



Journal of Official Statistics vol. 39, 2 (June 2023)

Effects of Changing Modes on Item Nonresponse in Panel Surveys.....139-149
Oliver Lipps, Marieke Voorpostel and Gian-Andrea Monsch

Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach.....151-172
Daniela Marella

Design and Sample Size Determination for Experiments on Nonresponse Followup using a Sequential Regression Model.....173-202
Andrew M. Raim, Thomas Mathew, Kimberly F. Sellers, Renee Ellis and Mikelyn Meyers

Estimating Intra-Regional Inequality with an Application to German Spatial Planning Regions...203-228
Marina Runge

Constructing Building Price Index Using Administrative Data.....229-251
Masahiro Higo, Yumi Saita, Chihiro Shimizu and Yuta Tachi

From Quarterly to Monthly Turnover Figures Using Nowcasting Methods.....253-273
Daan Zult, Sabine Krieg, Bernd Schouten, Pim Ouwehand and Jan van den Brakel

Effects of Changing Modes on Item Nonresponse in Panel Surveys

Oliver Lipps¹, Marieke Voorpostel¹, and Gian-Andrea Monsch²

To investigate the effect of a change from the telephone to the web mode on item nonresponse in panel surveys, we use experimental data from a two-wave panel survey. The treatment group changed from the telephone to the web mode after the first wave, while the control group continued in the telephone mode. We find that when changing to the web, “don’t know” answers increase moderately from a low level, while item refusal increases substantially from a very low level. This is the case for all person groups, although socio-demographic characteristics have some additional effects on giving a don’t know or a refusal when changing mode.

Key words: telephone; web; don’t know; item refusal.

1. Introduction

Several household panel surveys have explored the feasibility of changing from an (expensive) interviewer-based survey mode to the (cheaper) web mode (Voorpostel et al. 2021). One concern with such a change is that the web mode produces more item nonresponse (INR) than interviewer-based modes such as telephone surveys (Bowling 2005; De Leeuw 2005; Guzy and Leitgöb 2015). Reasons for this are in particular the absence of an interviewer to motivate the respondent and to provide help to find a substantive answer. Differences between interviewer-based modes and the web on INR has been mostly researched for cross-sectional surveys (e.g., Lipps and Monsch 2022). The problem in a cross-sectional design is that it is difficult to separate effects from selection into a survey mode and measurement issues (Vannieuwenhuyze and Loosveldt 2013).

In this article, we overcome this shortcoming by using experimental data from a two-wave panel, in which the treatment group changes from the telephone to the web survey mode while the control group keeps the telephone mode. This allows us to investigate individual changes for each variable using a difference-in-difference design. Our research question is whether changing the mode leads to more or less “don’t know” (DK) or item refusal (REF) in the second wave and whether our findings differ across different socio-demographic groups.

¹Foundation for Research in Social Sciences (FORS), University of Lausanne, Bâtiment Géopolis, CH – 1015 Lausanne, Switzerland. Emails: oliver.lipps@fors.unil.ch and marieke.voorpostel@fors.unil.ch

²Swiss Federal Statistical Office Espace de l’Europe 10, CH - 2010 Neuchâtel, Switzerland Email: gianandrea.monsch@bfs.admin.ch

Acknowledgments: This study has used data collected by the Swiss Household Panel, which is based at the Swiss Centre of Expertise in the Social Sciences (FORS) and financed by the Swiss National Science Foundation.

1.1. What Affects Item (Non-)responses?

The literature mentions four main factors that affect INR: the respondent, the question (Lipps and Monsch 2022), the interviewer (Silber et al. 2021), and the survey mode. In this research, we focus on effects from the respondent and in particular the survey mode.

As for “respondent characteristics”, older, very young, lower educated, and those with a higher response burden people have more problems with question comprehension and have a higher likelihood of giving an INR (Fricker et al. 2005). Holbrook et al. (2006) suggest that the way survey questions are constructed relies on cultural conceptions of the dominant cultural group such that respondents from language and cultural minorities have more difficulties to understand survey questions and to give adequate answers. Finally, question comprehension and motivation of respondents is related to INR (Kleiner et al. 2015) because motivation decreases satisficing and increases the cognitive effort of the interviewee when responding to survey questions (Krosnick 1991; Shoemaker et al. 2002).

While respondent characteristics are well established factors in the literature on INR, research on *mode* effects, most importantly, the presence or absence of an interviewer, is relatively new. The presence of an interviewer generally lowers INR because the interviewer motivates respondents to complete the task and can provide more explanation when a question is not understood (De Leeuw and Hox 2018; Groves et al. 2011; Silber et al. 2021). See Gooch and Vavreck (2019) for an exception.

However, an interviewer may also exert social pressure to give a substantively valid answer even if the respondent does not want or cannot answer a question (Chang and Krosnick 2010). INR often remains “hidden” in interviewer-based surveys as respondents, instead of not answering a question, may provide an answer that is socially accepted, for example a mid-scale response (Lipps and Monsch 2022; Sturgis et al. 2014). The mode of interview also affects responses through variation in other characteristics, such as the pace of the interview, presentation (visual or auditory), and the timing of the interview (Christian et al. 2007). Although more time and flexibility for web respondents and the possibility to reread the question may decrease INR, the evidence points to the opposite (Fricker et al. 2005). While an interviewer may classify a response as DK or REF without explicitly proposing this option to the respondent, in web surveys, the DK response category is often offered to the respondent to account for the fact that DK may be a valid response. Skipped questions are then interpreted as refusals (REF). Presenting DK and/or REF answers as visually explicit options is an important reason for obtaining a high rate of INR in the web mode.

1.2. Distinguishing DK and REF in Telephone and Web Surveys and by Respondent Characteristics

DK and REF are different types of INR, which result from different “disruptions of the cognitive response process” (Silber et al. 2021, 1; Beatty and Herrmann 2002). In the cognitive response process (Tourangeau and Rasinski 1988), respondents need to “understand” the question, retrieve adequate *information* to answer the question, evaluate whether they “are willing” to give this answer, and finally find the right category on the response scale. Beatty and Herrmann (2002) distinguish different states of the retrieval step, ranging from answers being easily available and retrievable to not being known. DK

answers will be given when a question is not well understood or a valid answer cannot be provided (Chang and Krosnick 2010; Young 2012), while REF will occur if the respondent has understood the question but does not want to give a valid answer. Often, REF results from social desirability if the respondent considers the valid answer to be inadequate (Krumpal 2013). Studies using unfolding brackets when answers about income are not given provide interesting information about the difference between DK and REF (Juster and Smith 1997). After an INR, bracket questions start by asking if the income is greater than x (e.g., EUR 5,000), then – once a substantive answer is given (e.g., no) – by asking if the income is greater than y (e.g., EUR 3,000), and so on. After all brackets (usually three or four) are answered, the ultimate income range is relatively small. Juster and Smith (1997) find that while 80% of initial DK respondents completed all brackets, this was the case for only 40% of initial REF respondents. Such findings show that DK answers may be truly valid but more substantive values can be retrieved if help or motivation is provided.

Self-administered modes generally provide more INR in particular to questions that are difficult or do not relate to respondents' life circumstances (Greene et al. 2008). More specifically, Lipps and Monsch (2022) found more DK answers in the web mode if calculation is needed, suggesting that interviewers can motivate and help respondents to answer such questions.

There is not much evidence about how socio-demographic characteristics interact with survey modes. Lipps and Monsch (2022) find that older, better educated, and male respondents give fewer DK responses in web surveys than in telephone surveys, while the contrary holds for foreigners from countries that speak a different language than the survey language. REFs in telephone surveys are more common among foreigners, and the more educated. Younger people, the more educated and men refuse answering a question less often in web surveys than in telephone surveys.

Given these still inconclusive findings, apart from an increase in REF answers in web surveys, we do not have clear-cut hypotheses about possible mode effects for different socio-demographic person groups. However, we expect more DK after a switch from telephone to web, especially from less educated respondents or those less familiar with the question topic.

2. Data and Method

2.1. Data

We use data from a two-wave pilot of the Swiss Household Panel (SHP) comparing telephone to web. The SHP is a longitudinal study that interviews randomly selected households and their members on an annual basis since 1999, predominantly by telephone (Tillmann et al. 2021). In preparation of the third refreshment sample, which was launched in 2020, a mode experiment conducted in 2017–2018 compared the standard telephone-based recruitment and fieldwork strategy with two web alternatives. This article is based on only a part of the data collected in the experiment. For the complete mode experiment, see Voorpostel et al. (2020).

The SHP conducts interviews with the household reference person (HRP) to collect information on the household and its members (household questionnaire) as well as with

all household members to complete an individual questionnaire. We include all HRPs who participated in a telephone interview in wave 1 of the pilot. Part of this sample completed the household questionnaire in the second wave again by telephone (tel-to-tel group) and part of the sample completed a web survey (tel-to-web group).

Figure 1 gives an overview of the modes of the tel-to-tel group and the tel-to-web group in both waves.

While Lipps and Monsch (2022) used data from the individual questionnaire of wave 1 of this experiment, we use data from the household questionnaire answered by the HRP in wave 1 and 2 in this article, because only for the household questionnaire did a mode-switch take place. Unlike Lipps and Monsch (2022), we are not able to distinguish question characteristics such as length or complexity of the questions, as the questions included in the household questionnaire were mostly factual questions. The sample for the study was a simple random sample of individuals stratified by region, drawn from a sampling frame based on population registers maintained by the Swiss Federal Statistical Office. The households of the sampled individuals were randomly assigned to one of the treatment groups. The sampled individual was approached first as an HRP, although in both waves and treatment groups, households were free to select an alternative HRP than the one initially approached. Although face-to-face and web were offered as alternatives if no telephone number was available and to initial refusals, our study only includes HRPs who participated by telephone in the first wave.

Response rates, RRI (AAPOR 2016) in the first wave on the household level were 53% for the tel-to-tel group ($n = 1\,214$ responding households) and (a not significantly different on the 5% level) 52% for the tel-to-web group ($n = 342$ responding households). The reason that the sample of the tel-to-tel group is much larger is that telephone on the household level in both waves was used in the control group, but also in one of the experimental groups. These two groups were combined in this analysis. All households that completed at least the grid in the first wave and that had not left the study were re-approached for the second wave. Response rates on the household level in wave 2 were 77% for the tel-to-tel group (935 households), and (a significantly different) 74% for the tel-to-web group (253 households).

We analyse INR by HRPs who participated in both waves and who answered in the assigned mode ($N = 967$ households). In addition, we only keep households with the same HRP in both waves ($N = 889$ households). These comprise 699 households in the tel-to-tel group and 190 households in the tel-to-web group. All analyses are done using Stata 16 SE.

2.2. Measures

We analyze all questions asked in the household questionnaire. The questions are typical for large household panel surveys and comprise questions in the domains of the division of





	Group 1 (telephone-to-telephone)	Group 2 (telephone-to-web)
Wave 1		
Wave 2		

Fig. 1. Research design for the two-wave pilot study of the SHP_IV (adapted from Voorpostel et al. 2020).

unpaid work, outsourcing of household tasks, household income, material deprivation, and housing. For each variable we code whether the HRP gave a (1) substantive answer, a (2) don't know (DK), or a (3) item refusal (REF). Then, we calculate for each HRP and for each question the following transitions between the two waves:

- Substantive answer to substantive answer (subs-to-subs)
- Substantive answer to DK (subs-to-DK)
- Substantive answer to REF (subs-to-ref)
- INR (DK or REF) in the first wave (INR-to-any)

INR in the latter group is not further split out according to answers in the second wave, because there are only few INR in the first wave. We provide univariate descriptive statistics distinguished by mode for the dependent variables in [Table 1](#).

All transitions are significantly different across the two groups ($\Pr(|T| > |t|) < 0.05$). In particular, changing to the web mode produces significantly more substantive answers to change to both DK and REF than when keeping the telephone mode. Considering the tel-to-tel group alone, there are fewer transitions from a substantive answer to INR (transitions 2 and 3) than starting from an INR in the first wave (transition 4). This suggests positive panel conditioning effects (e.g., [Sun et al. 2019](#)).

3. Multivariate Models

The main independent research variable is the transition to web versus repeating a telephone interview. The regression models described below include socio-demographic variables associated with survey participation and panel attrition ([Roberts and Vandenplas 2017](#)). The goal is to (partially) control for differences in the sample composition between the two groups. Except for education, all socio-demographic variables are included in the sampling frame and do not have missing values. First, we included whether the household has a registered landline (1 = yes, 0 = no (reference category)), gender (1 = male, 0 = female (reference category)), old (1 = over 54 years, 0 = 54 years or younger (reference category)), foreigner (1 = foreign nationality, 0 = Swiss nationality (reference category)), education (1 = tertiary level, 0 = lower than tertiary level (reference category)), and finally the number of question asked. Our substantive interest is the different effect of these variables on INR in the tel-to-tel and the tel-to-web group. In [Table 2](#), we show descriptive statistics of the independent variables by survey mode in the second wave.

Although there are only small insignificant differences (on the 5% level) in sample composition by mode, with one exception. The exception concerns the dummy variable

Table 1. Distribution of dependent variables (Transitions) by mode in the second wave.

Transition	Tel.-to-tel.		Tel.-to-web	
	mean	s.e.	mean	s.e.
1) Substantive answer to substantive answer	.986	.0010	.969	.0042
2) Substantive answer to DK	.004	.0004	.011	.0016
3) Substantive answer to REF	.001	.0005	.014	.0032
4) Item non-response in the first wave	.009	.0008	.006	.0010

Note: N(tel- to-tel.) = 699 households, N(tel.-to-web) = 190 households.

Table 2. Distribution of independent variables by mode in second wave.

Variable	Tel-to-tel.		Tel.-to-web	
	mean	s.e.	mean	s.e.
Registered landline at home	.73	.02	.71	.03
Male	.42	.02	.48	.04
Foreign nationality	.13	.01	.13	.02
Tertiary education level	.30	.02	.31	.03
Old (> 54 years)*	.52	.02	.39	.04
Number of questions	45.97	.20	45.87	.40

Note: N(tel.-to-tel.) = 699 households, N(tel.-to-web) = 190 households.

*: mean(tel.-to-tel.) statistically (5%) different from mean (tel.-to-web).

indicating whether the HRP is older than 54: the tel-to-tel group is more likely to include these older respondents in wave 2 compared with the tel-to-web group.

Like [Lipps and Monsch \(2022\)](#), to analyze the effects of respondent characteristics in the tel-to-tel and the tel-to-web conditions, we transform the data to a “cross-classified” structure (see [Fielding and Goldstein 2006](#)). Thus, we account for the crossed structure of the two levels respondents and questions: questions are clustered in respondents and respondents are clustered in questions. This structure allows modeling respondent and question variance in the same two-level model and to disentangle respondent and question effects on INR. Cross-classified models can be estimated using Bayesian Markov Chain Monte Carlo (MCMC) methods (see [Browne 2019](#)). To compare nested models, we use the Bayesian Deviance Information Criterion (DIC), which is an MCMC penalized goodness of fit measure, equivalent to the Akaike Information Criterion (AIC) used in maximum likelihood estimation. A smaller DIC means a better model fit. We use the `runmlwin` ([Leckie and Charlton 2013](#)) command in Stata version 16, which estimates the models by using the MLwiN software.

After dropping variables that were answered by fewer than 30 respondents, the data set consists of 40,766 respondent-question combinations (32,066 telephone and 8,700 web), including 92 variables and between 32 and 889 respondents per variable. Since each of the measured variables has potentially four nominal categories (subs-to-subs, subs-to-DK, subs-to-REF, INR-to-any), we would have to use a multinomial logit model (using, e.g., subs-to-subs as reference category). Because the cross-classified model with a multinomial distribution does not converge, we resort to using three binary models, modeling subs-to-DK, subs-to-REF, and INR-to-any (versus subs-to-subs) separately. We start with the base models (only estimating variance components), then add the socio-demographic variables, and finally add the interaction of the socio-demographic variables with tel-to-web (as opposed to tel-to-tel) to test whether the mode effects vary by the socio-demographic variables. [Table 3](#) presents the results.

We interpret for each of the dependent variables (subs-to-DK, subs-to-REF, INR-to-any) only the best-fitting model, that is, the model with the lowest DIC statistics. This is for subs-to-DK and for subs-to-REF the model which includes the interaction variables (c), and for INR-to-any the model which includes the socio-demographic variables only (b).

Web in the second wave increases the likelihood to provide a REF or a DK rather than again giving a substantive answer compared to telephone. For both telephone and web, we find that

Table 3. Beta-coefficients from cross-classified logit models.

Variables	Sub-to-subs (ref.)	Subs-to-DK	Subs-to-REF	INR-to-any
(a) base models (variance components)				
Constant		-6.326**	-5.875**	-6.166**
Variance (Respondent level)		1 (fixed)	1 (fixed)	1 (fixed)
Variance (Question level)		2.641**	1.065**	3.730**
DIC		2250	1898	2920
(b) main effects				
Web		1.165**	2.636**	-0.253
Registered landline		0.0728	0.0781	0.225
Male		-0.849**	0.436*	-0.524**
Foreign Nationality (Ref.: Swiss)		0.126	0.603**	0.601**
Tertiary education level		-0.615**	-0.637**	0.0573
Old (Ref.: < 55 years old)		0.0914	0.368	0.694**
Number of questions answered		-0.0329**	-0.0544**	-0.00750
Constant		-4.867**	-5.093**	-6.213**
Variance (Respondent level)		1 (fixed)	1 (fixed)	1 (fixed)
Variance (Question level)		2.828**	1.280**	3.950**
DIC		2149	1642	2860
(c) interaction effects				
Web		2.911*	2.421**	-0.903
Registered landline		-0.184	1.190	0.229
Male		-1.011**	0.489	-0.515**
Foreign Nationality (Ref.: Swiss)		0.0427	0.0570	0.625**
Tertiary education level		-0.619**	0.899**	-0.0162
Old (Ref.: < 55 years old)		0.761**	1.972**	0.755**
Number of questions answered		-0.0157	-0.112**	-0.00849
Web X Registered landline		0.619	-1.265	0.0565
Web X Male		0.353	-0.0449	-0.115
Web X Foreigner Nationality		0.227	0.523	-0.0537
Web X Tertiary education level		0.144	-2.228**	0.519
Web X Old		-1.660**	-1.936**	-0.297
Web X Number of questions		-0.0362	0.0727**	0.0125
Constant		-5.827**	-5.219**	-6.267**
Variance (Respondent level)		1 (fixed)	1 (fixed)	1 (fixed)
Variance (Question level)		2.907**	1.274**	3.955**
DIC		2130	1601	2866
N (respondent-question combinations)		40'280	40'215	40'393

Note: ** $p < 0.01$, * $p < 0.05$. Data: Swiss Household Panel. Shaded: best-fitting models (lowest DIC).

men give fewer transitions to DK and fewer INR from the onset than women, and foreigners more INR. Better educated HRP provide less DK but more REF than lower educated and older HRP more DK, REF, and initial INR than younger HRP. The number of questions is negatively associated with REF. Note that these effects may change if only main effects models (b) are considered (e.g., the effect of high education on DK changes sign).

As for the effect when transferred from telephone to the web as compared to keeping the telephone mode (interactions), no clear pattern emerges. Higher educated and older HRP who change to the web increase their REF less than lower educated and younger HRP. Older HRP are also less likely to switch to DK. This suggests that older HRP are more likely to provide a substantive answer when switched to the web. Finally, more questions lead to a higher likelihood to refuse in the tel-to-web group than in the tel-to-tel group. The comparatively high variance on the question level shows the importance to model respondent and question variance separately, using crossed models.

4. Conclusion

Based on data from the two-wave SHP IV pilot study, which incorporated a mode experiment including a control (tel-to-tel) and a treatment group (tel-to-web), we explored the impact of the mode change on INR in the household questionnaire. We find from simple descriptive statistics that compared to the telephone mode, the web survey produces more transitions from subs-to-DK (tel-to-tel: 0.4%, tel-to-web: 1.1%) and considerably more transitions from subs-to-REF (tel-to-tel: 0.1%, tel-to-web: 1.4%). This is in line with most findings from cross-sectional studies (e.g., [De Leeuw and Hox 2018](#)). While cross-sectional studies may suffer from selection effects, our study using longitudinal study clearly shows that this increase is a mode effect.

Furthermore, in addition to recent research that uses cross-sectional data ([Lipps and Monsch 2022](#)) showing that survey modes affect both DK and REF to different extents, we now show that this holds in a longitudinal context: people with different socio-demographic characteristics have a different tendency to switch to DK or REF when transferred to the web in the second wave. However, there is no clear pattern emerging.

While offering DK and REF may drive up INR, it produces DK answers that are likely to reflect not knowing the answer to the question more accurately, rather than unwillingness to answer. As an additional exploratory examination of what drives DK answers, we analyzed the median time needed for the HRP to reply to one question (minimum monthly income needed to make ends meet). Respondents who gave a substantive answer in both waves took slightly less time to answer this question in the second wave compared to the first (33 to 29 seconds in the tel-to-tel group, 34 to 30 seconds in the tel-to-web group), suggesting a learning effect. To the contrary, a DK was given much faster in the second wave compared to the time it took them to give a substantive answer in the first wave (50 to 36 seconds in the tel-to-tel group, 36 to 14 seconds in the tel-to-web group). This suggests that a DK does not result from a long search process, especially in the web mode. This points to the need to probe the respondent after giving a DK or using unfolding brackets although this may lengthen the interview duration considerably ([De Leeuw et al. 2016](#)).

Being able to isolate DK from REF may be of substantive interest for researchers analyzing the data. Reasons for more REF in web surveys may be due to less motivation and guidance but probably also the tendency to give a more polite answer to an interviewer. It may be that both DK and REF decrease once the web is an even more established and accepted survey mode, people are more familiar with it, or more motivation, interaction, or entertainment is provided (e.g., [Keusch 2020](#)). We find support for this idea by the much lower likelihood for higher educated HRPs to give a REF. Higher educated respondents may be less burdened by a web survey than less well-educated people. Contrary to our expectations, we do not find less educated respondents to switch more from a substantive answer to a DK in the web. This shows that our web version produces more DK because this category is explicitly offered and that respondents with different education levels do not selectively give or hide a DK according to the mode.

This study has some limitations. First, the sample size of the telephone-to-web group is small. Second, the rate of INR is rather small, as questions from the household questionnaire are most often factual and not very prone to INR. Third, there are probably still small

selective attrition effects such that HRP who are less internet affine in the telephone-to-web group are more likely to attrite after the first (telephone) wave. We identified an association between (older) age and attrition (Voorpostel et al. 2020) but there are most likely other characteristics that remained unobserved. A higher drop-out of older respondents in the telephone-to-web group with probably more INR rates in the second wave was likely to decrease INR differences between telephone and web in the second wave due to attrition. Finally, since we cannot observe the counterfactual, we do not know whether our findings are an effect of changing modes or an effect of the use of web. These limitations withstanding, we believe that our design is well-suited to analyze effects on INR when changing mode in a panel survey. We show that switching to the web is possible without risking a severe increase in DK or REF. Future research could use larger samples in the web mode and different (more complex, more sensitive, more non-factual) questions with higher likelihoods of DK or REF already in an interviewer-administered mode. This will provide more insights into the trade-off between more REF in the web mode and more substantive but untrue answers in interviewer-administered modes.

