(k)}) = \min\left\{h(\sigma_i^{2*}) / h(\sigma_i^{2(k)}), 1\right\}$;

(3) Generate u from Uniform (0, 1), if $u < \alpha(\sigma_i^{2*}, \sigma_i^{2(k)})$, the candidate σ_i^{2*} is accepted, $\sigma_i^{2(k+1)} = \sigma_i^{2*}$; otherwise σ_i^{2*} is rejected, and set $\sigma_i^{2(k+1)} = \sigma_i^{2(k)}$.

https://doi.org/10.2478/jos-2020-0031

Results - Example 2

Survey Methodology

Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling

SURVEY METHODOLOGY

https://doi.org/10.2478/jos-2020-0031

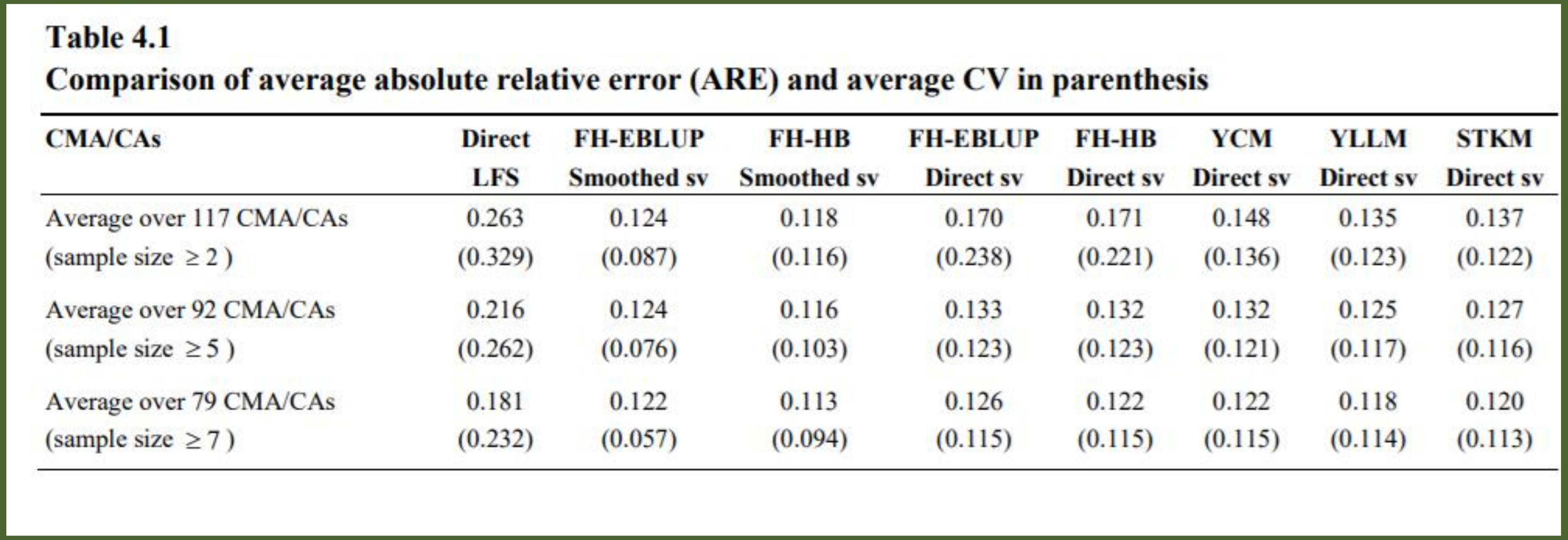
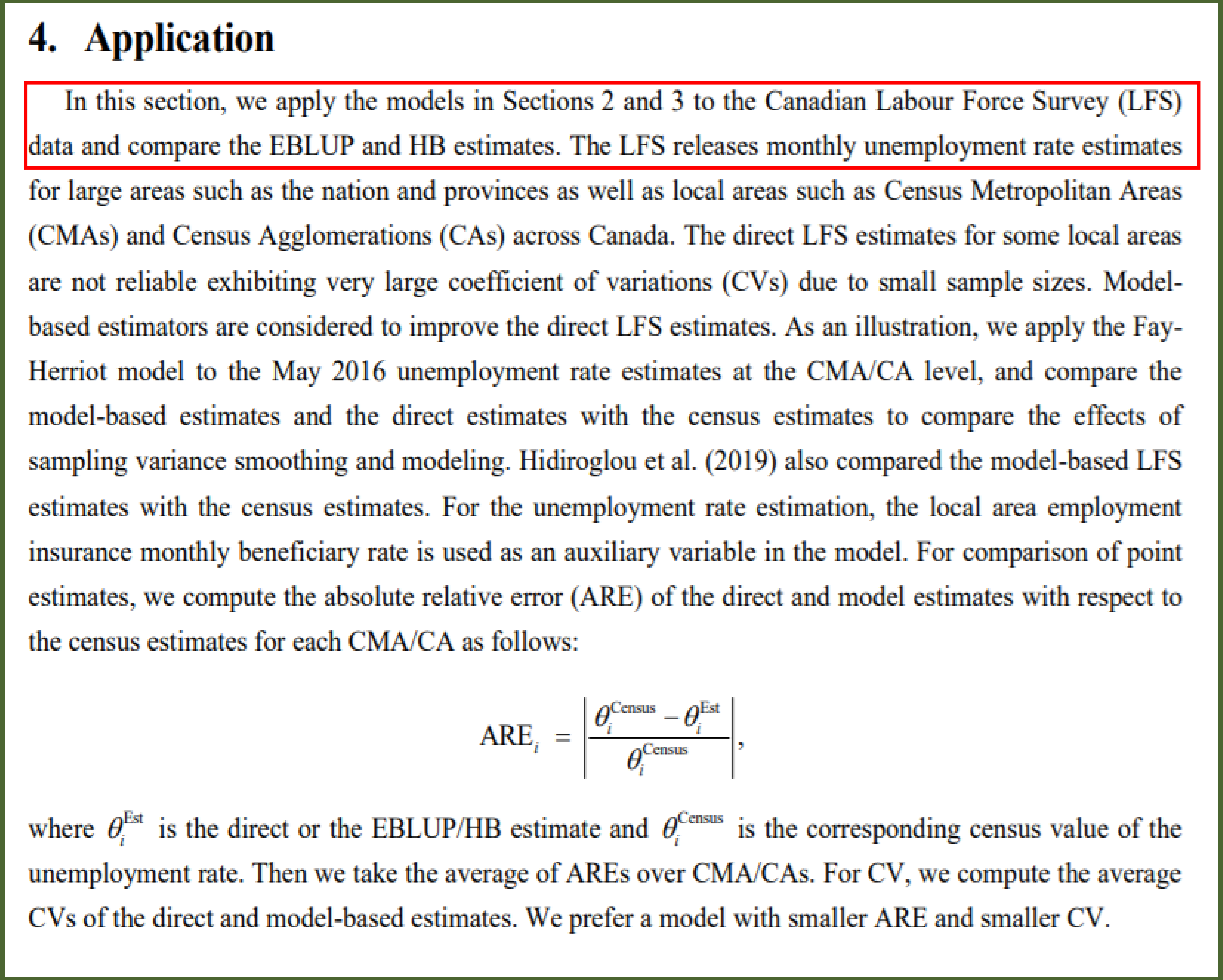


Table 4.1
Comparison of average absolute relative error (ARE) and average CV in parenthesis