5. References

- AAPOR (The American Association for Public Opinion Research). 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, (9th edition). Illinois: AAPOR. Available at: http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed November 2022).
- Beatty, P., and D. Herrmann. 2002. "To Answer or not to Answer: Decision Processes Related to Survey Item Non-response." In *Survey Nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and J. Little. (pp. 71–85). New York: Wiley.
- Bowling, A. 2005. "Mode of Questionnaire Administration can have Serious Effects on Data Quality." *Journal of Public Health* 27: 281–291. DOI: <https://doi.org/10.1093/pubmed/fdi031>.
- Browne, W.J. 2019. "MCMC Estimation in MLwiN, v3.03." Centre for Multilevel Modelling, University of Bristol. Available at: <http://www.bristol.ac.uk/cmm/media/software/mlwin/downloads/manuals/3-03/mcmc-web.pdf> (accessed November 2, 2022)
- Chang, L., and J.A. Krosnick. 2010. "Comparing Oral Interviewing with Self-Administered Computerized Questionnaires. An Experiment." *Public Opinion Quarterly* 74(1): 154–167. DOI: <https://doi.org/10.1093/poq/nfp090>.
- Christian, L.M., D.A. Dillman, and J.D. Smyth. 2007. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, M. Brick, E.D. de Leeuw, L. Japac, P.J. Lavrakas, M.W. Link, and R.L. Sangster: 250–275. Hoboken, Wiley.
- De Leeuw, E.D. 2005. "To Mix or not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21(2): 233–255. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/to-mix-or-not-to-mix-data-collection-modes-in-surveys.pdf>.
- De Leeuw, E.D., and J. Hox. 2018. "Internet Surveys as Part of a Mixed-Mode Design." In *Social and behavioral research and the Internet: Advances in applied methods and*

- research strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. 45–76. New York: Routledge.
- De Leeuw, E.D., J. Hox, and A. Boeve. 2016. “Handling do-not-know Answers: Exploring new Approaches in Online and Mixed-mode Surveys.” *Social Science Computer Review* 34(1): 116–132. DOI: [10.1177/0894439315573744](https://doi.org/10.1177/0894439315573744).
- Fielding, A., and H. Goldstein. 2006. “Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review,” University of Birmingham. Available at: <http://www.bristol.ac.uk/cmm/team/hg/full-publications/2006/cross-classified-models-review-for-dfes.pdf> (accessed November 2022)
- Fricker, S., M. Galesic, R. Tourangeau, and T. Yan. 2005. “An Experimental Comparison of Web and Telephone Surveys.” *Public Opinion Quarterly* 69(3): 370–392. DOI: <https://doi.org/10.1093/poq/nfi027>.
- Gooch, A., and L. Vavreck. 2019. “How Face-to-Face Interviews and Cognitive Skill Affect Item Non-response: A Randomized Experiment Assigning Mode of Interview.” *Political Science Research and Methods* 7(1): 143–162. DOI: <https://doi.org/10.1017/psrm.2016.20>.
- Greene, J., H. Speizer, and W. Wiitala. 2008. “Telephone and Web: Mixed-Mode Challenge.” *Health services research* 43(1): 230–248. DOI: [10.1111/j.1475-6773.2007.00747.x](https://doi.org/10.1111/j.1475-6773.2007.00747.x).
- Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2011. *Survey Methodology*. Hoboken: Wiley.
- Guzy, N., and H. Leitgöb. 2015. “Assessing Mode Effects in Online and Telephone Victimization Surveys.” *International Review of Victimology* 21(1): 101–131. DOI: <https://doi.org/10.1177/0269758014547995>.
- Juster, F.T., and J.P. Smith. 1997. “Improving the Quality of Economic Data: Lessons from the HRS and AHEAD.” *Journal of the American Statistical Association* 92(440): 1268–1278. DOI: [10.1080/01621459.1997.10473648](https://doi.org/10.1080/01621459.1997.10473648).
- Holbrook, A., Y.I. Cho, and T. Johnson. 2006. “The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties.” *Public Opinion Quarterly* 70(4): 565–595. DOI: <https://doi.org/10.1093/poq/nfl027>.
- Keusch, F. 2020. “Gamification in Web Surveys.” In *SAGE Research Methods Foundations*, edited by P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, and R.A. Williams. 1–8. London: Sage.
- Kleiner, B., O. Lipps, and E. Ferrez. 2015. “Language Ability and Motivation Among Foreigners in Survey Responding.” *Journal of Survey Statistics and Methodology* 3(3): 339–360. DOI: <https://doi.org/10.1093/jssam/smv015>.
- Krosnick, J.A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5: 231–236. DOI: <https://doi.org/10.1002/acp.2350050305>.
- Krumpal, I. 2013. “Determinants of Social Desirability Bias in Sensitive Surveys: a Literature Review.” *Quality & Quantity* 47(4): 2025–2047. DOI: <https://doi.org/10.1007/s11135-011-9640-9>.
- Leckie, G., and C. Charlton. 2013. “runmlwin – A Program to Run the MLwiN Multilevel Modelling Software from Within Stata.” *Journal of Statistical Software* 52: 1–40. DOI: <https://doi.org/10.18637/jss.v052.i11>.

- Lipps, O., and G.-A. Monsch. 2022. "Effects of Question Characteristics on Item Non-Response in Telephone and Web Survey Modes." *Field Methods*. DOI: <https://doi.org/10.1177/1525822X221115838>.
- Roberts, C., and C. Vandenplas. 2017. "Estimating Components of Mean Squared Error to Evaluate the Benefits of Mixing Data Collection Modes." *Journal of Official Statistics* 33(2): 303–334. DOI: <http://dx.doi.org/10.1515/JOS-2017-0016>.
- Shoemaker, P.J., M. Eichholz, and E.A. Skewes. 2002. "Item Nonresponse: Distinguishing Between Don't know and Refuse." *International Journal of Public Opinion Research* 14(2): 193–201. DOI: <https://doi.org/10.1093/ijpor/14.2.193>.
- Silber, H., J. Roßmann, T. Gummer, S. Zins, and K.W. Weyandt. 2021. "The Effects of Question, Respondent and Interviewer Characteristics on Two Types of Item Nonresponse." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. DOI: <https://doi.org/10.1111/rssa.12703>.
- Sun, H., R. Tourangeau, and S. Presser. 2019. "Panel Effects: Do the Reports of Panel Respondents Get Better or Worse over Time?" *Journal of Survey Statistics and Methodology* 7(4): 572–588. DOI: <https://doi.org/10.1093/jssam/smy021>.
- Sturgis, P., C. Roberts, and P. Smith. 2014. "Middle Alternatives Revisited: How the Neither/nor Response Acts as a Way of Saying "I don't know." *Sociological Methods & Research* 43(1): 15–38. DOI: <https://doi.org/10.1177/0049124112452527>.
- Tillmann, R., M. Voorpostel, E. Antal, N. Dasoki, H. Klaas, U. Kuhn, F. Lebert, G.-A. Monsch, and V.-A. Ryser. 2021. "The Swiss Household Panel (SHP)." *Jahrbücher für Nationalökonomie und Statistik* . DOI: <https://doi.org/10.1515/jbnst-2021-0039>.
- Tourangeau, R., and K.A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological bulletin* 103(3): 299–314. DOI: <https://doi.org/10.1037/0033-2909.103.3.299>.
- Vannieuwenhuyze, J.T., and G. Loosveldt. 2013. "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects." *Sociological Methods & Research* 42(1): 82–104. DOI: <https://doi.org/10.1177/0049124112464868>.
- Voorpostel, M., U. Kuhn, R. Tillmann, G.-A. Monsch, E. Antal, V.-A. Ryser, F. Lebert, H.S. Klaas, and N. Dasoki. 2020. *Introducing Web in a Refreshment Sample of the Swiss Household Panel: Main Findings From a Pilot Study*. FORS Working Paper Series, paper 2020-2. Lausanne: FORS. Available at: https://serval.unil.ch/resource/serval:BIB_7C300C68F2D1.P001/REF (accessed November 2022)
- Voorpostel, M., O. Lipps, and C. Roberts. 2021. "Mixing Modes in Household Panel Surveys: Recent Developments and new Findings." In *Advances in longitudinal survey methodology*, edited by P. Lynn. 204–226. Hoboken: Wiley. DOI: <https://doi.org/10.1002/9781119376965.ch9>.
- Young, R. 2012. *Don't Know Responses in Survey Research*. Unpublished dissertation. Department of Sociology: Pennsylvania State University. Available at: https://etda.libraries.psu.edu/files/final_submissions/7219 (accessed November 2022)

Received July 2021

Revised February 2022

Accepted November 2022

Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach

Daniela Marella¹

Large amount of data are today available, that are easier and faster to collect than survey data, bringing new challenges. One of them is the nonprobability nature of these big data that may not represent the target population properly and hence result in highly biased estimators. In this article two approaches for dealing with selection bias when the selection process is nonignorable are discussed. The first one, based on the empirical likelihood, does not require parametric specification of the population model but the probability of being in the nonprobability sample needed to be modeled. Auxiliary information known for the population or estimable from a probability sample can be incorporated as calibration constraints, thus enhancing the precision of the estimators. The second one is a mixed approach based on mass imputation and propensity score adjustment requiring that the big data membership is known throughout a probability sample. Finally, two simulation experiments and an application to income data are performed to evaluate the performance of the proposed estimators in terms of robustness and efficiency.

Key words: Big data; informative sample; mass imputation.

1. Introduction

The main characteristic of big data sources is that they provide us with detailed information often in real time, since they are generated in an automated way using information technology systems or sensors. This results in massive datasets of very large volume and in a huge variety of forms of data. For an overview on big data, see [Beresevicz et al. \(2018\)](#). Large amount of data are therefore available, that are easier and faster to collect than the standard data sources as census and survey data, bringing new opportunities and challenges, see [Pfeffermann \(2015\)](#).

However, if on one hand big data represent potentially new data sources, on the other we need to know how much they can help the inferential process and under which assumptions. From the statistical inference point of view, what really matters is the way these data are generated. Big data sources are nonprobability samples, which often fail to represent the target population properly then the analysis results may be subject to selection biases, see [Elliott and Valliant \(2017\)](#) and [Meng \(2018\)](#). In this article, this concern is addressed.

Let A and B be two data sources, where B is a nonprobability sample while A is an independent probability sample. We assume that (\mathbf{x}, y) are available from B while \mathbf{x} is available from survey data A , where \mathbf{x} is a vector of p auxiliary variables and y is the

¹University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome, 00185, Italy. Email: daniela.marella@uniroma1.it

Acknowledgment: The author is grateful to the referees for very careful reading of the manuscript and thoughtful comments.

variable of interest. Generally speaking, there are three possible methods to draw reliable statistical inference from nonprobability samples. The first method is the so called propensity score adjustment, see [Rosenbaum and Rubin \(1983\)](#). In this approach the unknown probability of selection for the units in B is estimated from sample A (propensity or sampling score) by the covariates \mathbf{x} . The second approach is based on calibration. That is, information on the auxiliary variables in sample B is calibrated with that in the population or at least can be estimated from the probability sample A , see [Kott \(2006\)](#) and [DiSogra et al. \(2011\)](#). The third approach is mass imputation where imputed values of y are created for all units in the probability sample A . Then, an estimator of the parameter of interest based on imputed data is computed. Survey data integration for combining a probability sample with a nonprobability sample is also discussed in [Yang et al. \(2021a\)](#) where a formal framework for mass imputation is developed and asymptotic results for the k nearest neighbor estimator are established. The nearest neighbor imputation estimator of [Rivers \(2007\)](#) is also covered as a special case. Finally, [Yang et al. \(2021b\)](#) propose a doubly robust estimator of the finite population mean using the estimated propensity scores as well as an outcome linear regression model. The double robustness entails that the final estimator is consistent for the true value if either the probability of selection into the nonprobability sample or the outcome model is correctly specified, not necessarily both.

All the aforementioned methods assume that the selection mechanism for sample B is ignorable after controlling on \mathbf{x} . Since selection mechanism and nonresponse are closely related, it is essentially the missing at random (MAR) assumption of [Rubin \(1976\)](#). However, the MAR assumption is not always realistic because survey participation may be related to the survey topic of interest. For instance, we might expect that the selection process (self-selection) to be nonignorable on Twitter data, since the propensity to tweet (sample inclusion probability) might depend on the particular subject, which will often be related to the target variable. When the inclusion probabilities are related to the value of the target outcome variable even after conditioning on the model covariates, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population. This allows the possibility that being in the sample or analogously being a respondent depends in some stochastic way on the variable of interest y . It is essentially the not missing at random (NMAR) assumption of [Rubin \(1976\)](#).

If MAR assumption does not hold, then we can build a NMAR model for the selection mechanism and estimate the model parameters, see [Chang and Kott \(2008\)](#) and [Riddles et al. \(2016\)](#). Existing approaches for parameter estimation for a propensity score model under nonignorable nonresponse can be classified as fully parametric approaches or method of moments approaches. A fully parametric approach, which makes parametric assumptions about the population distribution of the study variable, is considered in [Beaumont \(2000\)](#). Also, the [Heckman \(1979\)](#) selection model approach is a fully parametric approach in the sense that the outcome regression model and the response model are linked by a joint normal distribution on the error terms of the two models. In [Galimard et al. \(2018\)](#) an imputation model for missing binary data with NMAR mechanism from Heckman's model using a onestep maximum likelihood estimator is derived. These fully parametric approaches can be used to estimate the parameters in the

response model, but the estimates can be very sensitive to failure of the assumed model. The method of moments approach does not directly use the outcome model while the response model is assumed to be specified. In [Chang and Kott \(2008\)](#) and [Kott and Chang \(2010\)](#) propensity score weighting for nonignorable missing mechanism is introduced together with instrumental variable calibration. The authors extended the notion of calibration weighting by allowing the number of explanatory variables in the assumed response model to be less than the number of calibrations variables. Instead of the fully parametric or the calibration approach, [Riddles et al. \(2016\)](#) consider an alternative modeling approach that uses parametric model assumptions about the study variable among the respondents only. Such a modeling approach has been considered in [Pfeffermann and Sikov \(2011\)](#).

Evidently, accounting for nonignorable selection mechanism is a major undertaking and the present article attempts to address this challenge. In this article two approaches for dealing with selection bias when the selection process is nonignorable are discussed. The first one, based on the empirical likelihood, does not require parametric specification of the population model but the probability of being in the nonprobability sample needed to be modeled. An important advantage of this approach is that it facilitates the use of calibration constraints that can help to correct for selection bias in nonprobability samples. That is, auxiliary information known for the population or estimable from the probability sample A can be incorporated as calibration constraints, thus enhancing the precision of the estimators. The success of the proposed approach depends on proper modeling of the unknown selection probabilities. However, the resulting sample model can be tested based on the observations in nonprobability sample by standard test statistics. Then, model diagnostics are more feasible and the method is less sensitive to failure of the assumed selection model. The approach relies on work by [Feder and Pfeffermann \(2019\)](#) for dealing with problems such as observational studies, informative sampling and nonignorable nonresponse. Such an approach has also been proposed to deal with the statistical matching problem under nonignorable sampling and nonresponse in [Marella and Pfeffermann \(2021\)](#).

The second one is a mixed approach based on mass imputation and propensity score adjustment. We consider the case when additionally the membership to the nonprobability sample B can be determined throughout the probability sample A , as in [Yang et al. \(2021a\)](#). First of all, imputed values \tilde{y} are created for all units in A and the selection probabilities for units in B are estimated from A by (\mathbf{x}, \tilde{y}) . Next, the sample empirical likelihood is maximized by a two steps estimation procedure.

The article is organized as follows. In Section 2 the basic setup in a fully parametric context is briefly introduced. In Section 3 a semiparametric approach based on the empirical likelihood (EL) is discussed and its performance is evaluated by a simulation study in Sections 5 and 6. The failure in proper modeling the unknown selection probabilities is investigated in Subsection 6.2.1. The robustness of the approach to violations of the population normality assumption is evaluated in Subsection 6.2.2, where skewed and binary data are considered. Furthermore, an application to income data is presented in Section 7. In Section 4 the mixed approach is described. Its performance is assessed by a simulation study in Section 8. Section 9 draws final conclusions.

2. Adjusting for Selection Bias: Basic Setup

The main challenges in using nonprobability samples are under-coverage and self selection. In the sequel we assume that the target population is fully covered, then the inclusion probabilities are nonzero for all the population units. Suppose that we have two independent samples A and B selected from a finite population of size N generated from a joint probability distribution function (pdf) $f(\mathbf{x}, y; \boldsymbol{\theta})$, governed by a vector parameter $\boldsymbol{\theta}$. Let y be the study variable and $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ the vector of p auxiliary variables. Let B be a nonprobability sample of size n_B , such as a voluntary sample or a self-selected sample, and A an independent probability sample. We assume that (\mathbf{x}, y) are available from B while \mathbf{x} is available from survey data A . Then, B contains rich information on (\mathbf{x}, y) but the sampling mechanism is unknown while the sample A , representing the finite population, does not observe the study variable of interest. Let δ_i be the sample inclusion indicator, that is, a Bernoulli random variable taking value $\delta_i = 1$ if population unit $i \in B$, $\delta_i = 0$ otherwise. The sampling mechanism for the nonprobability sample B is ignorable (noninformative) after controlling on \mathbf{x} if,

$$P(\delta_i = 1 | \mathbf{x}_i, y_i) = P(\delta_i = 1 | \mathbf{x}_i), \quad (1)$$

for each \mathbf{x}_i . Unfortunately, the ignorability condition is a strong assumption and it is not verifiable based on the observed data. If the sampling mechanism for sample B is not ignorable, the inclusion probabilities are related to the value of the target outcome variable y even after conditioning on the model covariates \mathbf{x} , then the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population, see [Pfeffermann and Sverchkov \(2009\)](#) and [Pfeffermann \(2011\)](#) for discussion of the notion of informative sampling. This is equivalent to assume that the sample B is subject to not missing at random (NMAR) nonresponse, by which the response probabilities depend in some stochastic way on the study variable of interest.

In this section an approach of reducing the selection bias associated with the nonprobability sample B is briefly illustrated in a parametric context. In Section 3 the use of the EL is proposed. From [Pfeffermann et al. \(1998\)](#), the marginal sample pdf of (\mathbf{x}_i, y_i) for $i \in B$ is defined as,

$$f_B(\mathbf{x}_i, y_i; \boldsymbol{\theta}, \boldsymbol{\gamma}_B) = \frac{P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)}{P(\delta_i = 1; \boldsymbol{\theta}, \boldsymbol{\gamma}_B)} f_p(\mathbf{x}_i, y_i; \boldsymbol{\theta}), \quad (2)$$

where $f_p(\mathbf{x}_i, y_i; \boldsymbol{\theta})$ is the population pdf governed by $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}_B$ represents any additional parameters defining the sample distribution, resulting from the sampling process. Under independence between observations corresponding to different sampling units, the sample likelihood can be approximated by the product of the sample pdfs over the corresponding sample observations. Hence, the sample likelihood is,

$$L_B(\boldsymbol{\theta}, \boldsymbol{\gamma}_B) = \prod_{i=1}^{n_B} f_B(\mathbf{x}_i, y_i; \boldsymbol{\theta}, \boldsymbol{\gamma}_B). \quad (3)$$

The probabilities $P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)$ appearing in the sample pdf (2) needed to be modeled. To this aim, a parametric model indexed by the unknown parameter $\boldsymbol{\gamma}_B = (\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_y)'$

of length $p + 1$ can be assumed, which is allowed to depend on the observed data (the outcome and auxiliary variables). Formally,

$$P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B) = g(\boldsymbol{\gamma}'_x \mathbf{x}_i + \gamma_y y_i) \quad (4)$$

for some known function g , taking values in the range $[0, 1]$.

Remark 1 The sample inclusion probabilities in B may depend on several unobserved variables and yet, by definition of the sample pdf, one only needs to model the probabilities $P(\delta_i = 1 | \mathbf{x}_i, y_i)$. As discussed and illustrated in subsequent sections, the resulting sample model can be tested based on the observations in B .

Modeling the probabilities by the logistic or probit functions is common, but notice that in our case the probabilities depend also on the study variable y . Then, the two models, the population model $f_p(\mathbf{x}_i, y_i; \boldsymbol{\theta})$ and the parametric model (4), define the model holding for the observed units in B . Notice that, the sample likelihood in Equation (3) only depends on the observed data in sample B . Furthermore, it needs to be maximized with respect to the population and selection model parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma}_B)$. Thus, the unknown sampling parameters $\boldsymbol{\gamma}_B$ are estimated jointly from the likelihood.

Remark 2 If the main target of inference is the mean of y (μ_y), after having estimated $\boldsymbol{\theta}$, the following estimators can be computed,

$$\hat{\mu}_y = E_p(y_i; \hat{\boldsymbol{\theta}}), \quad \hat{\mu}_{y,H} = \frac{\sum_{i \in B} y_i / \hat{P}(\delta_i = 1 | \mathbf{x}_i, y_i)}{\sum_{i \in B} 1 / \hat{P}(\delta_i = 1 | \mathbf{x}_i, y_i)}. \quad (5)$$

where $\hat{\mu}_{y,H}$ is the Hájek estimator, see Hájek (1964). Large differences between the two estimators may indicate misspecification of either the population model or the parametric model (4). Notice that both the estimators take into account the informative sampling design in B since $\hat{P}(\delta_i = 1 | \mathbf{x}_i, y_i)$ instead of the propensity scores $\hat{P}(\delta_i = 1 | \mathbf{x}_i)$ are used.

However, the maximization of sample likelihood in Equation (3) with respect to $(\boldsymbol{\theta}, \boldsymbol{\gamma}_B)$ can be complicated numerically and result in unstable estimates, depending on the population model and the model assumed for the selection probabilities. One may also face identifiability or practical identifiability problems, see Pfeffermann and Landsman (2011) and Lee and Berger (2001). For this reason, we propose in the next section the use of the empirical likelihood approach.

3. Adjusting for Selection Bias: A Semiparametric Approach

The approach described in Section 2 is fully parametric, since it makes parametric assumptions about both the population distribution of the study variable and the selection mechanism. In this section we propose a semiparametric approach based on the use of the EL which enables estimating the parameter $\boldsymbol{\gamma}_B$, governing the sampling process, without specifying the population model. The EL combines the robustness of nonparametric methods with the efficiency of the likelihood approach, see Owen (2001, 2013) and references therein. The EL is essentially the likelihood of the multinomial distribution, where the parameters are the point masses assigned to the distinct sample values. An important advantage of the empirical likelihood approach is that it facilitates the use of

calibration constraints. That is, auxiliary information on known population means for some auxiliary variables can be incorporated by placing additional constraints on the maximization process. See [Chaudhuri et al. \(2010\)](#) for details of the constrained estimation procedure and the asymptotic properties of the resulting empirical likelihood estimators. Last, but not least, not requiring to specify the population model the approach is more robust and often easier to implement.

The basic idea of the empirical likelihood approach is to approximate the population distribution by a multinomial model with probabilities $p_i^{xy} = Pr(\mathbf{x}_i, y_i)$, which support is given by the empirical observations $\{(\mathbf{x}_i, y_i), i = 1, \dots, n_B\}$. This means that a multinomial probability is assigned just to the observed values in sample B . Notice that, the statement regarding the support is a basic assumption underlying the EL approach which can be justified by having sufficiently large sample. Then, the sample distribution in B is,

$$p_{i,B}^{xy} = \frac{P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)}{P(\delta_i = 1; \{p_i^{xy}\}, \boldsymbol{\gamma}_B)} p_i^{xy}, \quad (6)$$

where $P(\delta_i = 1; \{p_i^{xy}\}, \boldsymbol{\gamma}_B) = \sum_{i \in B} P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B) p_i^{xy}$. The sample EL based on B is thus,

$$EL_B(\{p_i^{xy}\}, \boldsymbol{\gamma}_B) = \prod_{i \in B} p_{i,B}^{xy} = \prod_{i \in B} \frac{P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)}{\sum_{i \in B} P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B) p_i^{xy}} p_i^{xy}. \quad (7)$$

Then, the semiparametric approach defines the sample EL and combines it with a parametric model for the probabilities $P(\delta_i = 1 | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)$, as specified in Equation (4).

As previously stressed, an important advantage of the EL approach is that it facilitates the use of calibration constraints that can help to correct for selection bias in nonprobability samples. Specifically, known population means of auxiliary variables related to the study variable and measured for the nonprobability sample B can be incorporated by placing additional constraints (calibration constraints) on the maximization process. For instance, in the simulation study of Section 5 the constraint,

$$\sum_{i \in B} \mathbf{x}_i \sum_{\{i \in B: \mathbf{X} = \mathbf{X}_i\}} p_i^{xy} = \boldsymbol{\mu}_x \quad (8)$$

is considered, where the population mean $\boldsymbol{\mu}_x$ of \mathbf{x} is assumed known. Then, the likelihood (7) must be maximized with respect to $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$ under the constraints,

$$p_i^{xy} \geq 0, \sum_{i \in B} p_i^{xy} = 1, \quad (9)$$

for all i , and the calibration constraint (8). One only needs the estimates of the multinomial population model parameters $\{p_i^{xy}\}$ and thus, we may consider $\boldsymbol{\gamma}_B$, as nuisance parameter. In order to write the likelihood in Equation (7) as only a function of the unknown probabilities $\{p_i^{xy}\}$, we adopt the profile likelihood approach. We use some initial estimates for the set of probabilities $\{p_i^{xy}\}$ and we solve the constrained maximization problem by first computing the profile likelihood of $\boldsymbol{\gamma}_B$ and then maximizing the profile likelihood over $\boldsymbol{\gamma}_B$. For a given $\boldsymbol{\gamma}_B$, we then maximize the resulting likelihood under the constraints in Equations (8) and (9) with respect to the unknown probabilities $\{\hat{p}_i^{xy}\}$,

yielding $\{p_i^{xy}\}$. This completes the first iteration in the estimation process. In the second iteration, we consider the estimates $\{\hat{p}_i^{xy}\}$ as known, re-estimate the parameters γ_B , and then the unknown probabilities $\{p_i^{xy}\}$. The iterations continue until convergence. See [Feder and Pfeffermann \(2019\)](#) for conditions guaranteeing the convergence of the maximization process. There are cases where a solution does not exist. An example is where all the observed values of a constraining variable are greater (or smaller) than its known population mean. Furthermore, a combination of multivariate constraints can also preclude a solution. For instance, when the sum of two variables used in the constraints is greater for all the observed units than the sum of the corresponding population means.

Remark 3 The simulation study has been carried out by using the software R, [R Core Team \(2021\)](#). The maximization with respect to γ_B can be performed by using the R numerical optimization function `optim`. For a given γ_B , the maximization with respect to $\{p_i^{xy}\}$ can be carried out by using the function `emplik` in the R package `mev`, see [Belzile et al. \(2022\)](#). See [Owen \(2013\)](#) for related theory and further details.

Notice that, inference on the unknown model parameters is based on the sample EL which requires that the corresponding sample model is identifiable. The sample model is not identifiable if there is more than one combination of a population model and a sampling mechanism yielding the same sample model. See [Pfeffermann and Landsman \(2011\)](#) and references therein for conditions guaranteeing the identifiability of the sample model. Notice that, for a given parameter γ_B and without any constraints the EL is not identifiable. In the proposed approach the empirical likelihood is maximized under a set of calibration constraints. Then, the main question is how the survey variables defining the constraints should be chosen. As in [Chang and Kott \(2008\)](#), such variables should be correlated as highly as possible with y and \mathbf{x} because otherwise they provide little or no information on the probabilities $P(\delta_i = 1 | \mathbf{x}_i, y)$.

Remark 4 If μ_x is unknown but a probability sample A is available then the auxiliary information in sample B can be calibrated with that in sample A . Then, in Equation (8) the mean vector μ_x can be replaced by its Horvitz-Thompson estimator computed from sample A . Formally,

$$\sum_{i \in B} \mathbf{x}_i \sum_{\{i \in B: \mathbf{X} = \mathbf{x}_i\}} p_i^{xy} = \frac{1}{N} \sum_{j \in A} d_j \mathbf{x}_j, \tag{10}$$

where d_j is the sampling weight associated to the j th unit in sample A . Constraint (10) is used in the application to income data of Section 7.

The success of the proposed approach depends on proper modeling of the unknown selection probabilities for sample B , that is the estimates can be sensitive to failure of the assumed model. However, once the parameters ($\{p_i^{xy}\}, \gamma_B$) have been estimated, the null hypothesis that the sample model fits the sample data can be tested successfully by classical test statistics, because the sample model refers to the observed data. An overview of the plausible test statistics that can be used for assessing the goodness of fit of the sample pdf is in [Pfeffermann \(2011\)](#). For instance, in the simulation study of Section 5 the Kolmogorov-Smirnov test has been used to compare the theoretical and the empirical sample pdfs of y . The asymptotic distribution of test statistic and correct critical values have been obtained by use of parametric bootstrap, as established theoretically by [Babu and Rao \(2004\)](#).

Finally, the bias and the standard deviation of the population parameters estimates can be obtained by resampling method. Formally, once the estimated model has been validated M bootstrap samples can be selected from it and for each bootstrap sample the unknown parameters can be estimated according the proposed approach. Then, bootstrap estimates of bias and standard deviation can be computed.

4. A Mixed Approach Based on Mass Imputation and Propensity Score Adjustment

In this section a data integration approach for combining the nonprobability sample B with an independent probability sample A is described. It is a mixed approach based on mass imputation and propensity score adjustment requiring that we can observe δ_i , the B sample inclusion indicator, from the probability sample A . That is, among the elements in the sample A , it is possible to obtain the membership information from the nonprobability sample B , as in [Kim and Wang \(2019\)](#). As stressed in [Yang et al. \(2021a\)](#), the key insight is that the subsample of units in probability sample A with the membership information ($\delta_i = 1$) constitutes a second phase sample from B , which acts as a new population. Clearly, this condition is more plausible in the big data context where the nonprobability sample B is so large that any probability sample A is bound to overlap with it.

As previously stressed, unlike the usual imputation for missing data analysis, in mass imputation imputed values for all units in the probability sample A are created. The mass imputation methods and their statistical properties are discussed in [Yang et al. \(2021a\)](#). The nearest neighbor imputation estimator of [Rivers \(2007\)](#) is also covered as a special case. The parameter of interest is μ_y . The proposed approach can be described by the following steps:

- Step 1. Create imputed values \tilde{y}_i for all units $i \in A$ by nearest neighbor method. The basic idea is to find the nearest neighbor in sample B to create an imputed value of y for each unit in sample A . Formally, the unit $k \in B$ closest to unit $i \in A$ is determined by the Euclidean distance based on the auxiliary variables \mathbf{x} and the corresponding y value from this unit is used as the imputed value.
- Step 2. Regress the membership indicator δ against (\mathbf{x}, \tilde{y}) in sample A , estimate the selection probabilities for all units in B and compute their inverse. Let \tilde{w}_i be the estimated sample weight (pseudo-weight) for the i th unit in B , for $i = 1, \dots, n_B$.
- Step 3. Estimate the parameter μ_y by
 - (3.1) the Horvitz-Thompson estimator

$$\mu_W = \frac{1}{N} \sum_{i \in B} \tilde{w}_i y_i. \quad (11)$$

- (3.2) the maximum sample EL estimator (μ_{EL}). Formally, once the sampling weights \tilde{w}_i in B are computed a two steps estimation procedure can be applied in the maximization of the EL (7). More specifically, since

$$P(\delta_i = 1 | \mathbf{x}_i, y_i; \mathcal{Y}_B) \approx \frac{1}{E_B(\tilde{w}_i | \mathbf{x}_i, y_i; \mathcal{Y}_B)}, \quad (12)$$

from Equation (6) the EL (7) becomes,

$$EL_B(\{p_i^{xy}\}, \boldsymbol{\gamma}_B) = \prod_{i=1}^{n_B} p_{i,B}^{xy} \approx \prod_{i=1}^{n_B} \frac{E_B(\tilde{w}_i; \{p_i^{xy}\}, \boldsymbol{\gamma}_B)}{E_B(\tilde{w}_i | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)} p_i^{xy}. \quad (13)$$

Then, in the first step the expectations displayed in Equation (13) are estimated from the observed data, using classical model fitting procedures. Specifically, the expectation $E_B(\tilde{w}_i | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)$ could be estimated by regressing the sampling weights \tilde{w}_i against (\mathbf{x}_i, y_i) . See, for example, [Pfeffermann and Sverchkov \(2009\)](#) and [Pfeffermann \(2011\)](#) for examples of regression models that can be used for this purpose, depending on the problem at hand. In the second step, fixing the unknown parameters $\boldsymbol{\gamma}_B$ featuring in these expectations at their estimated values allows to maximize the EL in Equation (13) only with respect to the parameter $\{p_i^{xy}\}$ indexing the population pdf, thus simplifying and stabilizing the maximization process.

The basic idea of the proposed method is to create predicted values for y in the probability sample A . In order to accomplish this in Step 1 the covariates \mathbf{x} are used, then the predictions are based on the ignorability assumption of the selection mechanism acting in the nonprobability sample B . A class of nonparametric imputation procedures based on k -nearest neighbors methods (kNN), including 1NN, is discussed in [Marella et al. \(2008\)](#), where both theoretical and simulation results are obtained. Furthermore, a nonparametric technique based on local linear regression is discussed in [Conti et al. \(2008\)](#). In Step 2 the inclusion probabilities in B are computed by applying the estimated regression of δ on (\mathbf{x}, \tilde{y}) to the observed values (\mathbf{x}, y) in B .

When auxiliary information is available it can be incorporated into the method to avoid the ignorability assumption in Step 1 and to improve the quality of the imputed values \tilde{y} in sample A . For instance, auxiliary information may refer to a set of proxy variables $\mathbf{z} \subset \mathbf{x}$ expected to behave similarly to the variable of interest. Under this circumstance better predicted values can be obtained in Step 1. Furthermore, the proxy variables \mathbf{z} , if sufficiently associated with y , can help studying the relationship between y and δ and in particular, help verifying or refuting the ignorability assumption.

In Section 8 a simulation study is employed to investigate the performance of the proposed method when the selection process is nonignorable, comparing it with other existing methods. As discussed in Section 9, new developments of the present work include the use of proxy variables.

5. Simulation Study 1

In order to evaluate the performance of the approach discussed in Sections 2 and 3 in its parametric and semiparametric form, a simulation experiment is performed. Suppose that the primary target of inference is to estimate μ_y . The simulation study consists of the following steps:

Step 1 Generate a population of $N = 1,000,000$ observations (\mathbf{x}_i, y_i) , where (x, y) has a bivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_x, \mu_y)'$ and variance covariance matrix (V-C matrix) $\boldsymbol{\Sigma}$ ($\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for short). Specifically, the marginal distribution of x is $\mathcal{N}(\mu_x, \sigma_x^2)$ with $\mu_x = 5$ and $\sigma_x^2 = 1$. The conditional distribution of y given x

is $\mathcal{N}(\mu_{y|x}, \sigma_{y|x}^2)$, with $\mu_{y|x} = \beta_0 + \beta_1 x$, $\beta_0 = \mu_y - \beta_1 \mu_x$, $\beta_1 = \sigma_{xy} / \sigma_x^2$, $\sigma_{y|x}^2 = \sigma_y^2 - \beta_1^2 \sigma_x^2$. We assume that $\beta_0 = 2$, $\beta_1 = 1$, $\sigma_{y|x} = 2$.

Step 2 Draw a sample B from the population generated in Step 1 by a Poisson sampling design with expected sample size $E(n_B) = 0.2N$ and sample inclusion probabilities given by,

$$E_p(\pi_i | x_i, y_i; \boldsymbol{\gamma}_B) = \kappa \exp\{\gamma_x x_i + \gamma_y y_i\}, \quad (14)$$

where $\boldsymbol{\gamma}_B = (\gamma_x, \gamma_y)'$ is the sampling model parameter and κ guarantees that the expectation is less or equal to one. We use different sampling model parameters $\boldsymbol{\gamma}_B$, so as to distinguish between informative and noninformative samples. From [Marella and Pfeffermann \(2019\)](#) the joint sample pdf $f_B(x_i, y_i)$ is $\mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, with mean vector $\boldsymbol{\mu}_B = (\mu_x + (\gamma_x + \beta_1 \gamma_y) \sigma_x^2, \mu_y + \beta_1 \gamma_x \sigma_x^2 + \gamma_y \sigma_y^2)'$ and V-C matrix $\boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}$, that is the sample V-C matrix is the same as for the population distribution. Then, the sample model and the population model are in the same family and only differ in the mean ($\boldsymbol{\mu}_B \neq \boldsymbol{\mu}$).

Step 3 For computational reasons, as in [Kim and Wang \(2019\)](#), we generate 500 samples S of size $n_S = 2,000$ from sample B drawn in Step 2 by a simple random sampling (*srs*). In *srs* the selection probabilities are equal for all units in sample B and the sample S can also be regarded as a set of independent and identically distributed observations from the sample model $f_B(x_i, y_i)$.

The population model parameters are estimated by parametric and semiparametric approach.

Parametric approach. For each sample S drawn in Step 3, the population model parameters $(\mu_x, \sigma_x, \beta_0, \beta_1, \sigma_{y|x}, \mu_y)$ are estimated under the following scenarios:

Scenario 1 The sample B and then each sample S are simply treated as simple random samples (*srs*). The estimates of the population parameters are denoted by $\{\hat{\mu}_{x,I}, \hat{\sigma}_{x,I}, \hat{\beta}_{0,I}, \hat{\beta}_{1,I}, \hat{\sigma}_{y|x,I}, \hat{\mu}_{x,I}\}$, where I means that the selection mechanism acting in B is ignored.

Scenario 2 The sample likelihood in Equation (3) is maximized with respect to the population parameters and the sampling parameters $\boldsymbol{\gamma}_B$. The estimates of the population parameters are denoted by $\{\hat{\mu}_{x,P}, \hat{\sigma}_{x,P}, \hat{\beta}_{0,P}, \hat{\beta}_{1,P}, \hat{\sigma}_{y|x,P}, \hat{\mu}_{x,P}\}$, where P stands for parametric approach.

Semiparametric approach. In what follows we assume knowledge of the population mean μ_x (Equation (8)). Hereafter the calibration constraint. For each sample S drawn in Step 3, the population model parameters $\{p_i^{xy}\}$ are estimated under the following scenarios:

Scenario 3 The sample B and then each sample S are simply treated as simple random samples but the knowledge of μ_x is assumed, so as to enhance the precision of the estimator for the mean μ_y . Formally, the EL under the independent and identically distributed assumption,

$$EL(\{p_i^{xy}\}) = \prod_{i \in B} p_i^{xy}, \quad (15)$$

is maximized under the constraints in Equation (9) and the calibration constraint in Equation (8). Denote by $\hat{\mu}_{y,ISP}$ the estimate of μ_y , where *ISP*

means that such an estimate is obtained under the semiparametric approach by ignoring the selection mechanism acting in B .

Scenario 4 The sample empirical likelihood in Equation (7) is maximized with respect to $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$ under the constraints in Equation (9) and the calibration constraint in Equation (8). Denote by $\hat{\mu}_{y,ISP}$ the estimate of μ_y , where SP stands for semiparametric approach.

In order to evaluate the performance of the proposed approach as the informativeness of sampling design acting in B changes, in scenarios 2 and 4 we assume that the model (14) for the inclusion probabilities $P(\delta_i = 1 | \mathbf{x}_i, y_i)$ is known and the sample EL is maximized with respect to the sampling and the population parameters $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$. The robustness of the semiparametric approach to misspecification of the selection model $P(\delta_i = 1 | \mathbf{x}_i, y_i)$ is assessed by a sensitivity analysis in Subsection 6.2.1. Finally, the robustness to violations of the population normality assumption is evaluated in Subsection 6.2.2. Notice that we generated the population values only once, so as to assess the design-based properties of the various estimation procedures.

6. Results of Simulation Study 1

In this section the simulation results obtained by the parametric approach (Subsection 6.1) and by the semiparametric approach (Subsection 6.2) are reported.

6.1. Simulation Results for the Parametric Approach

We start by studying the effect of ignoring the sampling mechanism used for drawing the sample B in the parametric approach. This is done by comparing the estimates of the population parameters under the scenarios 1 and 2, described in Section 5. In Table 1 the bias (B), the standard deviation (Sd) and the root mean square error (RMSE) of the estimates $\hat{\mu}_{y,I}$, $\hat{\mu}_{y,P}$ over the 500 samples are reported, for different $\boldsymbol{\gamma}_B$ coefficients so to distinguish between informative and noninformative samples. In Table 2 and 3 the mean and the standard deviation of the remaining parameters $(\mu_x, \sigma_x, \beta_0, \beta_1, \sigma_{y|x})$ over the 500 samples are presented. As stated previously, since $\Sigma_B = \Sigma$ it follows that $\hat{\sigma}_{x,I} = \hat{\sigma}_{x,P}$, $\hat{\beta}_{1,I} = \hat{\beta}_{1,P}$, and $\hat{\sigma}_{y|x,I} = \hat{\sigma}_{y|x,P}$, for details see Marella and Pfeffermann (2019). Then, in Table 2 and 3 just the means of $(\hat{\sigma}_{x,I}, \hat{\beta}_{1,I}, \hat{\sigma}_{y|x,I})$ and the corresponding standard deviations are reported.

As results in Table 1 show, for $\boldsymbol{\gamma}_B = (0,0)'$ the estimate $\hat{\mu}_{y,I}$ coincides with $\hat{\mu}_{y,P}$ since the sampling process acting in B is ignorable, $B(\hat{\mu}_{y,I}) = B(\hat{\mu}_{y,P}) = 0$. When $\boldsymbol{\gamma}_B \neq (0,0)'$ the sampling design is informative and the bias in $\hat{\mu}_{y,I}$ (last two rows in Table 1), coming from

Table 1. Bias (B), standard deviation (Sd) and RMSE of $\hat{\mu}_{y,I}$ and $\hat{\mu}_{y,P}$ over the 500 samples for different $\boldsymbol{\gamma}_B$ coefficients. True parameter is $\mu_y = 7$.

$\boldsymbol{\gamma}_B$	$B(\hat{\mu}_{y,I})$	$B(\hat{\mu}_{y,P})$	$Sd(\hat{\mu}_{y,I})$	$Sd(\hat{\mu}_{y,P})$	$RMSE(\hat{\mu}_{y,I})$	$RMSE(\hat{\mu}_{y,P})$
(0,0)	0.00	0.00	0.05	1.43	0.05	1.43
(0, 0.5)	2.16	-0.11	0.04	2.02	2.16	2.02
(0.25, 0.5)	2.26	-0.28	0.04	2.06	2.26	2.08

Table 2. Mean of the estimates of $(\mu_x, \sigma_x, \beta_0, \beta_1, \sigma_{y|x})$ under scenarios 1 and 2, over the 500 samples for different γ_B coefficients. True parameters are $\mu_x = 5, \sigma_x = 1, \beta_0 = 2, \beta_1 = 1, \sigma_{y|x} = 2$.

γ_B	$\bar{\hat{\mu}}_{x,I}$	$\bar{\hat{\mu}}_{x,P}$	$\bar{\hat{\sigma}}_{x,I}$	$\bar{\hat{\beta}}_{0,I}$	$\bar{\hat{\beta}}_{0,P}$	$\bar{\hat{\beta}}_{1,I}$	$\bar{\hat{\sigma}}_{y x,I}$
(0,0)	5.00	5.00	1.00	2.00	2.00	1.00	2.00
(0,0.5)	5.43	4.89	0.98	4.65	2.83	0.83	1.82
(0.25, 0.5)	5.62	4.85	0.95	5.21	3.23	0.72	1.80

Table 3. Standard deviation of the estimates of $(\mu_x, \sigma_x, \beta_0, \beta_1, \sigma_{y|x})$ under scenarios 1 and 2, over the 500 samples with different γ_B coefficients.

γ_B	$Sd(\hat{\mu}_{x,I})$	$Sd(\hat{\mu}_{x,P})$	$Sd(\hat{\sigma}_{x,I})$	$Sd(\hat{\beta}_{0,I})$	$Sd(\hat{\beta}_{0,P})$	$Sd(\hat{\beta}_{1,I})$	$Sd(\hat{\sigma}_{y x,I})$
(0,0)	0.02	0.62	0.02	0.22	1.10	0.04	0.03
(0, 0.5)	0.02	0.63	0.02	0.23	1.68	0.04	0.03
(0.25, 0.5)	0.02	0.65	0.01	0.25	1.86	0.04	0.03

the bias affecting $\hat{\beta}_{0,I}$ and $\hat{\mu}_{x,I}$ (last two rows in Table 2), increases considerably. Then, ignoring the sample selection process in sample B affects negatively the quality of the estimates of μ_y . The estimator $\hat{\mu}_{y,I}$ works poorly, even though $\hat{\mu}_{y,I}$ has the smallest standard deviation as shown in Table 1, a well known phenomenon from other studies, see Marella and Pfeffermann (2019). Furthermore, the larger is the informativeness of the sampling process the larger will be the bias in $\hat{\mu}_{y,I}$. Finally, the bias of the estimates $\hat{\mu}_{y,P}$ for $\gamma_B \neq (0, 0)'$ reduces since scenario 2 takes into account the selection mechanism acting in B . Same consideration holds for the estimates of the other population parameters, see Table 2 and 3.

Figure 1 (left) exhibits the population pdf, the sample pdf and the estimated sample pdf of y for one of the 500 samples, for the case $\gamma_B = (0.25, 0.5)'$. As can be seen, the sample pdf is very different from the population pdf, but the distribution of the estimated pdf is close to the true population distribution. Finally, with regard to the variable of interest y we test the model fitted for the sample units by Kolmogorov-Smirnov (KS) test statistic given by

$$KS_Y = \max_{y \in B} |\hat{F}_{emp}(y) - F_B(y; \hat{\theta}, \hat{\gamma}_B)|, \tag{16}$$

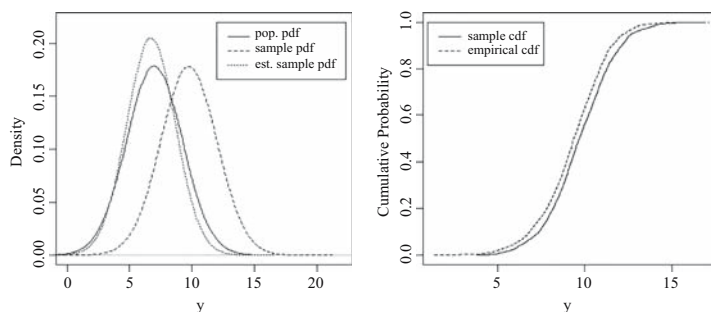


Fig. 1. Population pdf, sample pdf and estimated sample pdf of y (left); sample cdf and empirical cdf of y (right), for $\gamma_B = (0.25, 0.5)'$.

where $\hat{F}_{emp}(y) = \frac{1}{n_B} \sum_{i \in B} I(y_i \leq y)$ is the empirical cumulative distribution (cdf), $I(y_i \leq y)$ is the indicator function taking the value 1 if $y_i \leq y$ and 0 otherwise, and

$$F_B(y; \hat{\theta}, \hat{\gamma}_B) = \int_{-\infty}^y f_B(y_i; \hat{\theta}, \hat{\gamma}_B) dy_i, \tag{17}$$

is the sample cdf. The asymptotic distribution of test statistic in Equation (16) and correct critical values can be obtained by use of parametric bootstrap, as established theoretically by Babu and Rao (2004) and applied in Pfeffermann (2011). Specifically, first of all $M = 1,000$ samples are generated from the estimated sample model. Next, for each bootstrap sample the unknown parameters and the corresponding test statistic are computed. The empirical distribution of test statistic provides approximate critical values for the null distribution. In Figure 1 (right), the estimated sample cdf and the empirical cdf of y are reported. The KS statistic is 0.069 and the critical value corresponding to a significance level $\alpha = 0.05$ is 0.197. Then, the null hypothesis that the estimated model fits the sample data in B is not rejected.