CMA/CAs	Direct LFS	FH-EBLUP Smoothed sv	FH-HB Smoothed sv	FH-EBLUP Direct sv	FH-HB Direct sv	YCM Direct sv	YLLM Direct sv	STKM Direct sv
Average over 117 CMA/CAs	0.263	0.124	0.118	0.170	0.171	0.148	0.135	0.137
(sample size ≥ 2)	(0.329)	(0.087)	(0.116)	(0.238)	(0.221)	(0.136)	(0.123)	(0.122)
Average over 92 CMA/CAs	0.216	0.124	0.116	0.133	0.132	0.132	0.125	0.127
(sample size ≥ 5)	(0.262)	(0.076)	(0.103)	(0.123)	(0.123)	(0.121)	(0.117)	(0.116)
Average over 79 CMA/CAs	0.181	0.122	0.113	0.126	0.122	0.122	0.118	0.120
(sample size ≥ 7)	(0.232)	(0.057)	(0.094)	(0.115)	(0.115)	(0.115)	(0.114)	(0.113)



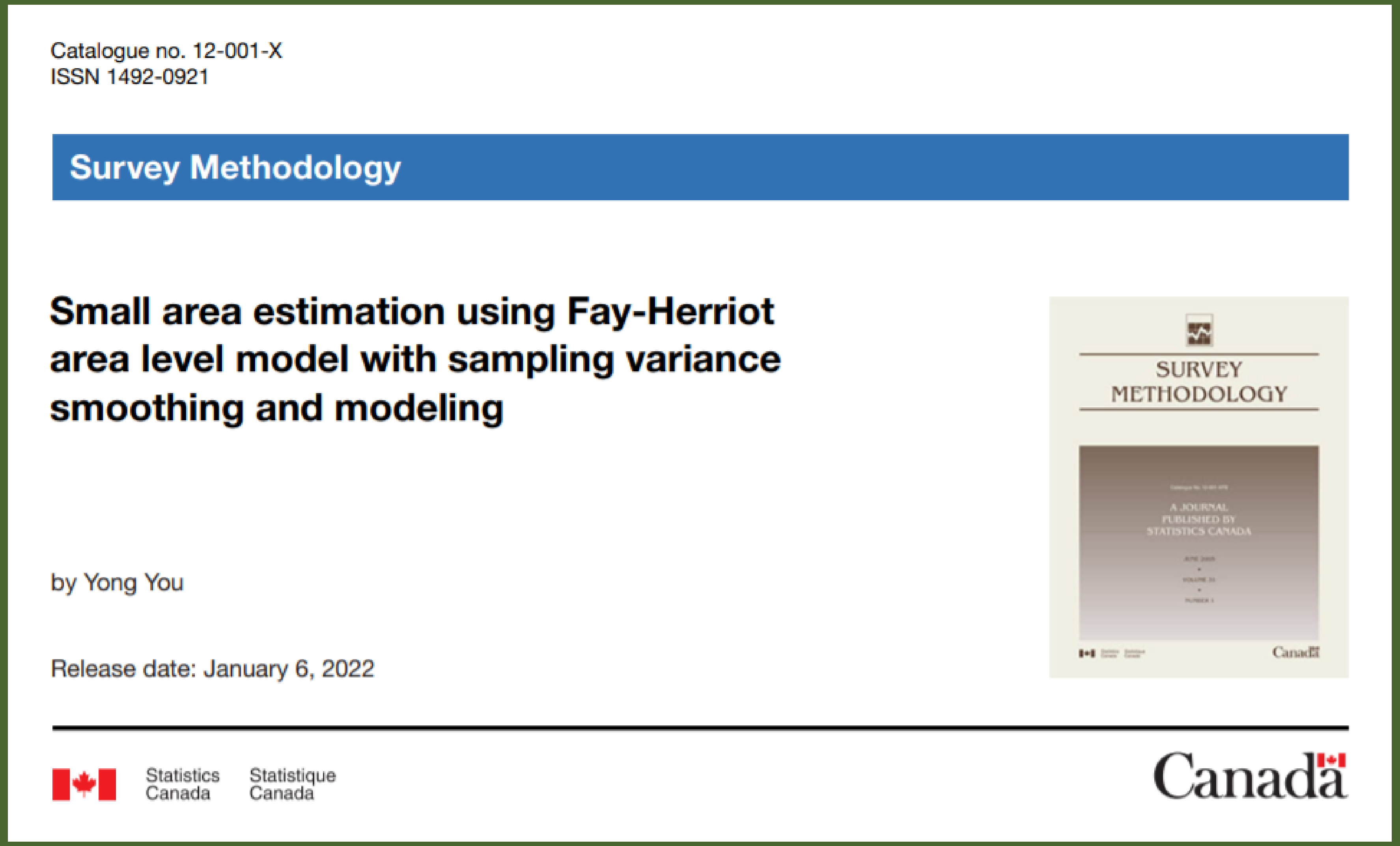
4. Application

In this section, we apply the models in Sections 2 and 3 to the Canadian Labour Force Survey (LFS) data and compare the EBLUP and HB estimates. The LFS releases monthly unemployment rate estimates for large areas such as the nation and provinces as well as local areas such as Census Metropolitan Areas (CMAs) and Census Agglomerations (CAs) across Canada. The direct LFS estimates for some local areas are not reliable exhibiting very large coefficient of variations (CVs) due to small sample sizes. Model-based estimators are considered to improve the direct LFS estimates. As an illustration, we apply the Fay-Herriot model to the May 2016 unemployment rate estimates at the CMA/CA level, and compare the model-based estimates and the direct estimates with the census estimates to compare the effects of sampling variance smoothing and modeling. Hidiroglou et al. (2019) also compared the model-based LFS estimates with the census estimates. For the unemployment rate estimation, the local area employment insurance monthly beneficiary rate is used as an auxiliary variable in the model. For comparison of point estimates, we compute the absolute relative error (ARE) of the direct and model estimates with respect to the census estimates for each CMA/CA as follows:

$$\text{ARE}_i = \left| \frac{\theta_i^{\text{Census}} - \theta_i^{\text{Est}}}{\theta_i^{\text{Census}}} \right|,$$

where θ_i^{Est} is the direct or the EBLUP/HB estimate and θ_i^{Census} is the corresponding census value of the unemployment rate. Then we take the average of AREs over CMA/CAs. For CV, we compute the average CVs of the direct and model-based estimates. We prefer a model with smaller ARE and smaller CV.