6.2. Simulation Results for the Semiparametric Approach

In this section we proceed to estimate μ_y by the semiparametric approach described in Section 3. The robustness with respect to misspecification of the parametric model for $P(\delta_i = 1 | \mathbf{x}_i, y_i)$ and to violations of the population normality assumption is evaluated in Subsections 6.2.1 and 6.2.2, respectively. Table 4 shows the bias (B), the standard deviation (Sd) and the RMSE of the estimates $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$ obtained under scenarios 3 and 4, respectively.

Notice that, if sample B is treated as a simple random sample the estimates of μ_y obtained maximizing the EL in Equation (15) under the constraints in Equation (9) match the estimates $\hat{\mu}_{y,I}$ obtained under scenario 1. The conclusions of Table 4 are similar to those obtained from Table 1. When $\gamma_B = (0, 0)'$ the estimates $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$ are equal since the sampling design is not informative, $B(\hat{\mu}_{y,ISP}) = B(\hat{\mu}_{y,SP}) = 0$. When $\gamma_B \neq (0, 0)'$ the estimates $\hat{\mu}_{y,ISP}$ ignoring the sampling process show as light reduction in the bias compared to the estimates $\hat{\mu}_{y,I}$ (scenario 1) because of the introduction of the calibration constraint in Equation (8) in the EL maximization (scenario 3). As results in the Table 4 show, the estimates $\hat{\mu}_{y,SP}$ obtained by maximizing the sample EL (7) under the constraints in Equation (9) and the calibration constraint in Equation (8) (scenario 4) are characterized by lower selection bias and standard deviation illustrating the good performance of our proposed methodology. Finally, for the sample in Figure 1 the goodness of fit of the estimated model to the observed data is tested by the KS statistic in Equation (16). Its value is 0.098 and the critical value corresponding to a significance level $\alpha = 0.05$ is 0.332. Then, the null

Table 4. Bias (B), standard deviation (Sd) and RMSE of $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples with different γ_B coefficients. True parameter is $\mu_y = 7$.

γ_B	$B(\hat{\mu}_{y,ISP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,ISP})$	$Sd(\hat{\mu}_{y,SP})$	$RMSE(\hat{\mu}_{y,ISP})$	$RMSE(\hat{\mu}_{y,SP})$
(0, 0)	0.00	0.00	0.04	0.05	0.04	0.05
(0,0.5)	1.78	-0.02	0.07	0.09	1.78	0.09
(0.25,0.5)	1.74	-0.08	0.17	0.11	1.75	0.14

hypothesis that the estimated model fits the sample data is not rejected. The results in the Table 4 suggest that μ_y can be estimated almost unbiasedly and with acceptable standard error estimates when external auxiliary information is incorporated in the EL maximization, as the estimates $\hat{\mu}_{y,SP}$ and their standard deviations show.

6.2.1. Misspecification of the Selection Model

As previously stated, the EL approach does not require to specify the population pdf while the relationship between the probabilities $P(\delta_i = 1 | \mathbf{x}_i, y_i)$ and the variables (\mathbf{x}, y) is parametrically specified, see Equation (4). Hence, its performance depends on how well the assumed parametric model describes the unknown selection mechanism acting in B . In this section a sensitivity analysis is performed to assess the impact on μ_y estimate due to misspecification of the selection model. First of all, suppose that the sample B is selected by a Poisson sampling design with expected sample size $E(n_B) = 0.2N$ and unknown selection probabilities given by Equation (14) with $\boldsymbol{\gamma}_B = (0.25, 0.5)'$. Next, 500 samples S of size $n_S = 2000$ are drawn from B by a *srs* (Step 1–3, Section 5). Let us assume that the probabilities $P(\delta_i = 1 | x_i, y_i)$ are modeled by:

Model A: a linear logistic model

$$P(\delta_i = 1 | x_i, y_i) = \text{logit}^{-1}(\gamma_x x_i + \gamma_y y_i). \tag{18}$$

Model B: a quadratic logistic model. In Equation (18) x is squared and y is linear;

Model C: a quadratic logistic model. In Equation (18) both x and y are squared.

For each sample S , the estimates of μ_y under scenario 4 and models A-C are computed. Table 5 reports the bias (B) and the standard deviation (Sd) of such estimates over the 500 samples. Finally, in Table 6 the corresponding RMSEs are computed.

As results in Table 5 show, $B(\hat{\mu}_{y,SP})$ increases from 0.89 (model A) to 1.53 (model B) with a reduction in the standard deviation from 0.36 to 0.18. An additional increase is obtained under model C where $B(\hat{\mu}_{y,SP}) = 1.78$. Recall that under a correct specification of the selection model $B(\hat{\mu}_{y,SP}) = -0.08$ (see Table 4). Hence, as results in Table 5 show, the reduction of the bias in estimating μ_y depends on proper modeling the probabilities $P(\delta_i = 1 | x_i, y_i; \boldsymbol{\gamma}_B)$. The larger is the distance between the true selection model and the assumed selection model the lower will be the performance of the semiparametric

Table 5. Bias (B) and standard deviation (Sd) of $\hat{\mu}_{y,SP}$ over the 500 samples under models A-C, for $\boldsymbol{\gamma}_B = (0.25, 0.5)'$. True parameter is $\mu_y = 7$.

Model A		Model B		Model C	
$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,SP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,SP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,SP})$
0.89	0.36	1.53	0.18	1.78	0.16

Table 6. RMSE of $\hat{\mu}_{y,SP}$ over the 500 samples under models A-C, for $\boldsymbol{\gamma}_B = (0.25, 0.5)'$

RMSE ($\hat{\mu}_{y,SP}$)		
Model A	Model B	Model C
0.96	1.54	1.79

approach in removing the bias in μ_y estimator. Same consideration holds for the RMSEs in Table 6. The RMSE value of $\hat{\mu}_{y,SP}$ under model A is lower than that of comparison models, implying that the model A is better.

As previously discussed, the combined model can be tested based on the observations in sample B by standard test statistics because the sample model refers to the observed data. For instance, with regard to the sample used in Figure 1, after having modeled the probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i)$ by the logistic model in Equation (18) (model A) the goodness of fit of the estimated model is tested by the KS statistic. The KS statistic is 0.166 and the critical value corresponding to a significance level $\alpha = 0.05$ is 0.244, then the null hypothesis that the estimated model fits the sample data is not rejected. Recall that the KS statistic when the model for the selection probabilities is assumed known is 0.098 (critical value 0.332) much smaller than 0.166 (critical value 0.244) when the model (18) is assumed. The same consideration holds for model B. Finally, under model C the null hypothesis is rejected. Specifically, the KS statistic is 0.129 and the critical value is 0.112. Finally, setting the significance level $\alpha = 0.01$ both models B and C are rejected. Notice that, the relative bias under model A is 13%. A further reduction in the bias can be obtained introducing additional calibration constraints in the empirical likelihood maximization.

6.2.2. Violations of the Population Normality Assumption

In this section we employ a simulation study to assess the impact associated to violations of the normality assumption on the proposed EL approach. With this regard, two population pdfs are considered:

1. Generate a population of $N = 1,000,000$ observations (\mathbf{x}_i, y_i) , where x has a Gamma distribution with shape 3 and scale 1 and $\log(y|x)$ is normal with parameters $\theta_{y|x} = \beta_0 + \beta_1 \mathbf{x}_i$ with $\beta_0 = 0.1$, $\beta_1 = 0.2$, and $\sigma_{y|x}^2 = 0.3$. A sample B is selected by a Poisson sampling design with expected sample size $E(n_B) = 0.2N$ and sample inclusion probabilities given by Equation (14) where $\boldsymbol{\gamma}_B = (0.25, 0.5)'$.
2. As in Feder and Pfeffermann (2019), generate a population of $N = 1,000,000$ observations (\mathbf{x}_i, y_i) , where x has a Gamma distribution with parameters (2, 2). For each \mathbf{x}_i a binary outcome y_i is generated with $P(y_i = 1|x_i; \boldsymbol{\beta}) = \text{logit}^{-1}(-0.8 + 0.8\mathbf{x}_i)$ where $\boldsymbol{\beta} = (-0.8, 0.8)$. Next, a value of a design variable z is generated as $z_i = \max [(x_i + 1.1)(2y_i + 1) + v_i; 0.01]$ where v_i follows a uniform distribution $(-0.2, 0.2)$.

The sample B is drawn by a Poisson sampling with inclusion probability,

$$\pi_i = \min(200000z_i^{-1} / \sum_{j=1}^N z_j^{-1}, 0.9999). \tag{19}$$

Finally, 500 samples S of size 2,000 are drawn from B by a *srs* and the logistic model in Equation (18) is used to model the selection probability $P(\delta_i = 1|x_i, y_i)$. Results are shown in Table 7 where the bias (B), the standard deviation (Sd) and the RMSE of the estimates $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, obtained under scenarios 3 and 4, are reported.

As Table 7 shows, if the nonprobability sample B is treated as a simple random sample the estimates $\hat{\mu}_{y,ISP}$ obtained maximizing the EL in Equation (15) under the constraints in Equations (8) and (9) are biased. The bias is 1.21 (relative bias 48.4%) and -0.26 (relative bias -0.32%) for the lognormal and the binary case, respectively. A reduction in

Table 7. Bias (B), standard deviation (Sd) and RMSE of $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples for $\gamma_B = (0.25, 0.5)'$. True parameter is $\mu_y = 2.5$ for the lognormal variable and $\mu_y = 0.8$ for the binary variable.

Population	$B(\hat{\mu}_{y,ISP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,ISP})$	$Sd(\hat{\mu}_{y,SP})$	$RMSE(\hat{\mu}_{y,ISP})$	$RMSE(\hat{\mu}_{y,SP})$
Lognormal	1.21	0.58	0.09	0.05	1.21	0.58
Binary	-0.26	-0.14	0.02	0.01	0.26	0.14

the bias is obtained when the selection process in B is taken into account as the estimates $\hat{\mu}_{y,SP}$ show. More specifically, in the lognormal case the bias decreases to 0.58 (relative bias 0.23), in the binary case to -0.14 (relative bias -0.18). A further reduction in the bias can be obtained introducing additional calibration constraints in the empirical likelihood maximization.

7. An Application to Income Data

In this section the approach based on the EL is applied to real sample data. In Italy, reliable information on households income (y) is provided by the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy (Banca d'Italia) every two years. Its main goal is to study the economic status of Italian households, focusing on income and wealth. The sample for the SHIW survey is drawn in two stages, with municipalities and households as, respectively, the primary and secondary sampling units. The primary units are stratified by region and population size. Bigger municipalities (with more than 40,000 inhabitants) are all included in the sample, while the smaller towns are selected using a probability proportional to size sampling (PPS). The individual households to be interviewed are then selected by simple random sampling. In the present article we use the 2010 wave, whose sample consists of 7,951 households and 387 municipalities. The variable of interest is the household income, defined as the combined disposable annual income of all the people living in the household. The average annual household income in 2010 is $\mu_y = \text{EUR } 32,714$, as published by Bank of Italy (Banca d'Italia 2012). To reproduce the situation where a nonprobability sample B and a probability sample A are available the following procedure has been implemented:

1. A sample B is selected from SHIW according to a Poisson sampling design with expected sample size $E(n_B) = 2,000$ and unknown inclusion probabilities proportional to $(y_i - \min_i y_i + 10)$.
2. Suppose that the Household Budget Survey (HBS) run by Italian National Institute of Statistics, (ISTAT) in 2010 (sample A) which consists of 22,227 households is available. The HBS uses a sampling design similar to SHIW and collects detailed information on sociodemographic characteristics and expenditures on a disaggregated set of commodities (durable and nondurable). Let $\mathbf{x} = (x_1, x_2)$ be the available auxiliary variables, where x_1 is the household size and x_2 is the monthly expenditure on food. Furthermore, let 2.4 and 507.46 be the estimates of the average size of households and the monthly mean expenditure on food in 2010, respectively, as obtained from HBS. Then, we can add the calibration constraint in Equation (10) where,

$$\frac{1}{N} \sum_{i \in A} d_i x_{1i} = 2.4, \quad \frac{1}{N} \sum_{i \in A} d_i x_{2i} = 507.46. \tag{20}$$

Step 3 The probabilities $P(\delta_i = 1 | \mathbf{x}_i, y_i)$ are modeled by the logistic model $\text{logit}^{-1}(\boldsymbol{\gamma}_x' \mathbf{x}_i + \gamma_y y_i)$ with $\boldsymbol{\gamma}_x = (\gamma_{x1}, \gamma_{x2})'$ and the average annual household income μ_y is estimated from B by the EL under scenarios 1,3,4 described in Section 5. We recall that under scenario 1 the selection process in B is not taken into account. Scenario 3 is as scenario 1 but we add the calibration constraint in Equation (10). Under scenario 4 we maximize the EL in Equation (7) under the constraints in Equation (9) and the calibration constraint in Equation (10), respectively.

Step 4 Steps 1–3 are repeated 500 times.

For one of the 500 samples B , Figure 2 shows the weighted kernel density of y estimated from SHIW for the purpose of benchmark comparison and the kernel density estimate of y estimated from B . The bandwidth selection rule is as proposed in Sheather and Jones (1991). As clearly seen, both the distributions are right-skewed but the B sample pdf is very different from the SHIW pdf. In Table 8 the bias (B) and the standard deviation (Sd) of the estimates $\hat{\mu}_{y,I}$, $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$ obtained under scenarios 1,3,4, over the 500 samples, are reported. Furthermore, in Table 9 the corresponding RMSEs are computed.

As results in Table 8 show, the estimate $\hat{\mu}_{y,ISP}$ ignoring the sampling process in B shows a slight reduction in the bias compared to the estimates $\mu_{y,I}$ because of the introduction of the calibration constraint in Equation (10). A larger reduction in the bias is obtained when the selection process in B is taken into account as the estimates $\hat{\mu}_{y,SP}$ show. Finally, the RMSE of $\hat{\mu}_{y,SP}$ is lower than that of comparison estimators as shown in Table 9.

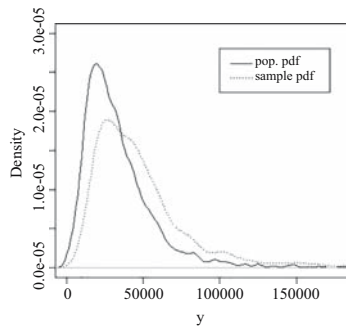


Fig. 2. Income pdf from SHIW data set, sample pdf from B .

Table 8. Bias (B) and standard deviation (Sd) of $\hat{\mu}_{y,I}$, $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples. True parameter is $\mu_y = 32,714$.

$B(\hat{\mu}_{y,I})$	$B(\hat{\mu}_{y,ISP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,I})$	$Sd(\hat{\mu}_{y,ISP})$	$Sd(\hat{\mu}_{y,SP})$
15803.01	13331.07	-2527.36	420.49	328.01	261.54

Table 9. RMSE of $\hat{\mu}_{y,I}$, $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples

$RMSE(\hat{\mu}_{y,I})$	$RMSE(\hat{\mu}_{y,ISP})$	$RMSE(\hat{\mu}_{y,SP})$
15808.60	13335.10	2540.86

8. Simulation Study 2

In this section we evaluate the performance of the mixed approach described in Section 4. We generate a finite population $\{\mathbf{x} = (x_{1i}, x_{2i}), \mathbf{y} = (y_{1i}, y_{2i}): i = 1, \dots, N\}$ with size $N = 1,000,000$ where y_1 is a continuous outcome while y_2 is a binary outcome. From the finite population we select a sample B where the inclusion indicator $\delta_i \sim \text{Ber}(p_i)$ with p_i the inclusion probability for unit i . We obtain a representative sample $A = \{x_{1i}, x_{2i}, d_i\}$ of size $n = 1,000$ using *srs*. As in [Yang et al. \(2021a\)](#), for generating the finite population we consider the following models,

$$y_{1i} = 1 + x_{1i} + x_{2i} + \alpha_i + \varepsilon_i, \quad (21)$$

$$P(y_{2i} = 1 | x_{1i}, x_{2i}; \alpha_i) = \text{logit}^{-1}(1 + x_{1i} + x_{2i} + \alpha_i), \quad (22)$$

where $x_1 \sim N(1, 1)$, $x_2 \sim \text{Exp}(1)$, $\alpha \sim N(0, 1)$, $\varepsilon \sim N(0, 1)$, and x_1 , x_2 , α and ε are mutually independent. The variable α induces the dependence of y_1 and y_2 even adjusting for x_1 and x_2 . The point biserial correlation coefficient between y_1 and y_2 is 0.32. For the inclusion probability in B , we consider the following logistic linear model,

$$p_i = \text{logit}^{-1}(y_{1i}). \quad (23)$$

The expected size of the subsample of units in probability sample A with the membership information ($\delta_i = 1$) is 873. Notice that, in the logistic regression there should be an adequate number of outcomes per predictor variable to avoid an overfit model. [Agresti \(2007\)](#) suggests that there should be ten outcomes for each independent variable. However, the issue has not been definitively settled. We compare the following estimators:

1. $\hat{\mu}_{HT}$, the Horvitz-Thompson estimator assuming that (y_{1i}, y_{2i}) are observed in sample A for the purpose of benchmark comparison.
2. $\hat{\mu}_{NN}$, the nearest neighbor imputation estimator where the imputed values (y_{1i}, y_{2i}) are obtained by nearest neighbor method, as described in section 4.
3. $\hat{\mu}_{RC}$, the regression calibration estimator based on $\hat{\mu}_{NN}$ with calibration variables $H(\delta, \mathbf{x}, \mathbf{y}) = (\delta, 1 - \delta, \delta \mathbf{x}, \delta \mathbf{y})'$, as described in [Yang et al. \(2021a\)](#).
4. $\hat{\mu}_W$, the Horvitz-Thompson estimator with weights \tilde{w}_i obtained by regressing the membership indicator δ against $(x_{1i}, x_{2i}, \tilde{y}_{1i}, \tilde{y}_{2i})$ in sample A , as described in section 4 (Step 3, point 3.1),
5. $\hat{\mu}_{EL}$, the estimator based on maximization of empirical likelihood, as described in section 4 (Step 3, point 3.2).

The simulation is based on 1,000 Monte Carlo runs. [Table 10](#) summarizes the simulation results with biases, standard deviations and coverage rates of 95% confidence intervals using asymptotic normality of the aforementioned estimators. All the results are multiplied by 100. The population means of y_1 and y_2 are 3 and 0.89, respectively.

First of all, if sample B is simply treated as a simple random sample the bias is 33.1% and 2.3% for the mean of y_1 and y_2 , respectively. In [Table 10](#) the estimators $\hat{\mu}_{NN}$ and $\hat{\mu}_{RC}$ have the larger bias, but $\hat{\mu}_{RC}$ has a smaller standard error than $\hat{\mu}_{NN}$. Recall that both estimators implicitly assume that the selection mechanism for sample B is ignorable. With regard to the mean of y_1 , $\hat{\mu}_W$ has the smaller bias (13.3%) followed by $\hat{\mu}_{EL}$

Table 10. Bias (B_h), standard deviation (Sd_h), RMSE and coverage rate of 95% confidence interval (CR_h) for the population mean of y_h , $h = 1, 2$, based on 1,000 Monte Carlo samples. The population means of y_1 and y_2 are 3 and 0.89, respectively. All the results are multiplied by 100.

Estimator	B_1	Sd_1	$RMSE_1$	CR_1	B_2	Sd_2	$RMSE_2$	CR_2
$\hat{\mu}_{HT}$	0.2	6.2	6.2	95.4	0.0	1.0	1.0	96.6
$\hat{\mu}_{NN}$	20.2	5.6	21.0	93.8	1.2	0.9	1.5	96.6
$\hat{\mu}_{RC}$	20.1	1.8	20.2	96.0	1.3	0.4	1.4	95.4
$\hat{\mu}_W$	13.3	1.9	13.4	95.8	-0.4	0.3	0.5	93.4
$\hat{\mu}_{EL}$	16.12	0.1	16.1	99.4	-0.02	0.3	0.3	94.4

(16.12%) even if $\hat{\mu}_{EL}$ has a smaller standard error and a larger coverage rate ($CR_1 = 99.4\%$). The opposite occurs for the mean of y_2 , the bias is $B_2 = -0.4\%$ for $\hat{\mu}_W$ against $B_2 = -0.02\%$ for $\hat{\mu}_{EL}$. The coverage rates are all close to the nominal level. In conclusion, both μ_W and μ_{EL} reduce the bias in estimating μ_y respect to $\hat{\mu}_{NN}$ and $\hat{\mu}_{RC}$. The Horvitz-Thompson estimator ($\hat{\mu}_W$) seems to perform better for continuous variables while the estimator based on maximization of empirical likelihood ($\hat{\mu}_{EL}$) according to the two steps procedure described in Section 4 seems to show a better performance for binary variables.

9. Concluding Remarks

In this article two approaches for reducing selection bias when the selection process is non-ignorable are proposed. The first one based on EL requires to model parametrically the unknown selection probabilities and to maximize the sample likelihood with respect to the sampling and the population parameters. Auxiliary information known for the population or estimable from a probability sample can be incorporated in the maximization process, thus enhancing the precision of the estimators. As previously stressed, the success of the proposed approach depends on proper modeling of the unknown selection probabilities. However, the resulting sample model can be tested from the data by standard test statistics, see Subsection 6.2.1. A broad simulation study illustrates the good performance of the EL approach also when skewed and binary data are considered, see Subsection 6.2.2. Finally, the proposed approach can be extended to the multivariate case when several variables of interest are considered. For variables selection in modeling $P(\delta_i = 1 | \mathbf{x}_i, \mathbf{y}_i)$ see Variyath et al. (2010) and Chen et al. (2022). We obviously hope that other researchers will apply our proposed approach with appropriate modifications required for their data.

The second one is a mixed approach based on mass imputation and propensity score adjustment. It requires that the membership to nonprobability sample can be determined throughout the probability sample A . As indicated by the results in Section 8, the method seems to show a good performance in terms of bias, standard error and confidence interval coverage probabilities. Empirical studies with alternative population and selection models are needed to further ascertain the results of the mixed approach obtained in the present article. Finally, new theoretical developments of the present work include the use of proxy variables that can help studying the relationship between y and δ and in particular, help verifying or refuting the ignorability assumption.

10. References

- Agresti, A. 2007. *An Introduction to Categorical Data Analysis* (second edition). John Wiley & Sons, Inc., Hoboken: New Jersey.
- Babu, G.J., and C.R. Rao. 2004. "Goodness-of-Fit Tests when Parameters are Estimated." *Sankhyā. Series A* 66(1): 63–74. DOI: <https://doi.org/10.2307/25053332>.
- Banca d'Italia. 2012. *Supplement to the Statistical Bulletin, Sample Surveys, Household income and wealth in 2010*: 12(6). Available at: <https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/bil-fam2010>.
- Beaumont, J.F. 2000. "An Estimation Method for Nonignorable Nonresponse." *Survey Methodology* 26(2): 131–136. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000002/article/5532-eng.pdf?st=WJWdN_31.
- Belzile, L., J.L. Wadsworth, P.J. Northrop, S.D. Grimshaw, J. Zhang, M.A. Stephens, A.B. Owen, and R. Huser. 2022. *mev: Modelling Extreme Values*. R package version 1.14 Available at: <https://CRAN.R-project.org/package=mev> (accessed June 2022).
- Beresewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio and M. Karlberg. 2018. *An overview of methods for treating selectivity in big data sources*. Statistical Working Papers, Eurostat. Available at: <https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-tc-18-004> (accessed July 2022).
- Chang, T., and P.S. Kott. 2008. "Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model." *Biometrika* 95(3): 555–571. DOI: <https://doi.org/10.1093/biomet/asn022>.
- Chaudhuri S., M.S. Handcock, and M.S. Rendall. 2010. *A conditional empirical likelihood approach to combine sampling design and population level information*. Technical report No. 3/2010, National University of Singapore, Singapore. Available at: https://cpb-us-w2.wpmucdn.com/blog.nus.edu.sg/dist/0/14452/files/2020/10/tr03_2010.pdf (accessed July 2022).
- Chen, C., M. Wang, R. Wu, and R. Li. 2022. "A Robust Consistent Information Criterion for Model Selection Based on Empirical Likelihood." *Statistica Sinica* 32: 1205–1223. DOI: <https://doi.org/10.5705/ss.202020.0254>.
- Conti P.L., D. Marella, and M. Scanu. 2008. "Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators." *Computational Statistics & Data Analysis* 53(2): 354–365. DOI: <https://doi.org/10.1016/j.csda.2008.07.041>.
- DiSogra, C., C. Cobb, E. Chan, and J. M. Dennis. 2011. "Calibrating non-probability internet samples with probability samples using early adopter characteristics." In *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*. Miami Beach, Florida, July 30-August 4, 2011: 4501–4515. Alexandria, VA: American Statistical Association. Available at: <http://www.asasrms.org/Proceedings/y2011/Files/30270468925.pdf> (accessed June 2022).
- Elliott, M., and R. Valliant. 2017. "Inference for non-probability samples." *Statistical Science* 32(2): 249–264. DOI: <https://doi.org/10.1214/16-STS598>.
- Feder, M., and D. Pfeffermann. 2019. *Statistical Inference Under Non-ignorable Sampling and Non-response. An Empirical Likelihood Approach*. Working paper. University of Southampton. Available at: <https://eprints.soton.ac.uk/378245/> (accessed July 2022).

- Galimard J.E., S. Chevret, E. Curis, and M. Resche-Rigon. 2018. "Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors." *BMC Medical Research Methodology* 18(90). DOI: <https://doi.org/10.1186/s12874-018-0547-1>.
- Hájek, J. 1964. "Asymptotic theory of rejective sampling with varying probabilities from a finite population." *The Annals of Mathematical Statistics* 35(4): 1491–1523. DOI: [10.1214/aoms/1177700375](https://doi.org/10.1214/aoms/1177700375).
- Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153–161. DOI: [http://dx.doi.org/10.2307/1912352](https://dx.doi.org/10.2307/1912352).
- Kim, J.K., and Z. Wang. 2019. "Sampling techniques for big data analysis in finite population inference." *International Statistical Review* 87(S1): S177–S191. DOI: <https://doi.org/10.1111/insr.12290>.
- Kott, P.S. 2006. "Using calibration weighting to adjust for nonresponse and coverage errors." *Survey Methodology* 32(2): 133–142. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9547-eng.pdf?st=B2aZNvo0>.
- Kott, P.S., and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105(491): 1265–1275. DOI: <https://doi.org/10.1198/jasa.2010.tm09016>.
- Lee, J., and J.O. Berger. 2001. "Semiparametric Bayesian analysis of selection models." *Journal of the American Statistical Association* 96(456): 1397–1409. DOI: <https://doi.org/10.1198/016214501753382318>.
- Marella D., M. Scanu, and P.L. Conti. 2008. "On the matching noise of some nonparametric imputation procedures." *Statistics & Probability Letters* 78(12): 1593–1600. DOI: <https://doi.org/10.1016/j.spl.2008.01.020>.
- Marella, D., and D. Pfeffermann. 2019 "Matching Information from two independent informative samples." *Journal of Statistical Planning and Inference* 203: 70–81. <https://doi.org/10.1016/j.jspi.2019.03.001>.
- Marella, D., and D. Pfeffermann. 2021 "Accounting for nonignorable sampling and nonresponse in statistical matching." *International Statistical Review*. Accepted for publication. DOI: <https://doi.org/10.1111/insr.12524>.
- Meng, X-L., 2018. "Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox and the 2016 US presidential election." *The Annals of Applied Statistics* 12(2): 685–726. DOI: <https://doi.org/10.1214/18-AOAS1161SF>.
- Owen, A.B. 2001. *Empirical Likelihood*. Chapman & Hall/CRC: New York.
- Owen, A.B. 2013. "Self-concordance for empirical likelihood." *Canadian Journal of Statistics* 41(3): 387–397. DOI: <https://doi.org/10.1002/cjs.11183>.
- Pfeffermann, D., A.M. Krieger, and Y. Rinott. 1998. "Parametric distribution of complex survey data under informative probability sampling." *Statistica Sinica* 8(4): 1087–1114.
- Pfeffermann, D., and M. Sverchkov. 2009. "Inference under Informative Sampling." In *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*, edited by D. Pfeffermann and C.R. Rao.: 455–487. North Holland.
- Pfeffermann, D. 2011. "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?" *Survey Methodology* 37(2): 115–136. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11602-eng.pdf?st=_vXrCcPb.

- Pfeffermann, D., and V. Landsman. 2011. "Are private schools really better than public schools? Assessment by methods for observational studies." *Annals of Applied Statistics* 5(3): 1726–1751. DOI: <https://doi.org/10.1214/11-AOAS456>.
- Pfeffermann, D., and A. Sikov. 2011. "Imputation and Estimation under Nonignorable Non-response in Household Surveys with Missing Covariate Information." *Journal of Official Statistics* 27(2): 181–209. Available at: <https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/imputation-and-estimation-under-nonignorable-nonresponse-in-household-surveys-with-missing-covariate-information.pdf>.
- Pfeffermann, D. 2015. "Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture." *Journal of Survey Statistics and Methodology* 3(4): 425–483. DOI: <https://doi.org/10.1093/jssam/smv035>.
- R Core Team 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: <http://www.R-project.org/>.
- Riddles, M.K., J. K. Kim, and J. Im. 2016. "A propensity-score-adjustment method for non-ignorable nonresponse." *Journal of Survey Statistics and Methodology* 4(2): 215–245. DOI: <https://doi.org/10.1093/jssam/smv047>.
- Rivers, D. 2007. "Sampling for web surveys." In Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings. Salt Lake City, Utah, July 29-August 2, 2007: 4127–4134. Alexandria, VA: American Statistical Association. Available at: http://www.websm.org/uploadi/editor/1368187629Rivers_2007_Sampling_for_web_surveys.pdf (accessed June 2022).
- Rosenbaum, P.R., and D.B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41–55. DOI: <https://doi.org/10.1093/biomet/70.1.41>.
- Rubin, D.B. 1976. "Inference and missing data." *Biometrika* 63(3): 581–592. DOI: <https://doi.org/10.1093/biomet/63.3.581>.
- Sheather, S.J., and M.C. Jones. 1991. "A reliable data-based bandwidth selection method for Kernel density estimation." *Journal of the Royal Statistical Society. Series B-Statistical Methodology* 53(3): 683–690. DOI: <https://doi.org/10.2307/2345597>.
- Variyath, A. M., J. Chen, and B. Abraham. 2010. "Empirical likelihood based variable selection." *Journal of Statistical Planning and Inference* 140(4): 971–981. DOI: <https://doi.org/10.1016/j.jspi.2009.09.025>.
- Yang, S., J.K. Kim, and Y. Hwang. 2021a. "Integration of data from probability surveys and big found data for finite population inference using mass imputation." *Survey Methodology* 47(1): 29–58. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf?st=WLDQdr7>.
- Yang, S., J.K. Kim, and R. Song. 2021b. "Doubly Robust Inference when Combining Probability and Nonprobability Samples with High-dimensional Data." *Journal of the Royal Statistical Society. Series B-Statistical Methodology* 82(2): 445–465. DOI: <https://doi.org/10.1111/rssb.12354>.

Received July 2021

Revised April 2022

Accepted July 2022

Design and Sample Size Determination for Experiments on Nonresponse Followup using a Sequential Regression Model

Andrew M. Raim¹, Thomas Mathew¹, Kimberly F. Sellers¹, Renee Ellis¹, and Mikelyn Meyers¹

Statistical agencies depend on responses to inquiries made to the public, and occasionally conduct experiments to improve contact procedures. Agencies may wish to assess whether there is significant change in response rates due to an operational refinement. This work considers the assessment of response rates when up to L attempts are made to contact each subject, and subjects receive one of J possible variations of the operation under experimentation. In particular, the continuation-ratio logit (CRL) model facilitates inference on the probability of success at each step of the sequence, given that failures occurred at previous attempts. The CRL model is investigated as a basis for sample size determination— one of the major decisions faced by an experimenter—to attain a desired power under a Wald test of a general linear hypothesis. An experiment that was conducted for nonresponse followup in the United States 2020 decennial census provides a motivating illustration.

Key words: Continuation-ratio logit; design of experiments; general linear hypothesis; generalized linear models; embedded experiments.

1. Introduction

Sample surveys and censuses are heavily relied upon to measure characteristics of a population. These methods of data collection involving direct contact with members of the population provide the basis for most official statistics. A major and growing problem is nonresponse, which can occur for a variety of reasons, including inability to contact respondents or refusal to participate (e.g., Singer 2006). Missing responses can bias inference from the data when the underlying cause of nonresponse is associated with characteristics to be measured. Lohr (2010, chap. 8) provides further discussion on effects of nonresponse and summarizes a variety of techniques developed to reduce and adjust for missing responses; these include followup operations to make further contact attempts (“callbacks”), imputing missing responses, and adjusting estimates by weights based on response probabilities. The present article focuses on callbacks, which have been an effective strategy for improving response rates; (see Hansen and Hurwitz 1946; Politz and

¹ U.S. Census Bureau, Washington, D.C, 20233, U.S.A., Emails: andrew.raim@census.gov, thomas.mathew@census.gov, kimberly.flagg.sellers@census.gov, renee.ellis@census.gov, and mikelyn.v.meyers@census.gov

Acknowledgments: The authors thank Luke Larson, Kathleen Kephart, and Marcus Berger (Center for Behavioral Science Methods, U.S. Census Bureau) for useful discussions on the enumerator training experiment. We are also grateful to Jennifer Hutnick (Decennial Statistical Studies Division, U.S. Census Bureau) and Eric Slud (Center for Statistical Research and Methodology, U.S. Census Bureau), the associate editor, and the anonymous referees for insightful feedback regarding the manuscript. This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the authors and not those of the U.S. Census Bureau.

Simmons 1949; Deming 1953; Rao 1983; Särndal et al. 1992, sec. 15.4.2). Consideration has been given to the use of administrative records and other available sources of data to augment or replace field work in official statistics (e.g., Scheuren 1999; Morris et al. 2016; Daas et al. 2015; Brown et al. 2018). However, such use of administrative data presents its own challenges including lack of public availability and data structures that are not intended for this particular application (Davern et al. 2009; Molfino et al. 2017; Groves and Schoeffel 2018). With field work currently the primary method of data collection, measuring and improving response rates continue to be of major interest to statistical agencies.

One of the major data collection activities of the U.S. Census Bureau is the decennial census, which seeks to contact every household and group quarters in the United States and record basic information, such as the number of residents along with age and race. Census data are used to produce statistical summaries which are disseminated to the public. Households are initially invited to self-respond via mail or another convenient mode. Households which do not respond within a certain time period become part of the Nonresponse Followup (NRFU) operation. Here, enumerators attempt to personally contact the household and elicit a response. The specific contact strategy designed in the years leading up to the census typically includes in-person visits to the household. NRFU was the most expensive component of the 2010 decennial census, with a cost of about USD 1.6 billion (Walker et al. 2012).

A variety of experiments are typically conducted in the years leading up to the decennial census, and also within the census itself, to test whether modifications in the operation make significant changes to response rates. The National Research Council (2010) describes experiments carried out by the U.S. Census Bureau for decennial censuses between the years 1950 and 2010. One notable experiment that took place in the 2010 Census Program of Experiments and Evaluations (CPEX) is described in detail by Compton and Bentley (2012). This experiment was carried out within the 2010 decennial census to study the effect of reducing the number of callbacks in NRFU from six to either four or five, concluding that such a reduction may reduce operational costs and does not appear to have a detrimental effect on case completion. Here, experimental NRFU questionnaires (supporting four or five callbacks) were systematically distributed into case workloads among standard questionnaires (utilizing six callbacks). Conclusions were drawn from a series of cross-tabulations which provided a number of perspectives on response behavior. In this setting, a formal hypothesis testing approach may be desired to quantify uncertainty in conclusions, and can also be used to ensure adequate certainty (i.e., power of the test) in designing the experiment. Response rates over the sequence of callbacks may be of interest; therefore, we may wish to take the sequential process into account when formulating the test.

This article explores the use of a sequential regression model in measuring response rates where multiple callback attempts can be made to the same household. The continuation-ratio logit (CRL) model, also referred to as the sequential logit model, is a particular parameterization of the multinomial distribution which can be interpreted as a truncated sequence of dependent Bernoulli trials. This makes it a suitable extension of logistic regression when modeling the number of attempts required for a successful contact, rather than merely the occurrence of successful contact. In particular, we consider

a procedure for selecting a sample size in a study whose goal is to test a general linear hypothesis; this is used to detect whether two or more treatments in an experiment lead to significantly different response rates. When such effects vary over the sequence of attempts, CRL can express the situation while a model capturing only response or nonresponse cannot.

An experiment which was carried out within the 2020 decennial census serves as a motivating application of the CRL methodology. Enumerators hired by the agency are given formal training before participating in field operations. For the 2020 decennial census, the Census Bureau tested the inclusion of training for bilingual enumerators when administering the census questionnaire in their non-English (“target”) language(s) (Ellis et al. 2018). The agency did not provide such training prior to the 2020 census. The training took the form of a brief module added to the larger suite of training materials for bilingual, Spanish-speaking enumerators. The objective of additional training was to improve consistency in messaging and in the usage of official translations, rather than bilingual enumerators translating messages, and in some cases the survey itself, into Spanish in an impromptu fashion. Increased consistency may result in improved response rates and improved data quality for affected households (Pan and Lubkemann 2013). There was thought to be little disadvantage to deploying the new training module; it did not constitute a major cost when implemented as an experimental intervention, and a negative impact to response rates was not expected. However, it was of interest whether the training significantly improved response rates for affected households in the context of NRFU attempts. In an internal U.S. Census Bureau report, Ellis et al. (2018) detail the experiment carried out within the 2020 census NRFU operation to make this assessment. We emphasize that the experiment is presented in the current article to demonstrate our methodology and does not represent planning or analysis for the actual experiment, nor any position of the U.S. Census Bureau. An official report on the experiment, to be released to the public, is in progress at the time of this writing. Note that later stages of planning and execution of NRFU and other 2020 census operations coincided with the COVID-19 pandemic, but such complications will not be considered in this work. In the present article, we will consider the use of CRL models in two important aspects of experiment planning: to formulate a design which incorporates sequential contact attempts used in field operations, and to select a sample size with adequate statistical power to evaluate effectiveness of the training.

Sequential models such as CRL have been widely used in a variety of applications, including survival analysis (Cox 1972; Albert and Chib 2001), social science (Fullerton 2009), economics (Boes and Winkelmann 2006), and public health (Barboza and Dominguez 2016). CRL is also closely connected to stick-breaking processes used to fit Dirichlet process models in Bayesian analysis; for example, see Ghosal and Van der Vaart (2017, chap. 3) and Rigon and Durante (2021). Use for nonresponse in official statistics settings, however, appears to be relatively limited. Alho (1990) formulates a model for nonresponse based on CRL for the purpose of adjusting survey estimates to avoid bias. A similar approach was taken later by Wood et al. (2006). Fienberg (2007, chap. 6) provides an overview of CRL in the context of contingency tables, while Agresti (2013, chap. 8) provides an overview in the context of multinomial regression. Tutz (1991) explores connections between models for sequential data (including CRL) and those for ordinal

data such as the proportional odds model. [Tutz \(1991\)](#) also establishes sequential models as multivariate generalized linear models (GLMs). More recently, [Peyhardi et al. \(2015\)](#) and [Tutz \(2022\)](#) discuss proportional odds models (POMs), CRL, and extensions within larger families of models for categorical data.

Sample size calculation is the subject of a large literature; the following brief summary features a few examples to help give context for the present work. [Chow et al. \(2017\)](#) provide a general reference for sample size calculation in a number of non-regression settings. [Self and Mauritsen \(1988\)](#) consider power calculations for a score test in the context of a GLM; there are several important features in this work which appear in later references. These authors partition the regression coefficients into a parameter of interest whose value is specified in the null hypothesis, and a nuisance parameter which is estimated. Second, covariates are treated as random variables whose distribution must be considered. In particular, [Self and Mauritsen \(1988\)](#) assume categorical covariates. [Self et al. \(1992\)](#) explore a likelihood ratio test in the setting of GLMs and make use of an asymptotic expansion to compute power. [Shieh \(2000\)](#) extends [Self et al. \(1992\)](#) and removes the restriction that covariates must be categorical. [Shieh \(2005\)](#) studies a Wald test in GLMs; here an adjustment is made to the significance level to account for the large sample approximation. [Demidenko \(2007, 2008\)](#) consider a Wald test, but focus on a more specific case/control setting in logistic regression with binary covariates. [Lyles et al. \(2007\)](#) explore Wald and likelihood ratio tests in GLMs, assuming a general linear hypothesis which subsumes the partitioning of test and nuisance parameters. These authors propose a computational approach which allows a specified distribution of the covariates to be studied without requiring derivations for each new setting. [Bush \(2015\)](#) summarizes many of the previously referenced works and investigates them by simulation.

The present work focuses on sample size determination under the CRL model to compare the impact of several treatments on response probability. The comparison is quantified by formulating a hypothesis based on CRL parameters and by fixing a test procedure. Therefore, an adequate sample size is one in which the test procedure achieves the desired power. A general linear hypothesis is utilized to incorporate a range of hypotheses which may be of interest in an experimental setting. The Wald test provides an explicit formula for the asymptotic power and facilitates power studies where the sample size, number of contact attempts included in the analysis, and any nuisance parameter are varied. The objective of this work differs from [Alho \(1990\)](#) and [Wood et al. \(2006\)](#), who adjust for nonresponse using a CRL model but ultimately seek to produce survey estimates for a population variable of interest. Another major departure from the referenced work is that we condition on covariates so that they are fixed throughout sample size determination. Possessing covariate information on the population of interest may be more realistic in an official statistics setting than in the clinical setting that pertains to most of the referenced literature. An additional departure is how we handle the “nuisance” part of the parameter which is not dictated by the test hypothesis; we take this to be fixed based on a priori information rather than estimated. To compute the power for a given departure from the null hypothesis, we utilize an optimization over the parameter space to ensure that the power calculations are conservative.

The remainder of the article is organized as follows. Section 2 presents the illustration motivated by the enumerator training experiment which requires a sample size to be

specified. Section 3 recalls the CRL model and basic inference using maximum likelihood estimation. Section 4 presents a method of sample size determination under the CRL model. Section 5 applies our sample size method to the illustration from Section 2; this permits the effects of interest—which define the hypothesis of the test whose power is used to justify a sample size—to vary both by treatment and enumeration attempt. Subsection 5.1 connects the study to the notation and testing framework; Subsection 5.2 presents simulation results comparing the empirical power of the test to the approximation described in Section 4; finally, Subsection 5.3 presents a power study using the illustration so that a sample size can be justified. A brief discussion in Section 6 concludes the article.

2. A Motivating Illustration

The enumerator training experiment described in Section 1 serves as a compelling motivation for this work. While some details from the actual experiment have been included for demonstration, others require further methodological development and are discussed in Section 6.

The experiment was envisioned to be carried out within the decennial census, hence its design must be compatible with census operations. It is worthwhile to review the major components of the experiment, such as the experimental subjects, treatments, and the meaning of “sample size.” A general reference for experimental design is [Oehlert \(2000\)](#). Experimental subjects here are Spanish-speaking households in the NRFU operation; these are not known with certainty until the actual NRFU operation is carried out, so we make use of estimates from previous operations in the planning phase. The number of households included in the study is therefore associated with the sample size, but is not something which we can directly manipulate in the design. Parameters of interest are probabilities of Spanish-speaking households to respond to the NRFU operation.

As experimenters, we can assign control (“no training”) or experimental (“training”) treatments to enumerators. For practical considerations which will now be described, we assign treatments at the level of area census office (ACO) rather than to individual enumerators. For this discussion, an ACO is considered to be a geographic delineation used in data collection for the census. Tracts from the standard (“tabulation”) geography can generally overlap with multiple ACOs; however, tracts intersecting the ACOs used in this study are contained strictly in one ACO. Enumerators associated with an experimental ACO will receive the new training, while those in a control ACO will not receive the new training. We cannot directly assign individual households to enumerators; instead, case assignments will be made dynamically based on enumerator availability and workloads ([U.S. Census Bureau 2019](#)). Under this system, each enumerator will visit multiple households, and a household may be visited by multiple enumerators. We wish to avoid situations of “contamination” where households in the study are visited by both trained and untrained enumerators. To minimize the risk of such occurrences, we ensure that control and experimental ACOs are geographically separated. After the data collection, any cases in which a household is visited by both trained and untrained enumerators will be discarded from the analysis.

The number of households in the sample is controlled via the ACOs we select for the experiment. This selection must be decided sufficiently in advance of field operations. To

minimize impact to operations, we would prefer a small number of ACOs which will provide adequate power. We have pre-selected ACOs from several metropolitan statistical areas (MSAs) in Dallas and Houston, Texas and Los Angeles, California as a starting point. Historically, these areas have had large numbers of residents who primarily speak Spanish and also a large expected workload for NRFU. [Table 1](#) displays the fourteen pre-selected ACOs: six in the Dallas area, six in Houston, and two in Los Angeles. All ACOs in Dallas are assigned to the control group, while Houston is assigned to the experimental group. Of the two ACOs in Los Angeles, one is assigned to the experimental group and the other to the control group.