Conclusions – Example 2



Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

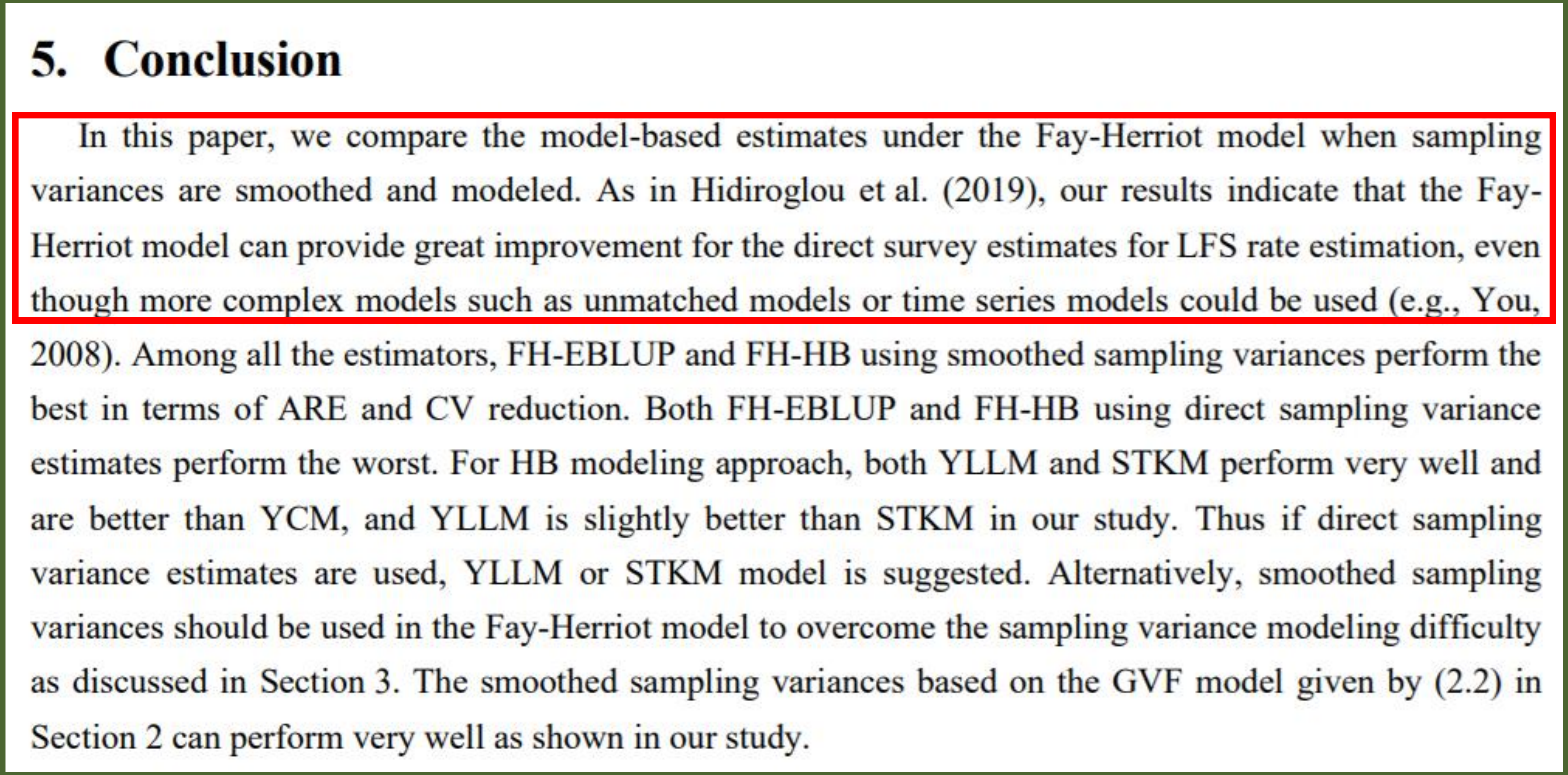
Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling

by Yong You

Release date: January 6, 2022

Statistics Canada Statistique Canada

Canadä



5. Conclusion

In this paper, we compare the model-based estimates under the Fay-Herriot model when sampling variances are smoothed and modeled. As in Hidiroglou et al. (2019), our results indicate that the Fay-Herriot model can provide great improvement for the direct survey estimates for LFS rate estimation, even though more complex models such as unmatched models or time series models could be used (e.g., You, 2008). Among all the estimators, FH-EBLUP and FH-HB using smoothed sampling variances perform the best in terms of ARE and CV reduction. Both FH-EBLUP and FH-HB using direct sampling variance estimates perform the worst. For HB modeling approach, both YLLM and STKM perform very well and are better than YCM, and YLLM is slightly better than STKM in our study. Thus if direct sampling variance estimates are used, YLLM or STKM model is suggested. Alternatively, smoothed sampling variances should be used in the Fay-Herriot model to overcome the sampling variance modeling difficulty as discussed in Section 3. The smoothed sampling variances based on the GVF model given by (2.2) in Section 2 can perform very well as shown in our study.

https://doi.org/10.2478/jos-2020-0031

General Comments

- State the conclusion concisely and avoid overstatements
- Concise and informative headings (and subheadings) helps organizing the manuscript
- Be careful with notation and definition of variables
- Check the journal’s word limit for a manuscript

References

Journal of Young Investigators. (2005). Writing Scientific Manuscripts, a guide for undergraduates.

https://ugr.ue.ucsc.edu/sites/default/files/jyi_guide_to_scientific_writing.pdf