We gathered additional data from the Census Bureau Planning Database ([Erdman and Bates 2017](#)) for the selected ACOs, including the total number of households (HH_Total), percent of Spanish speakers (Pct_Spanish), and percent of self-responders (Pct_Selfresp). We obtain a rough estimate of the count of relevant households in each ACO using the formula

$$\text{HH-Target} = \text{HH-Total} \times \text{Pct-Spanish}/100 \times (1 - \text{Pct-Selfresp}/100), \quad (1)$$

and truncating to an integer. Calculation (1) is carried out at the tract level, then aggregated to the ACO level. This provides a total sample size of up to 380,018 households; although this represents a small proportion of households in the United States, it seems to be quite a large number to use in an experiment. A formal power analysis will reveal whether or not it is sufficient.

Table 1. ACOs under consideration for the experiment.

Pair	Area	Group	Tracts	Percent		HH counts	
				Spanish	Selfresp	Total	Target
1	Dallas	Ctrl	176	6.8	62.8	352,347	11,900
1	Houston	Expt	136	21.0	44.1	253,932	33,305
2	Dallas	Ctrl	163	14.2	48.5	293,170	24,847
2	Houston	Expt	148	10.1	47.9	278,782	18,412
3	Dallas	Ctrl	180	10.4	57.4	337,574	19,828
3	Houston	Expt	140	15.8	44.0	282,424	31,434
4	Dallas	Ctrl	170	24.9	41.2	277,452	43,271
4	Houston	Expt	122	21.6	41.3	240,950	36,575
5	Dallas	Ctrl	194	11.6	55.6	335,557	23,521
5	Houston	Expt	146	20.0	40.7	238,144	32,587
6	Dallas	Ctrl	235	4.0	66.3	482,153	8,084
6	Houston	Expt	91	8.0	61.3	268,572	9,525
7	LA	Ctrl	304	13.9	49.5	441,726	35,989
7	LA	Expt	355	16.1	48.5	496,564	50,740
Total			2,560			4,579,347	380,018

¹Total HH Counts, Percent Spanish, and Percent Self-Response are based on Planning Database variables Tot_Occp_Units_ACS_13_17, pct_Age5p_Spanish_ACS_13_17, and Self_Response_Rate_ACS_13_17, respectively, which are sourced from American Community Survey 5-year estimates for the year 2017.

²Percentages are based on ACOs counts which have been aggregated from tract data; Target HH Count cannot be reproduced via Equation (1) from here.

Of primary concern is whether the fourteen available ACOs will be adequate to test the effect of enumerator training on response rate. A secondary interest is in plotting power curves as the sample size is increased. To accomplish the second goal, we consider arranging the seven experimental and seven control ACOs into $I = 7$ pairs where each pair contains one ACO for each of the $J = 2$ treatments. The Los Angeles ACOs form one pair, while remaining pairs are constructed by matching an ACO from Houston with an ACO from Dallas where Pct_Spanish and Pct_Selfresp are most similar. After matching, pairs were randomly assigned indices $i = 1, \dots, I$ which are displayed in the “pair” column of Table 1. A power study may be carried out by first including only the pair of ACOs labeled $i = 1$, then the four total ACOs labeled either $i = 1$ or $i = 2$, and so on until all fourteen ACOs are included, to obtain a power curve with $I = 7$ points. We have described one possible way of characterizing an increasing number of households in this experiment; other variations will yield different power curves. If the ordering will be used for more than a demonstration—such as in limiting the scope of an experiment to control costs—stakeholders may need to establish an arrangement which prioritizes experimental desiderata and practicality while avoiding selection bias. In this setting, such a bias might occur if a particular arrangement were chosen to make the training exhibit a more or less favorable effect.

3. Continuation-Ratio Logit Model

To motivate the continuation-ratio logit (CRL) model, let $\{p_\ell\}$ denote a sequence of probabilities for $\ell \in \{1, 2, \dots\}$ with $p_\ell \in (0, 1)$. Define a discrete random variable W^* whose support is the set of positive integers $\{1, 2, \dots\}$ with probabilities $P(W^* = \ell) = p_\ell \prod_{b=1}^{\ell-1} (1 - p_b)$. The random variable W^* naturally represents a number of Bernoulli trials required to obtain the first success in a sequence of heterogeneous trials. In the special case of a common $p_\ell = p$, W^* follows a geometric distribution. For example, W^* could represent the number of attempts required for a successful enumeration of a given household, provided there were no upper limit on attempts. In practice, it may be reasonable to assume an upper bound L for the number of trials. For example, this would be suitable if L maximum number of attempts is prespecified for enumeration, or if results greater than L will be truncated in the observed data. Here, it is natural to consider truncating W^* to $W = W^* \cdot I(W^* \leq L) + (L + 1) \cdot I(W^* > L)$. With this construction, W has support $\{1, \dots, L + 1\}$ where the event $[W = L + 1]$ indicates that no response was observed in the first L attempts under consideration.

By this construction, W follows a CRL distribution which we will write as $W \sim CRL_L(\mathbf{p})$ with $\mathbf{p} = (p_1, \dots, p_L)$. We may write the overall probability of the sequence resulting in success at the ℓ th trial as

$$\pi_\ell \stackrel{\text{def}}{=} P(W = \ell) = p_\ell \prod_{b=1}^{\ell-1} (1 - p_b), \quad \ell = 1, \dots, L + 1, \tag{2}$$

with $p_{L+1} \equiv 1$. It can be shown that $\pi_1 + \dots + \pi_{L+1} = 1$ when defined in this way, emphasizing that W also follows a multinomial distribution. Using Equation (2), we can obtain a transformation from $(\pi_1, \dots, \pi_{L+1})$ to $(p_1, \dots, p_L, p_{L+1})$ using

$$p_\ell = \frac{\pi_\ell}{\pi_\ell + \dots + \pi_{L+1}}, \quad \ell, \dots, L + 1. \tag{3}$$

From Equation (3), it is clear that each $p_\ell = P(W = \ell \mid W \geq \ell)$ is the conditional probability of success on the ℓ th trial given that trials $1, \dots, \ell - 1$ were unsuccessful. The quantity from Equation (3) is also referred to as a discrete hazard rate in survival analysis (Ghosal and Van der Vaart 2017, chap. 3).

Now, consider a random sample $W_i \sim \text{CRL}_L(p_i)$ for $i = 1, \dots, n$ where W_i represents the outcome for the i th subject. A common truncation of L trials will be taken for all n subjects in this work. We are typically interested in the relationship between response probability and an independent variable $\mathbf{x}_{i\ell} \in \mathbb{R}^d$ which is provided for each $i = 1, \dots, n$ and may vary with trial $\ell = 1, \dots, L$. A logistic link can be used to explicitly make the connection

$$\text{logit}(p_{i\ell}) = \mathbf{x}_{i\ell}^\top \beta \Leftrightarrow p_{i\ell} = G(\mathbf{x}_{i\ell}^\top \beta), \quad (4)$$

where $G(x) = 1/(1 + e^{-x})$ denotes the inverse logit function, $\beta \in \mathbb{R}^d$ is a vector of unknown regression coefficients which are the objectives of our inference (and whose interpretation depends on the design chosen for $\mathbf{x}_{i\ell}$), and

$$\text{logit}(p_{i\ell}) = \log\left(\frac{p_{i\ell}}{1 - p_{i\ell}}\right) \equiv \log\left(\frac{\pi_{i\ell}}{\pi_{i,\ell+1} + \dots + \pi_{i,L+1}}\right).$$

Dependence of $p_{i\ell}$ on $\mathbf{x}_{i\ell}$ and β is omitted from the notation for brevity. The likelihood is

$$\mathcal{L}(\beta) = \prod_{i=1}^n \prod_{\ell=1}^{L+1} \left\{ G(\mathbf{x}_{i\ell}^\top \beta) \prod_{b=1}^{\ell-1} [1 - G(\mathbf{x}_{ib}^\top \beta)] \right\}^{I(w_i=\ell)}. \quad (5)$$

Remark 1.

A basic form of CRL often used in the literature assumes that $p_\ell(\mathbf{x}_i) = G(\gamma_{0\ell} + \mathbf{x}_i^\top \beta)$, so that attempt probabilities vary only with intercepts $\gamma_{01}, \dots, \gamma_{0L}$ (e.g. Tutz 1991, 2022). This assumption yields the following log odds-ratio for category ℓ , comparing covariate values \mathbf{x}_i and $\mathbf{x}_{i'}$:

$$\log \left[\frac{p_\ell(\mathbf{x}_i)}{1 - p_\ell(\mathbf{x}_i)} \Big/ \frac{p_\ell(\mathbf{x}_{i'})}{1 - p_\ell(\mathbf{x}_{i'})} \right] = (\mathbf{x}_i - \mathbf{x}_{i'})^\top \beta. \quad (6)$$

Interpretation is simplified with Equation (6) constant across attempts. However, this assumption may not be suitable for models of interest in the NRFU setting where attempt effects may be a function of independent variables, independent variable values can vary across attempts, and common effects may be shared across attempts. The selected link (4) is a variation of the discrete choice model (Agresti 2013, sec. 8.5) which provides this flexibility.

Remark 2.

The popular class of proportional odds models (POMs) is also used to link integer-valued outcomes to probabilities of a categorical distribution. A standard POM described by Agresti (2013, chap. 8) proposes L intercepts $\gamma_1 < \dots < \gamma_L$ and a regression $\mathbf{x}_i^\top \beta$. Outcomes V_i are assumed to be independently distributed for $i = 1, \dots, n$, with each V_i taking on values $1, \dots, L + 1$. Probabilities $\pi_{i\ell} = P(V_i = \ell)$ are linked to $\gamma_1, \dots, \gamma_L$ and β via

$$\pi_{i\ell} = G(\gamma_\ell + x_i^\top \beta) - G(\gamma_{\ell-1} + x_i^\top \beta), \quad \ell = 1, \dots, L + 1, \tag{7}$$

with $\gamma_0 \equiv -\infty$ and $\gamma_{L+1} \equiv \infty$. Taking V_i^* to be a logistic random variable with location 0 and scale 1, Equation 7, prescribes that $P(V_i = \ell) = P(\gamma_{\ell-1} < V_i^* - x_i^\top \beta \leq \gamma_\ell)$; therefore, a larger observation of V_i may be interpreted as the unobservable continuous variable $V_i^* - x_i^\top \beta$ having been larger. From this perspective, the POM is suitable as a model for ordinal data, while CRL is appropriate when outcomes represent a stage reached in a sequence of trials.

To facilitate the upcoming discussion, let $\mathcal{J} = ((1, 1), (1, 2), \dots, (n, L))$ denote pairs of indices (i, ℓ) ordered first by trial and then by observation. Let \mathbf{X} denote the $nL \times d$ design matrix with rows $x_{i\ell}^\top$ for $(i, \ell) \in \mathcal{J}$ and $g(x) = e^{-x}/(1 + e^{-x})^2$ denote the first derivative of $G(x)$. The score vector and Fisher information matrix of Equation (5) are routine to obtain, but are now stated as a result for reference and a derivation is given in the Appendix (Section 6).

Result 3.

Under likelihood Equation (5),

a. the score vector is

$$S(\beta) = \frac{\partial}{\partial \beta} \log \mathcal{L}(\beta) = \sum_{i=1}^n \sum_{\ell=1}^{L+1} [I(w_i = \ell)x_{i\ell} - I(w_i \geq \ell)G(x_{i\ell}^\top \beta)x_{i\ell}].$$

b. the Fisher information matrix is

$$\mathcal{I}(\beta) = \mathbf{X}^\top D_\beta \mathbf{X}, \text{ with } D_\beta = \text{Diag} \left\{ g(x_{i\ell}^\top \beta) \prod_{b=1}^{\ell-1} [1 - G(x_{i\ell}^\top \beta)] : (i, \ell) \in \mathcal{J} \right\}.$$

Using Result 3, maximum likelihood estimates (MLEs) for a given data set w_1, \dots, w_n and \mathbf{X} can be computed using scoring iterations

$$\beta^{(r+1)} = \beta^{(r)} + [\mathcal{I}(\beta^{(r)})]^{-1} S(\beta^{(r)}), \quad r = 1, 2, \dots,$$

starting from a given $\beta^{(0)}$ and continuing until an acceptable convergence criterion has been reached. It is possible, however, to recode CRL data as a logistic regression to facilitate computations. The observed w_i can be recoded as L binary variables (y_{i1}, \dots, y_{iL}) with

$$y_{i\ell} = \begin{cases} 1 & \text{if } \ell = w_i, \\ 0 & \text{if } \ell < w_i, \\ \text{NA} & \text{if } \ell > w_i, \end{cases} \tag{8}$$

so that Equation (5) can be rewritten as

$$\mathcal{L}(\beta) = \prod_{i=1}^n \prod_{\ell=1}^L \left\{ G(x_{i\ell}^\top \beta)^{y_{i\ell}} [1 - G(x_{i\ell}^\top \beta)]^{1-y_{i\ell}} \right\}^{I(y_{i\ell} \neq \text{NA})}, \tag{9}$$

where NA values are treated as missing values and excluded from the likelihood. Standard software packages, such as the glm function in R (R Core Team 2021) or PROC GENMOD

in SAS (SAS Institute Inc. 2018), can then be used to fit Equation (9) via the logistic regression

$$Y_{i\ell} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p_{i\ell}), \quad \text{logit}(p_{i\ell}) = x_{i\ell}^\top \beta, \quad \ell \in \{1, \dots, L\} \text{ and } i \in \{1, \dots, n\},$$

and obtain the MLE $\hat{\beta}$ for the CRL model. This is only a computational device: y_{i1}, \dots, y_{iL} are clearly not independent from Equation (8). Also note that a common β is used to estimate all trial probabilities $p_{i\ell}$. Standard software packages also produce a Hessian $\mathbf{H}(\hat{\beta})$, from which $-\mathbf{H}(\hat{\beta})$ and $-\mathbf{H}^{-1}(\hat{\beta})$ can serve as an estimate of $\mathcal{I}(\hat{\beta})$ and $\text{Var}(\hat{\beta})$, respectively, evaluated at $\hat{\beta}$. In a basic logistic regression setting, the Hessian is equivalent to the information matrix and does not depend on the y values (e.g., Agresti 2013, chap. 5). The logistic regression here, however, is carried out conditionally on $\{y_{i\ell}: y_{i\ell} \neq \text{NA}\}$ so that, in general, $\mathbf{H}(\hat{\beta})$ is not equal to $\mathcal{I}(\hat{\beta})$ computed by the CRL information matrix.

4. Hypothesis Testing and Sample Size Calculation

This section describes the hypothesis of interest, the assumed test procedure, and a method to conservatively compute power under a given amount of deviation from the null hypothesis. To handle a variety of testing problems that may arise in experiments, we will assume a general linear hypothesis setting (e.g., Myers 2000, chap. 3). Given a matrix $\mathbf{C} \in \mathbb{R}^{q \times d}$ with rank $q \leq d$ and vector $\mathbf{c}_0 \in \mathbb{R}^q$, consider the hypotheses

$$H_0 : \mathbf{C}\beta = \mathbf{c}_0 \quad \text{vs.} \quad H_1 : \mathbf{C}\beta \neq \mathbf{c}_0. \quad (10)$$

The matrix \mathbf{C} and vector \mathbf{c}_0 are specified by the experimenter to complete the formulation of Equation (10) and kept fixed throughout sample size calculation. Usually \mathbf{c}_0 will be taken as $\mathbf{0}$, but there is no additional difficulty in letting it be an arbitrary vector. A Wald test Equation (10) with significance level α is

$$\text{Reject } H_0 \text{ if } \mathcal{T} > \chi_q^2(1 - \alpha), \quad \text{where } \mathcal{T} = (\mathbf{C}\hat{\beta} - \mathbf{c}_0)^\top (\mathbf{C}\mathcal{I}^{-1}(\hat{\beta})\mathbf{C}^\top)^{-1} (\mathbf{C}\hat{\beta} - \mathbf{c}_0)$$

and $\chi_q^2(1 - \alpha)$ is the $1 - \alpha$ quantile of a chi-square distribution with q degrees of freedom. Let $V \sim N(\mu, \Sigma)$ denote that a random variable V has a multivariate normal distribution with mean μ variance Σ . For large samples, we approximately have that $\hat{\beta} \sim N(\beta, \mathcal{I}^{-1}(\beta))$, so that $(\mathbf{C}, \mathcal{I}^{-1}(\beta)\mathbf{C}^\top)^{-1/2} (\mathbf{C}\hat{\beta} - \mathbf{c}_0) \sim N(\lambda(\beta), \mathbf{I})$ with $\lambda(\beta) = (\mathbf{C}\mathcal{I}^{-1}(\beta)\mathbf{C}^\top)^{-1/2} (\mathbf{C}\beta - \mathbf{c}_0)$. This implies \mathcal{T} is distributed as a non-central chi-square with q degrees of freedom and non-centrality parameter $\psi(\beta) = \lambda(\beta)^\top \lambda(\beta) = (\mathbf{C}\beta - \mathbf{c}_0)^\top (\mathbf{C}\mathcal{I}^{-1}(\beta)\mathbf{C}^\top)^{-1} (\mathbf{C}\beta - \mathbf{c}_0)$. Let $F_{\mathcal{T}}(w; q, \psi)$ denote the cumulative distribution function (cdf) of this distribution. The power function of the test, which will be denoted ϖ , is then approximately

$$\varpi(\beta) = \mathbf{P}_\beta(\mathcal{T} > \chi_q^2(1 - \alpha)) = 1 - F_{\mathcal{T}}(\chi_q^2(1 - \alpha); q, \psi(\beta)). \quad (11)$$

Notice that Equation (11) takes the value $1 - \alpha$ when $\mathbf{C}\beta = \mathbf{c}_0$ so that the test has the desired significance level α under H_0 . The function $F_{\mathcal{T}}$ is readily computed using standard statistical software. When Equation (11) is suitable to approximate the power of the test, we can avoid more computationally demanding methods such as simulation to compute power empirically. This may be especially convenient in power studies, which we are ultimately looking to carry out, where power must be calculated many times. Expression

(11) was obtained using informal arguments; [Cordeiro et al. \(1994\)](#) provide a more rigorous justification under the closely-related setting of GLMs with $C = (I_q 0_{q \times (d-q)})$.

We make several remarks before proceeding. Although the non-centrality parameter $\psi(\beta)$ can be directly chosen to satisfy a given power ϖ , our purpose is to study ϖ through $\psi(\beta)$ as a function of the sample size. Also, $\psi(\beta)$ is not only a function of $C\beta - c_0$, but also depends on the entire vector β through $\mathcal{I}(\beta)$. The parameter of interest is $c = C\beta \in \mathbb{R}^q$, so that the number of parameters not involved in the hypothesis is $d_0 = d - q$. Let B be a $d_0 \times d$ matrix so that $B\beta$ represents the nuisance parameter whose value, say $B\beta = b_0$, is assumed to be given a priori. For example, b_0 may be available from a pilot study or studied via simulation. Furthermore, suppose $A = (B^T C)^T$ is a $d \times d$ nonsingular matrix. Note that both C and B are determined during the setup of the test, and do not depend on unknown parameters. The choice of B is not unique, and may be constructed from any d_0 vectors in \mathbb{R}^d which yield a nonsingular A . Some examples of B are indicated in Remark 5.1 in the context of the application taken up in Section 5. An experimenter may prefer to study power as a function of a scalar effect size rather than the d -dimensional β ; to this end, H_1 may be partitioned into subsets $S(c_0, \Delta) = \{\beta \in \mathbb{R}^d : \|C\beta - c_0\| = \Delta, B\beta = b_0\}$ characterized by effect size $\Delta > 0$. Each $S(c_0, \Delta)$ represents a set of β for which the power ϖ may vary. In view of these remarks, we shall proceed as follows. Given a fixed effect size $\Delta = \|C\beta - c_0\|$, we find the value $\tilde{\beta}$ of β which solves the optimization problem,

$$\text{minimize } \psi(\beta) = (C\beta - c_0)^T (C\mathcal{I}^{-1}(\beta)C^T)^{-1} (C\beta - c_0) \text{ subject to } \beta \in S(c_0, \Delta), \tag{12}$$

and evaluate the power at $\psi(\tilde{\beta})$ via Equation (11). Other options are possible, such as drawing β randomly from $S(c_0, \Delta)$ and evaluating an average or quantile of attained power values, but we will make use of the optimization in Equation (12) for the remainder of the article to ensure that the power calculation is conservative.

The constrained minimization problem in Equation (12) can be transformed to an unconstrained problem and solved using standard optimization software such as `optim` in R ([R Core Team 2021](#)); to do this, we proceed as follows. We can express $c = (c_1, \dots, c_q)$ using spherical coordinates (e.g., [Blumenson 1960](#)) as

$$\begin{aligned} c_1 &= c_{01} + \Delta \cos \phi_1, \\ c_2 &= c_{02} + \Delta \cos \phi_2 \sin \phi_1, \\ &\vdots \\ c_{q-1} &= c_{0,q-1} + \Delta \cos \phi_{q-1} \prod_{j=1}^{q-2} \sin \phi_j, \\ c_q &= c_{0q} + \Delta \sin \phi_{q-1} \prod_{\ell=1}^{q-2} \sin \phi_\ell, \end{aligned}$$

based on $\phi = (\phi_1, \dots, \phi_{q-1})$, where $\phi_j \in [0, \pi]$ for $j = 1, \dots, q - 2$ and $\phi_{q-1} \in [0, 2\pi)$. Here, $\pi = 3.14159\dots$ refers to the mathematical constant, not to be confused with (2). The scaling by Δ and shift by $c_0 = (c_{01}, \dots, c_{0q})$ ensures that c is an element on the

corresponding sphere. A second transformation $\phi_j = \pi G(\vartheta_j)$ for $j = 1, \dots, q - 2$ and $\phi_{q-1} = 2\pi G(\vartheta_{q-1})$ yields ϕ from an unconstrained $\vartheta \in \mathbb{R}^{q-1}$, where $G(x)$ again denotes the inverse logit function. Therefore, a candidate point $\vartheta \in \mathbb{R}^{q-1}$ from the optimizer is transformed to β via

$$(\mathbf{b}_0, \vartheta) \rightarrow (\mathbf{b}_0, \phi) \rightarrow \alpha = (\mathbf{b}_0, \mathbf{c}) \rightarrow \beta = \mathbf{A}^{-1}\alpha. \tag{13}$$

Such a β may be evaluated by the objective function in Equation (12) with the constraint omitted. Algorithm 1 summarizes the steps we have just described to compute this objective function.

An investigation to study the power of test in Equation (10) and determine sample size can therefore be carried out as follows. Determine samples $\mathcal{J}_1, \dots, \mathcal{J}_m \subseteq \{1, \dots, n\}$ of increasing size which are viable for the experiment. Also determine a grid $\{\Delta_1, \dots, \Delta_r\}$ of effect sizes to consider. For each combination of $\Delta \in \{\Delta_1, \dots, \Delta_r\}$ and $\mathcal{J} \in \{\mathcal{J}_1, \dots, \mathcal{J}_m\}$, solve optimization problem in Equation (12) by minimizing the objective function given in Algorithm 1. This yields $\hat{\beta}$, the corresponding non-centrality parameter $\psi(\hat{\beta})$, and the associated power via Equation (11) for each combination. This process allows the test’s power to be studied as a function of the underlying sample size. A sample size may then be selected to meet testing objectives, or it can be determined that no sample size under consideration meets the objectives.

5. Application

We now apply the sample size methodology from Section 4 to the experiment described in Section 2. First, Subsection 5.1 describes the experiment in notation to concretely connect it to Sections 3 and 4. Subsection 5.2

Algorithm 1. Objective function for minimization problem (12) using transformation (13) to satisfy constraints.

Input: ϑ a vector of dimension $q - 1$.
Input: X a $nL \times d$ matrix.
Input: B a $(d - q) \times d$ matrix.
Input: C a $q \times d$ matrix.
Input: \mathbf{b}_0 a vector of dimension $d - q$.
Input: \mathbf{c}_0 a vector of dimension q .
Input: Δ a nonnegative number.

1: **function** OBJECTIVE($\vartheta|X, B, C, \mathbf{b}_0, \mathbf{c}_0, \Delta$)
2: $\phi_j = \pi G(\vartheta_j)$ for $j = 1, \dots, q - 2$. ▷ Transform to angles.
3: $\phi_{q-1} = 2\pi G(\vartheta_{q-1})$
4: $c_j = c_{0j} + \Delta \cos \phi_j \prod_{\ell=1}^{j-1} \sin \phi_\ell$ for $j = 1, \dots, q - 1$. ▷ To sphere with center \mathbf{c}_0 and radius Δ .
5: $c_q = c_{0q} + \Delta \sin \phi_{q-1} \prod_{\ell=1}^{q-2} \sin \phi_\ell$.
6: $\alpha = (\mathbf{b}_0, c_1, \dots, c_q)$. ▷ To a coefficient in $S(\mathbf{c}_0, \Delta) \subseteq \mathbb{R}^d$.
7: $\mathbf{A} = (\mathbf{B}^\top \mathbf{C}^\top)^\top$.
8: $\beta = \mathbf{A}^{-1}\alpha$

9: Compute $\mathcal{I}(\beta)$ from \mathbf{X} and β using Result 3.3.
 10: return $\psi(\beta) = (\mathbf{C}\beta - \mathbf{c}_0)^\top (\mathbf{C}\mathcal{I}^{-1}(\beta)\mathbf{C}^\top) - 1(\mathbf{C}\beta - \mathbf{c}_0)$.
 11: **end function**

then reports a simulation study which compares the approximate power in Equation (11) to empirical power computed from repeated sampling. Finally, with guidance on when in Equation (11) adequately approximates the actual power, Subsection 5.3 presents a power study relying on Equation (11) for computational efficiency. From here, we can decide whether the fourteen initially selected ACOs will be sufficient to assess the impact of training on response rate.

5.1. Setup

Recall that there are $I = 7$ ACOs assigned to each of the $J = 2$ treatments. Let the index $j = 1$ indicate a control ACO receiving no training, while $j = 2$ indicates an experimental ACO which will receive the training. Let K_{ij} denote the household count `HH_Target` from Equation (1) for the i th ACO receiving the j th treatment. Let $W_{ijk} \sim \text{CRL}_L(\mathbf{p}_{ijk})$ indicate the number of contact attempts needed for a response at the i th ACO, j th treatment, and k th household for $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$, $k \in \{1, \dots, K_{ij}\}$, where $\mathbf{p}_{ijk} = (p_{ijk1}, \dots, p_{ijkL})$ are the associated probabilities of a response at each attempt conditional on failures at previous attempts via Equation (3). Recall that an observation of $w_{ijk} = L + 1$ indicates that no response was obtained in the first L attempts. We consider a basic model for response rate as

$$\text{logit}(p_{ijk\ell}) = \zeta_{j\ell} \tag{14}$$

$$= \mu + \tau_j + \delta_\ell + (\tau\delta)_{j\ell} \tag{15}$$

$$= \mathbf{s}_{j\ell}^\top \beta \tag{16}$$

Model formulation (14) uses unconstrained effects $\zeta_{11}, \dots, \zeta_{JL}$ to facilitate computations. Formulation (15) provides a more clear interpretation, with an intercept term μ , treatment effects τ_j which are of primary interest, contact attempt effects δ_ℓ , and effects $(\tau\delta)_{j\ell}$ for treatment-attempt interaction. Formulation (16) is a regression form of Equation (14) which can more readily be connected to Sections 3 and 4. To reparameterize from Equations (14) to (15), we assume constraints

$$\sum_{j=1}^J \tau_j = 0, \quad \sum_{\ell=1}^L \delta_\ell = 0, \quad \sum_{j=1}^J (\tau\delta)_{j\ell} = 0, \quad \sum_{\ell=1}^L (\tau\delta)_{j\ell} = 0, \tag{17}$$

and let $\zeta_{j\ell} = \mu + \tau_j + \delta_\ell + (\tau\delta)_{j\ell}$ so that

$$\frac{1}{JL} \sum_{j=1}^J \sum_{\ell=1}^L \zeta_{j\ell} = \mu, \quad \frac{1}{L} \sum_{\ell=1}^L \zeta_{j\ell} - \mu = \tau_j, \quad \frac{1}{J} \sum_{j=1}^J \zeta_{j\ell} - \mu = \delta_\ell.$$

Care should be taken when interpreting μ , τ_j , and δ_ℓ as they are averages of the raw $\zeta_{j\ell}$ parameters. There are $J - 1$ distinct parameters among the τ_j 's, $L - 1$ among the δ_ℓ 's, and

$(J - 1)(L - 1)$ among the $(\tau\delta)_{j\ell}$'s; with the addition of μ , there are a total of $(J - 1) + (L - 1) + (J - 1)(L - 1) + 1 = JL$ parameters. In particular, JL is equivalent to $2L$ with $J = 2$ treatments. To rewrite Equation (15) in the form Equation (16), let

$$\beta = (\mu, \tau_1, \delta_1, \dots, \delta_{L-1}, (\tau\delta)_{11}, \dots, (\tau\delta)_{1,L-1})$$

with $s_{j\ell}$ coded in the manner shown in Table 2. To emphasize the grouping of trials implied by the model, let $\mathcal{H}(j, \ell)$ represent the list of (i, j, k, ℓ) indices corresponding to the j th treatment and ℓ th attempt, so that $\mathcal{H}(j, \ell)$ contains $N_{j\ell} = L \sum_{i=1}^L K_{ij}$ elements, and write $\mathbf{p}_{\mathcal{H}(j,\ell)} = (p_{ijk\ell} : (i, j, k, \ell) \in \mathcal{H}(j, \ell))$. We can then rewrite Equation (16) as

$$\text{logit}(\mathbf{p}_{\mathcal{H}(j,\ell)}) = \mathbf{X}_{j\ell}\beta, \quad j = 1, \dots, J \text{ and } \ell = 1, \dots, L,$$

where $\mathbf{X}_{j\ell} = \mathbf{1}_{N_{j\ell}} \otimes s_{j\ell}^\top$ using the Kronecker product and a vector $\mathbf{1}_{N_{j\ell}}$ of $N_{j\ell}$ ones. Sample size determination will be based on a test of the general linear hypothesis (10) with $\mathbf{C} = (\mathbf{0}_{(JL-1) \times 1} \ \mathbf{I}_{JL-1})$ and $\mathbf{c}_0 = \mathbf{0}_{(JL-1) \times 1}$; that is, a test for the presence of any treatment effects, attempt effects, or their interactions. We will assume a significance level of $\alpha = 0.10$ for the test, which is a standard used by the U.S. Census Bureau (U.S. Census Bureau 2013). The choice of $\mathbf{B} = (\mathbf{1}_{(1 \times JL-1)})$ is taken so that $\mathbf{B}\beta = \mu$ is a scalar nuisance parameter. Subsection 5.3 will investigate the relationship between the sample size, the effect size $\Delta = \|\mathbf{C}\beta - \mathbf{c}_0\|$, and power ϖ of the test. Some discussion will be provided to interpret the achieved Δ .

Remark 1.

A number of other tests are possible using the general linear hypothesis in this setting. Note that the alternative hypothesis H_1 in each case is assumed to be the complement of the specified H_0 .

1. A test for interaction effects $H_0 : (\tau\delta)_{11} = \dots = (\tau\delta)_{1,L-1} = 0$ can be carried out using $\mathbf{c}_0 = \mathbf{0}$ and the $(L - 1) \times 2L$ matrix

$$\mathbf{C} = (0_{(L-1) \times (L+1)} \ \mathbf{I}_{L-1})$$

so that $\mathbf{C}\beta = ((\tau\delta)_{11}, \dots, (\tau\delta)_{1,L-1})$. One straightforward choice of a corresponding B would be the $(L + 1) \times 2L$ matrix $B = (I_{L+1} \ 0_{(L+1) \times (L - 1)})$ so that the nuisance parameter becomes $\mathbf{B}\beta = (\mu, \tau_1, \delta_1, \dots, \delta_{L-1})$.

Table 2. Choice of coding for each row of the design matrix \mathbf{X} for the regression, denoted $s_{j\ell}$; in Equation (16).

j	ℓ	Intercept	Treatment	Attempt	Treatment \times attempt
1	1	1	1	1	0
	2	1	1	0	1
	\vdots	\vdots	\vdots	\vdots	\ddots
	$L-1$	1	1	0	0
1	L	1	1	-1	-1
2	1	1	-1	1	0
	2	1	-1	0	1
	\vdots	\vdots	\vdots	\vdots	\ddots
	$L-1$	1	-1	0	0
2	L	1	-1	-1	-1

2. Simultaneously testing for treatment main effects and treatment-attempt interactions $H_0 : \tau_1 = (\tau\delta)_{11} = \dots = (\tau\delta)_{1,L-1} = 0$ can be accomplished using $c_0 = \mathbf{0}$ and the $L \times 2L$ matrix

$$C = \begin{pmatrix} 0 & 1 & 0_{1 \times (L-1)} & 0_{1 \times (L-1)} \\ 0_{(L-1) \times 1} & 0_{(L-1) \times 1} & 0_{(L-1) \times (L-1)} & I_{L-1} \end{pmatrix}$$

so that $C\beta = (\tau_1, (\tau\delta)_{11}, \dots, (\tau\delta)_{1,L-1})$. Here, taking B as the $L \times 2L$ matrix

$$B = \begin{pmatrix} 1 & 0 & 0_{1 \times (L-1)} & 0_{1 \times (L-1)} \\ 0_{(L-1) \times 1} & 0_{(L-1) \times 1} & I_{L-1} & 0_{(L-1) \times (L-1)} \end{pmatrix}$$

leads to nuisance parameter $B\beta = (\mu, \delta_1, \dots, \delta_{L-1})$.

3. Simultaneously testing for equality among main effects for the attempt and presence of treatment-attempt interactions $H_0 : \delta_1 = \dots = \delta_{L-1}$ and $(\tau\delta)_{11} = \dots = (\tau\delta)_{1,L-1} = 0$ can be accomplished using $c_0 = 0$ and the $(2L - 3) \times 2L$ matrix

$$C = \begin{pmatrix} 0_{(L-2) \times 2} & C_1 & 0_{(L-2) \times (L-1)} \\ 0_{(L-1) \times 2} & 0_{(L-1) \times (L-1)} & I_{L-1} \end{pmatrix}, \text{ where}$$

$$C_1 = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

has $L - 2$ rows and $L - 1$ columns, so that $C\beta = (\delta_1 - \delta_2, \dots, \delta_{L-2} - \delta_{L-1}, (\tau\delta)_{11}, \dots, (\tau\delta)_{1,L-1})$. The choice of $B = (I_3 \ 0_{3 \times (2L - 3)})$ yields the nuisance parameter $B\beta = (\mu, \tau, \delta_1)$.

The matrix $A = (B^T C^T)$ is nonsingular in all three cases. In Cases 1 and 2, A is a $d \times d$ identity matrix and a permutation of the identity matrix, respectively. The inverse A^{-1} can be constructed in Case 3 by noting that for $\alpha = A\beta = (\mu, \tau_1, \delta_1, \delta_1 - \delta_2, \dots, \delta_{L-2} - \delta_{L-1}, (\tau\delta)_{11}, \dots, (\tau\delta)_{1,L-1})$, the elements of β are recovered using $\mu = \alpha_1$, $\tau_1 = \alpha_2$, $\delta_1 = \alpha_3$, $\delta_\ell = \alpha_3 - \sum_{b=4}^{\ell+2} \alpha_b$ for $\ell = 2, \dots, L - 1$, and $(\tau\delta)_{1\ell} = \alpha_{\ell+L+1}$ for $\ell = 1, \dots, L - 1$.

Remark 2.

All independent variables in the illustration are constructed from (categorical) experimental factors, but this is not required in general. Other covariate data may also be used if available in advance of data collection. For example, suppose a strong predictor t_{ij} could be used to adjust the baseline response probability and was available for each ACO during planning. One source of data which might serve this purpose is the Low Response Score in the Planning Database (Erdman and Bates 2017). Here, model (15) could be extended to $\text{logit}(p_{ijk\ell}) = \mu + \tau_j + \delta_\ell + (\tau\delta)_{j\ell} + t_{ij}\psi$ to adjust for the variability in the outcome using coefficient $\psi \in \mathbb{R}$. The regression framework on the logit scale may be used to capture such covariates in addition to experimental factors.

Remark 3.

The CRL regression model assumes that independent variables x_{i1}, \dots, x_{iL} are fixed during the entire process in which response W_i is generated. Independent variables may vary with the attempt, as we have done in this section, but cannot depend on additional data collected during the sequence of trials. This corresponds to studies which are planned in advance and not altered during the course of data collection. In contrast, work on adaptive designs seeks to adjust contact strategies during an operation for purposes such as reducing operational costs or reducing burden to respondents (e.g., [Ashmead et al. 2017](#)). This can be aided by paradata collected while attempting to contact respondents, such as the nature of previous failures (e.g., a refusal to participate or a failure to make any contact). Here, binary regression models which evolve over time and allow time-varying independent variables, such as in [Slud and Kedem \(1994\)](#), might be considered over the CRL model. The adaptive design setting will not be considered further in this article, but is a topic of interest for future work.

It is important to consider the number of contact attempts L to be used in the model. Too few contact attempts can fail to capture the response behavior of interest, while too many will lead to an issue of sparse observations which we will now discuss. Although a high probability of response during each contact attempt is desirable from the perspective of data collection, enumerations during later attempts will be a more rare occurrence. In turn, corresponding counts will be close to zero, large sample properties used in Section 4 will not take effect, and consequently the power expression (11) will be inaccurate unless sample sizes are taken to be very large. To make this issue concrete, suppose H_0 is true so that the probability of a successful enumeration $p_{ijk\ell} \equiv p$, given that any attempts $1, \dots, \ell - 1$ failed, depends only on μ . We may then write the overall (unconditional) probabilities of enumeration as $\pi_{ijk\ell} = p \prod_{b=1}^{\ell-1} (1-p) = p(1-p)^{\ell-1}$. Values for $\pi_{ijk} = (\pi_{ijk1}, \dots, \pi_{ijkL})$ are shown in [Table 3](#) for $L = 5$ for several values of p under H_0 . It is clear that responses occurring after two attempts are quite common under small p but become increasingly rare events when p approaches 1. In practice, many factors can influence response probability across attempts, but consideration of the model under H_0 helps to serve as a guideline.

5.2. Simulation

[Table 3](#) emphasized that successful enumerations in later attempts can be quite rare in some circumstances: in particular, under H_0 when the baseline probability of success p for each attempt approaches 1. It is anticipated that large sample approximations used in Section 4 will fail when data in later categories become too uncommon. In this section, we will compare the empirical power of the Wald test to the approximate power computed via Equation (11). A simulation is carried out in R ([R Core Team 2021](#)) under the experimental design introduced in Subsection 5.1. When suitable, the approximation is greatly preferred to empirical power because it greatly facilitates computation of a power study such as the one given in Subsection 5.3.

Suppose there is $I = 1$ experimental ACO with K households and one control ACO with K households so that there are $J = 2$ treatments. We take $K \in \{10, 50, 200\}$ and consider CRL models of the form (15) which include $L \in \{1, 2, 3, 4\}$ attempts. For the baseline

Table 3. Probabilities $\pi_{ijk\ell}$ of a successful enumeration for attempts $\ell = 1, \dots, 5$ assuming constant probability p for each attempt. Category 6+ contains the leftover probability that enumeration occurs after attempt 5. This simplified model illustrates that responses occurring in later attempts may be rare.

p	Attempt					
	1	2	3	4	5	6+
0.05	0.05	0.0475	0.0451	0.0429	4.073E-2	7.738E-1
0.10	0.10	0.0900	0.0810	0.0729	6.561E-2	5.905E-1
0.15	0.15	0.1275	0.1084	0.0921	7.830E-2	4.437E-1
0.20	0.20	0.1600	0.1280	0.1024	8.192E-2	3.277E-1
0.25	0.25	0.1875	0.1406	0.1055	7.910E-2	2.373E-1
0.30	0.30	0.2100	0.1470	0.1029	7.203E-2	1.681E-1
0.35	0.35	0.2275	0.1479	0.0961	6.248E-2	1.160E-1
0.40	0.40	0.2400	0.1440	0.0864	5.184E-2	7.876E-2
0.45	0.45	0.2475	0.1361	0.0749	4.118E-2	5.033E-2
0.50	0.50	0.2500	0.1250	0.0625	3.125E-2	3.125E-2
0.55	0.55	0.2475	0.1114	0.0501	2.255E-2	1.845E-2
0.60	0.60	0.2400	0.0960	0.0384	1.536E-2	1.024E-2
0.65	0.65	0.2275	0.0796	0.0279	9.754E-3	5.253E-3
0.70	0.70	0.2100	0.0630	0.0189	5.670E-3	2.430E-3
0.75	0.75	0.1875	0.0469	0.0117	2.930E-3	9.766E-4
0.80	0.80	0.1600	0.0320	0.0064	1.280E-3	3.200E-4
0.85	0.85	0.1275	0.0191	0.0029	4.303E-4	7.594E-5
0.90	0.90	0.0900	0.0090	0.0009	9.000E-5	1.000E-5
0.95	0.95	0.0475	0.0024	0.0001	5.938E-6	3.125E-7

effect, we take $\text{logit}^{-1}(\mu) \in \{0.60, 0.75, 0.90\}$. For the departure from H_0 , we consider $\Delta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1\}$. Here, we explicitly choose the parameters to be

$$\beta = (\mu, \tau_1 = \Delta, \delta_1 = 0, \dots, \delta_{L-1} = 0, (\tau\delta)_{11} = 0, \dots, (\tau\delta)_{1,L-1} = 0).$$

so that Δ is entirely allocated to τ_1 . The simulation proceeds by drawing a sample $W_{ijk} \sim \text{CRL}_L(\mathbf{p}_{ijk})$ for $i = 1, j \in \{1, 2\}$, and $k \in \{1, \dots, K\}$, recoding W_{ijk} 's to $Y_{ijk\ell}$'s via (8), then fitting the (correctly specified) data-generating model (15) by a logistic regression with the R `glm` function. This is repeated $R = 1,000$ times for each simulation setting, yielding coefficient estimates $\hat{\beta}^{(r)}$ and corresponding covariance estimates $\hat{V}^{(r)} = \mathcal{I}^{-1}(\hat{\beta}^{(r)})$ for $r = 1, \dots, R$. We then compute Wald statistics

$$T^{(r)} = (\mathbf{C}\hat{\beta}^{(r)} - \mathbf{c}_0)^\top (\mathbf{C}\hat{V}^{(r)}\mathbf{C}^\top)^{-1} (\mathbf{C}\hat{\beta}^{(r)} - \mathbf{c}_0),$$

to obtain an empirical probability of rejection $\frac{1}{R} \sum_{r=1}^R I(T^{(r)} \geq \chi_q^2(1 - \alpha))$. Here, $\chi_q^2(1 - \alpha)$ denotes the $1 - \alpha = 0.90$ quantile of the χ^2 distribution with $q = JL - 1$ degrees of freedom, which is the critical value of the test. For some repetitions, the coefficients or the associated covariance estimates could not be fully computed. For example, this occurred when no outcomes were observed for an attempt ℓ in one or both of the treatments. These cases were recorded as $T^{(r)} = \text{NA}$ and excluded from the empirical power calculation. The approximate rejection probability in Equation (11) is also computed for each simulation setting; note that this does not make use of the simulation draws.

Tables 4, 5, 6, and 7 display the empirical power and approximated power using $L = 1, \dots, 4$, respectively, after carrying out the simulation. Respective entries across the two tables can be compared to check their agreement. Table 8 displays frequencies of $T^{(r)} = \text{NA}$ from the empirical power calculation; for example, a count of zero indicates that all samples in the given setting could be estimated. When $L = 1$, the empirical and approximate power closely agree when $\mu = \text{logit}(0.6)$, for all sample sizes K and all Δ .

Table 4. Empirical power (E) computed by simulation versus approximate power (A) computed via Equation (11) with $L = 1$ maximum attempts.

Power	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
E	10	0.60	0.1090	0.1020	0.1100	0.1840	0.1950	0.2970	0.4480	0.5810
		0.75	0.0450	0.0440	0.0670	0.0580	0.0860	0.1110	0.2200	0.2820
		0.90	0.0010	0.0020	0.0000	0.0020	0.0020	0.0050	0.0070	0.0240
E	50	0.60	0.0900	0.1280	0.2470	0.4290	0.6230	0.7710	0.9800	0.9990
		0.75	0.0940	0.1270	0.2300	0.3410	0.5420	0.6830	0.9450	0.9950
		0.90	0.0590	0.0800	0.1220	0.1550	0.2670	0.3600	0.6380	0.7730
E	200	0.60	0.0900	0.2520	0.6450	0.9030	0.9820	1.0000	1.0000	1.0000
		0.75	0.1160	0.2290	0.5380	0.8090	0.9680	0.9980	1.0000	1.0000
		0.90	0.0840	0.1520	0.3320	0.5700	0.7720	0.9160	0.9960	1.0000
A	10	0.60	0.1000	0.1081	0.1321	0.1707	0.2220	0.2830	0.4551	0.6142
		0.75	0.1000	0.1063	0.1250	0.1552	0.1951	0.2429	0.3796	0.5118
		0.90	0.1000	0.1030	0.1120	0.1264	0.1455	0.1683	0.2345	0.3013
A	50	0.60	0.1000	0.1403	0.2558	0.4248	0.6084	0.7665	0.9622	0.9963
		0.75	0.1000	0.1315	0.2226	0.3597	0.5183	0.6697	0.9098	0.9820
		0.90	0.1000	0.1152	0.1595	0.2290	0.3171	0.4149	0.6461	0.8025
A	200	0.60	0.1000	0.2570	0.6198	0.8963	0.9859	0.9990	1.0000	1.0000
		0.75	0.1000	0.2237	0.5308	0.8205	0.9586	0.9942	1.0000	1.0000
		0.90	0.1000	0.1602	0.3270	0.5491	0.7515	0.8868	0.9917	0.9996

Table 5. Empirical power (E) computed by simulation versus approximate power (A) computed via Equation (11) with $L = 2$ maximum attempts.

Power	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
E	10	0.60	0.0071	0.0213	0.0163	0.0123	0.0392	0.0484	0.1065	0.2326
		0.75	0.0034	0.0000	0.0022	0.0046	0.0082	0.0120	0.0294	0.0656
		0.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0034
E	50	0.60	0.0790	0.1290	0.2050	0.3670	0.5640	0.7400	0.9770	1.0000
		0.75	0.0580	0.0820	0.1270	0.2170	0.3650	0.5200	0.8880	0.9900
		0.90	0.0283	0.0263	0.0276	0.0443	0.0730	0.1313	0.2913	0.6145
E	200	0.60	0.0950	0.2190	0.5760	0.9120	0.9910	1.0000	1.0000	1.0000
		0.75	0.0970	0.1890	0.4680	0.7810	0.9520	0.9940	1.0000	1.0000
		0.90	0.0540	0.0870	0.1780	0.3290	0.5830	0.7630	0.9870	1.0000
A	10	0.60	0.1000	0.1061	0.1246	0.1554	0.1980	0.2510	0.4117	0.5713
		0.75	0.1000	0.1043	0.1170	0.1381	0.1669	0.2025	0.3111	0.4245
		0.90	0.1000	0.1018	0.1071	0.1158	0.1274	0.1415	0.1838	0.2284
A	50	0.60	0.1000	0.1312	0.2281	0.3871	0.5776	0.7514	0.9656	0.9975
		0.75	0.1000	0.1216	0.1881	0.2991	0.4428	0.5953	0.8735	0.9713
		0.90	0.1000	0.1090	0.1363	0.1814	0.2429	0.3169	0.5192	0.6857
A	200	0.60	0.1000	0.2293	0.5922	0.8993	0.9899	0.9996	1.0000	1.0000
		0.75	0.1000	0.1891	0.4572	0.7703	0.9442	0.9922	1.0000	1.0000
		0.90	0.1000	0.1368	0.2508	0.4325	0.6358	0.8036	0.9776	0.9983

Table 6. Empirical power (E) computed by simulation versus approximate power (A) computed via Equation (11) with $L = 3$ maximum attempts.

Power	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
E	10	0.60	0.0000	0.0000	0.0000	0.0000	0.0017	0.0017	0.0068	0.0408
		0.75	0.0000	0.0000	0.0000	0.0000	0.0045	0.0000	0.0000	0.0088
		0.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
E	50	0.60	0.0430	0.0631	0.1474	0.2590	0.4789	0.6839	0.9621	1.0000
		0.75	0.0087	0.0294	0.0496	0.0938	0.1553	0.3211	0.7169	0.9594
		0.90	0.0351	0.0272	0.0210	0.0210	0.0000	0.0227	0.0962	0.2537
E	200	0.60	0.0900	0.2080	0.5580	0.8750	0.9930	1.0000	1.0000	1.0000
		0.75	0.0480	0.1260	0.3490	0.6770	0.9179	0.9890	1.0000	1.0000
		0.90	0.0384	0.0369	0.0865	0.1832	0.3450	0.5611	0.9657	1.0000
A	10	0.60	0.1000	0.1051	0.1206	0.1467	0.1833	0.2298	0.3759	0.5286
		0.75	0.1000	0.1034	0.1134	0.1301	0.1531	0.1820	0.2725	0.3711
		0.90	0.1000	0.1014	0.1054	0.1119	0.1207	0.1315	0.1642	0.1991
A	50	0.60	0.1000	0.1261	0.2100	0.3550	0.5399	0.7193	0.9581	0.9967
		0.75	0.1000	0.1170	0.1704	0.2630	0.3899	0.5338	0.8304	0.9550
		0.90	0.1000	0.1068	0.1275	0.1623	0.2109	0.2711	0.4475	0.6083
A	200	0.60	0.1000	0.2111	0.5556	0.8831	0.9881	0.9995	1.0000	1.0000
		0.75	0.1000	0.1712	0.4035	0.7153	0.9197	0.9869	1.0000	1.0000
		0.90	0.1000	0.1279	0.2172	0.3698	0.5587	0.7351	0.9590	0.9959

Table 7. Empirical power (E) computed by simulation versus approximate power (A) computed via Equation (11) with $L = 4$ maximum attempts. A dash (—) indicates that no samples in this setting yielded valid estimates of the coefficients and their variances.

Power	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
E	10	0.60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		0.75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		0.90	—	—	—	—	—	—	0.0000	—
E	50	0.60	0.0150	0.0352	0.0506	0.1226	0.2513	0.4449	0.8755	0.9909
		0.75	0.0033	0.0071	0.0206	0.0326	0.1188	0.1741	0.5946	0.8889
		0.90	0.0000	0.0000	0.0000	0.0000	0.0000	—	0.0000	1.0000
E	200	0.60	0.0810	0.1650	0.5055	0.8660	0.9850	0.9990	1.0000	1.0000
		0.75	0.0242	0.0631	0.2335	0.5271	0.8302	0.9548	1.0000	1.0000
		0.90	0.1800	0.1282	0.0645	0.2250	0.3333	0.6129	1.0000	1.0000
A	10	0.60	0.1000	0.1044	0.1177	0.1403	0.1723	0.2133	0.3457	0.4900
		0.75	0.1000	0.1028	0.1112	0.1252	0.1446	0.1690	0.2471	0.3345
		0.90	0.1000	0.1011	0.1044	0.1099	0.1172	0.1261	0.1533	0.1827
A	50	0.60	0.1000	0.1225	0.1958	0.3266	0.5015	0.6814	0.9458	0.9952
		0.75	0.1000	0.1142	0.1592	0.2387	0.3513	0.4850	0.7883	0.9360
		0.90	0.1000	0.1056	0.1228	0.1518	0.1926	0.2440	0.4005	0.5525
A	200	0.60	0.1000	0.1968	0.5169	0.8584	0.9836	0.9993	1.0000	1.0000
		0.75	0.1000	0.1599	0.3636	0.6647	0.8916	0.9793	1.0000	1.0000
		0.90	0.1000	0.1231	0.1980	0.3303	0.5044	0.6804	0.9385	0.9924

When μ is increased to $\text{logit}(0.75)$, $K = 10$ becomes too small, and the empirical power is systematically smaller than the approximation. For this value of μ , $K = 50$ appears to be a sufficient number of households. When we further increase μ to $\text{logit}(0.9)$, $K = 50$ is no longer sufficient, but increasing to $K = 200$ is enough for the two power calculations to

Table 8. Count of NAs in each simulation setting when calculating empirical power. An NA indicates that estimates could not be computed for a particular simulation repetition; for example, this occurred when no outcomes reached the L th attempt in one of the two treatments.

L	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
1	10	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
	50	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
	200	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
2	10	0.60	11	15	16	26	30	29	61	110
		0.75	109	111	106	132	145	167	252	314
		0.90	576	589	572	587	583	613	649	708
	50	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	4
		0.90	10	10	23	29	28	48	80	144
	200	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
3	10	0.60	303	324	339	359	400	414	557	681
		0.75	804	800	794	773	776	795	853	887
		0.90	995	989	993	985	988	991	986	996
	50	0.60	0	1	3	4	4	13	49	150
		0.75	77	81	91	158	182	265	410	532
		0.90	827	852	855	856	866	868	896	933
	200	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	1	3	24	102
		0.90	244	269	306	356	371	435	592	747
4	10	0.60	772	754	794	789	803	832	892	930
		0.75	979	977	976	985	968	993	983	983
		0.90	1000	1000	1000	1000	1000	999	1000	1000
	50	0.60	69	90	110	168	236	292	510	669
		0.75	697	720	709	785	798	799	889	946
		0.90	993	998	998	997	999	1000	998	999
	200	0.60	0	0	1	0	3	5	59	203
		0.75	91	97	135	207	317	380	625	765
		0.90	950	961	969	960	973	969	980	991

agree. If we increase L to 2, $K = 10$ is no longer a sufficient number of households for any displayed setting of μ . $K = 50$ gives a sufficient power approximation when $\mu = \text{logit}(0.6)$, but not for the two larger values of μ . $K = 200$ is enough when $\mu = \text{logit}(0.6)$ or $\mu = \text{logit}(0.75)$. When $\mu = \text{logit}(0.90)$, however, we need a larger sample to use the approximation reliably.

The pattern becomes more severe as L increases, with larger K needed for a reasonably good approximation of the power for larger μ . Referring to Table 8, we notice that NA counts increase accordingly when L and μ are both larger. For example, in the case of $L = 3$ and $\mu = \text{logit}(0.90)$, it is rare to obtain valid estimates under $K = 10$, but slowly becomes more frequent as the number of households increases to $K = 50$ and to $K = 200$. Referring back to Table 3, we see that attempt 3 for $\mu = \text{logit}(0.90)$ has probability of about 0.009 under H_0 . Therefore, we expect that a sample size of approximately $K = 100$ will be needed to observe third attempts in both treatments, which is a minimum requirement to be able to use a model with $L = 3$.

Note that if H_0 were modified to test effects corresponding only to the L th enumeration attempt, a much larger sample size would be needed in many of these settings to achieve the same power, when such enumerations become rare events.

5.3. Sample Size for Illustration

With some insight into the quality of the approximation (11), we now present a power study using the fourteen ACOs from Table 1. To describe seven increasing sample sizes, let \mathcal{J}_1 contain the experimental and control ACOs with $i = 1$ corresponding to the “pair” column in Table 1, \mathcal{J}_2 contain those with $i \in \{1, 2\}$, . . . , and \mathcal{J}_7 contain those with $i \in \{1, . . . , 7\}$. For each $\mathcal{J} = \mathcal{J}_1, . . . , \mathcal{J}_7$ and each $\Delta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0\}$, the optimization problem in Equation (12) is solved to yield the minimizer $\beta = \tilde{\beta}(\Delta, \mathcal{J})$ and associated power $\varpi(\Delta, \mathcal{J})$. We repeat this using $L \in \{2, . . . , 5\}$ contact attempts and baseline response effect $\text{logit}^{-1}(\mu) \in \{0.75, 0.90\}$. Figure 1 displays the results as a grid of power curves. For this discussion, we will consider $\varpi = 0.80$ as a rough target for the power.

First, we give an upper bound on μ to decide on the largest L that can be supported by the model. Internal discussions with U.S. Census Bureau personnel have suggested that the baseline response probability μ might be larger than $\text{logit}(0.75)$ but should be no greater than $\text{logit}(0.90)$; therefore, Table 3 suggests modeling at most $L = 3$ attempts. With $L = 3$, using all fourteen available ACOs, we achieve nearly $\varpi = 1$ when $\mu = \text{logit}(0.75)$. Under $\mu = \text{logit}(0.90)$, we also achieve $\varpi \approx 1$ except under the smallest effect size in the study, $\Delta = 0.1$, where $\varpi \approx 0.77$ is achieved. Therefore, $\Delta = 0.1$ represents the smallest effect size we can detect using all fourteen available ACOs, modeling $L = 3$ contact attempts, achieving power $\varpi \approx 0.77$, and assuming $\mu = \text{logit}(0.90)$. Stakeholders of the experiment will likely need an intuitive interpretation of $\Delta = 0.1$ to decide if this provides a level of detection precise enough to be practically useful. To assist with interpretation, we can consider the extreme cases of the alternative hypothesis with effect size Δ , namely

$$\beta \in \{(\mu, \Delta, 0, . . . , 0), . . . , (\mu, 0, . . . , 0, \Delta)\}, \tag{18}$$

so that Δ is completely allocated to one of the coordinates of β aside from the intercept. Table 9 shows the $p_{ijk\ell}$ and $\pi_{ijk\ell}$ corresponding to each of the values in Equation (18), along with the value $\beta = (\mu, 0, . . . , 0)$ under H_0 . A comparison of each case (b)–(f) in Table 9 to case (a) suggests that $\Delta = 0.1$ corresponds to rather small changes in probabilities. Presented with this information, stakeholders may determine whether this level of detection is sufficiently precise for the experiment.

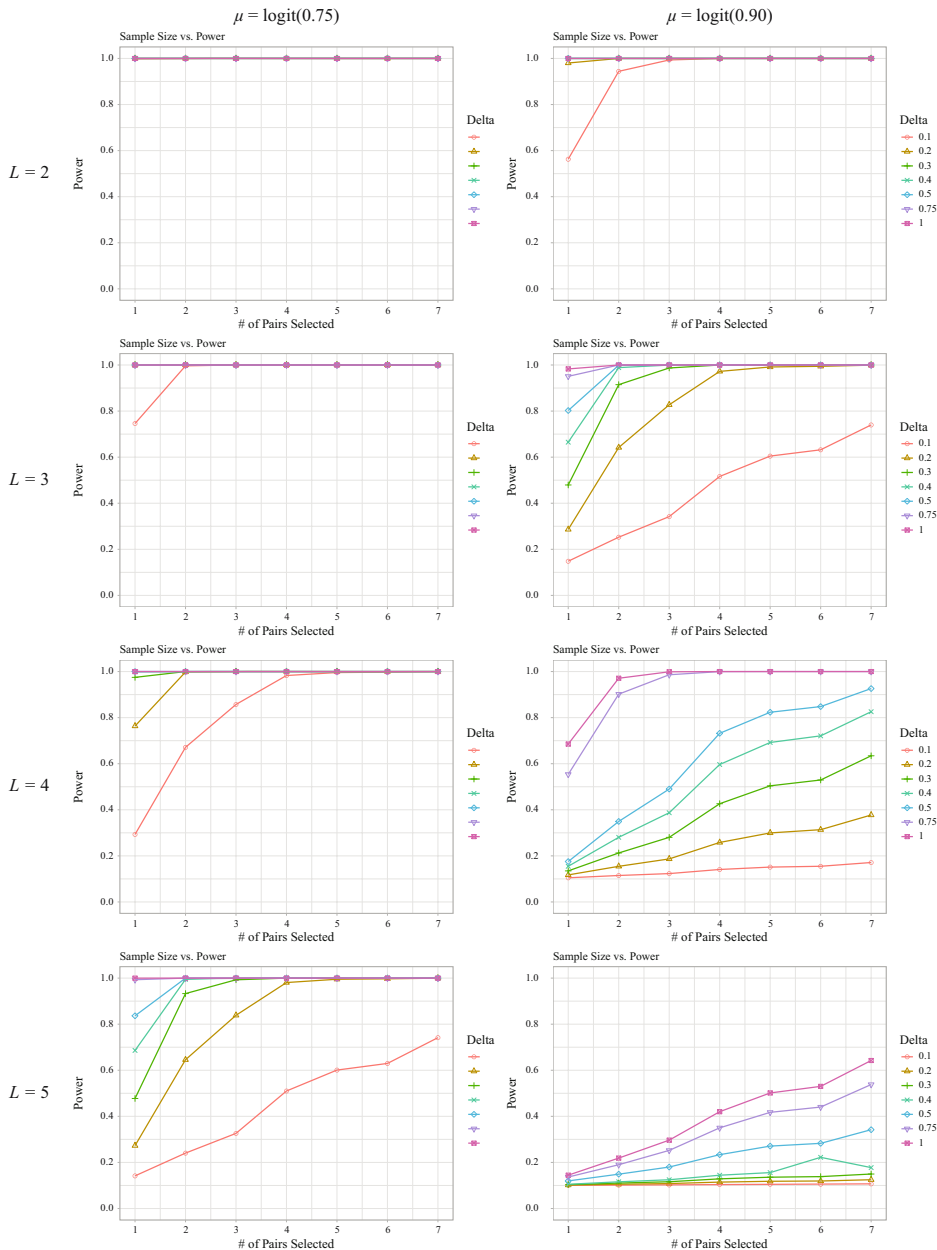


Fig. 1. Power study using the fourteen pre-selected ACOs in Dallas, Houston, and Los Angeles.

6. Discussion and Conclusions

Experiments assessing changes to response rates may involve multiple attempts to establish contact with households, persons, businesses, or other entities. Sequential models such as the continuation-ratio logit (CRL) provide a statistical framework for such experiments. Through an illustration based on an actual experiment for a new enumerator

Table 9. An aid to interpret effect size $\Delta = 0.1$ with baseline effect $\mu = \text{logit}(0.9)$ for probability of response. Case (a) represents the null hypothesis where all other effects are zero, while cases (b)–(f) place all of effect size Δ on one particular coordinate of β . Probabilities of success $P_{ijk1}, P_{ijk2}, P_{ijk3}$ corresponding to enumeration attempts 1, 2, and 3 from case (a) can be compared to each case (b)–(f) to visualize the differences that can be detected by the experiment. Similarly, overall probabilities of enumeration $\pi_{ijk1}, \dots, \pi_{ijk4}$ at attempts 1, ..., 4 from case (a) can be compared to each case (b)–(f). Recall that with $L = 4$, “enumeration at attempt 4” means that enumeration does not occur in the first three attempts.

(a) H_0							
j	P_{ijk1}	P_{ijk2}	P_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9000	0.9000	0.9000	0.9000	0.0900	0.0090	0.0010
2	0.9000	0.9000	0.9000	0.9000	0.0900	0.0090	0.0010
(b) $\tau_1 = \Delta$							
j	P_{ijk1}	P_{ijk2}	P_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9086	0.9086	0.9086	0.9086	0.0830	0.00758	0.000762
2	0.8906	0.8906	0.8906	0.8906	0.0974	0.01065	0.001308
(c) $\delta_1 = \Delta$							
j	P_{ijk1}	P_{ijk2}	P_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9086	0.9000	0.8906	0.9086	0.0822	0.00814	0.000999
2	0.9086	0.9000	0.8906	0.9086	0.0822	0.00814	0.000999
(d) $\delta_2 = \Delta$							
j	P_{ijk1}	P_{ijk2}	P_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9000	0.9086	0.8906	0.9000	0.0909	0.00814	0.000999
2	0.9000	0.9086	0.8906	0.9000	0.0909	0.00814	0.000999
(e) $(\tau\delta)_{11} = \Delta$							
j	P_{ijk1}	P_{ijk2}	P_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9086	0.9000	0.8906	0.9086	0.0822	0.00814	0.000999
2	0.8906	0.9000	0.9086	0.8906	0.0984	0.00994	0.000999
(f) $(\tau\delta)_{12} = \Delta$							
j	P_{ijk1}	P_{ijk2}	P_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9000	0.9086	0.8906	0.9000	0.0909	0.00814	0.000999
2	0.9000	0.8906	0.9086	0.9000	0.0891	0.00994	0.000999

training module, we have explored use of the CRL model in an experimental design to measure changes in response rates. The presented methodology was used to justify a sample size and provide intuition on effect sizes which could be detected in the experiment with a desired level of power. We focused on an experiment which would be carried out within the decennial census, but experiments carried out in a number of other official statistics settings, including surveys and operational tests which occur between decennial censuses (e.g., [National Research Council 2010](#)), could involve measuring response rates over multiple contacts.

Note that a complete run of the simulation in Subsection 5.2 took 28.4 minutes to complete on an Intel Core i7–2600 3.40 GHz workstation with four CPU cores (without explicitly running tasks in parallel); of this, 28.3 minutes was used for empirical power computation. This emphasizes the benefit of computing the power via Equation (11) in settings where it must be repeated many times, provided that the approximation holds. The power study presented in Section 4 used only the approximation and required slightly over seven days of total computing time on an enterprise server, enlisting multiple CPUs to reduce elapsed time; this study would have become impractical to carry out using the empirical power computation.

A likelihood ratio test can be considered in place of the Wald test using an approximate power expression (e.g., [Self et al. 1992](#)). Test procedures relying less on asymptotic approximation could also be considered, but may be onerous to scale to larger data sets if they rely heavily on computation. In the illustration, all covariates have been treated as known ahead of the experiment, but it would be desirable to account for uncertainty in the counts of housing units. This work has focused solely on unit-level nonresponse; item-level nonresponse may also be of interest in sample size calculation.

The issue of cost was not a primary concern in our illustration by virtue of it being embedded within the census; here the main operational costs involve developing and administering the training module to enumerators. Such costs would be insignificant compared to those of the fieldwork and data processing needed to carry out the census. Cost may be a much larger concern in other settings, however—especially when the amount of fieldwork is dictated by the extent of the experiment—and may provide a much stronger motivation to control the sample size. Such settings could inspire interesting variations of the testing problem. For example, the experimenter may wish to maximize power subject to a given budget. Furthermore, costs of operations may vary among different locations or household types so that an allocation of the budget must be determined.

The illustration featured several notable simplifications which may need to be addressed in a real-life experiment. The illustration assumed a common maximum number of attempts L across all households. Subsection 5.1 mentioned plans to dynamically assign enumerators to households during the 2020 census NRFU operation until attempts are exhausted; however, L itself is also subject to dynamic adjustment ([U.S. Census Bureau 2019](#)). To account for uncertainty during planning, it may be conceivable to formulate a model for L and extend the sample size methodology accordingly. Experimenters may also wish to define “success” more broadly than in-person contact by an enumerator, and may include contact by another mode such as phone call, contact with a proxy, or an implicit response via administrative records in lieu of contact. For example, [Ashmead et al. \(2017\)](#) consider a more holistic contact process in the context of the American Community Survey.

Therefore, it may be necessary to generalize the outcome model beyond simple sequences of trials to provide a more comprehensive notion of response. Furthermore, variations of the contact strategy—including the choice of L for NRFU attempts—may be of interest to compare using an experiment.

The ability to support mixed effects would be a desirable extension to this work. For example, our illustration grouped the ACOs into pairs so that one element receives the experimental treatment and the other receives the control treatment. This was merely used to construct samples of increasing sizes to plot power curves, but such a design would be especially desirable if ACOs within a pair were known to exhibit more similar response behavior than ACOs across pairs. Here, a random intercept for each pair may be appropriate to reduce overall uncertainty in the fixed effects of interest. Other random effects such as enumerator and enumerator-attempt interaction could be considered as well; however, their use in sample size determination would be complicated in a setting with dynamic workload allocation.

The testing problem we have used as the basis for sample size determination is meant to assess if the treatment has any significant impact. An experimenter may instead wish to test if the treatment has a positive impact so that it has been worth the effort and investment. Under these circumstances, it could be more useful to formulate a one-sided alternative which implies the need to develop a test under a multivariate one-sided alternative hypothesis. In order to develop such a test, it may be possible to appeal to literature on order restricted inference (e.g., [Silvapulle and Sen 2004](#), chap.9). Required sample sizes may also be reduced by utilizing a one-sided alternative; for example, a result due to [Praestgaard \(2012\)](#) shows that the likelihood ratio test under constraints has a larger power compared to a test that does not take into account the constraints. One-sided alternatives—and other possible extensions described in this section—may be relevant in practice to planning experiments and worth consideration in future work.

7. Appendix

Proof of Result 3.3. Write $\eta_{i\ell} = \mathbf{x}_{i\ell}^\top \beta$. To derive (a), first note that

$$\frac{\partial}{\partial \beta} \log p_{i\ell} = \frac{1}{p_{i\ell}} g(\eta_{i\ell}) \mathbf{x}_{i\ell} = (1 + e^{-\eta_{i\ell}}) \frac{e^{-\eta_{i\ell}}}{(1 + e^{-\eta_{i\ell}})^2} \mathbf{x}_{i\ell} = [1 - G(\eta_{i\ell})] \mathbf{x}_{i\ell}$$

and

$$\frac{\partial}{\partial \beta} \log(1 - p_{ib}) = -\frac{1}{1 - p_{ib}} g(\eta_{ib}) \mathbf{x}_{ib} = -\frac{1 + e^{-\eta_{ib}}}{e^{-\eta_{ib}}} \frac{e^{-\eta_{ib}}}{(1 + e^{-\eta_{ib}})^2} \mathbf{x}_{ib} = -G(\eta_{ib}) \mathbf{x}_{ib}.$$

We then have

$$\begin{aligned} \frac{\partial}{\partial \beta} \log \mathcal{L}(\beta) &= \frac{\partial}{\partial \beta} \sum_{i=1}^n \sum_{\ell=1}^{L+1} I(w_i = \ell) \left[\log p_{i\ell} + \sum_{b=1}^{\ell-1} \log(1 - p_{ib}) \right] \\ &= \sum_{i=1}^n \sum_{\ell=1}^{L+1} I(w_i = \ell) \left[[1 - G(\eta_{i\ell})] \mathbf{x}_{i\ell} - \sum_{b=1}^{\ell-1} G(\eta_{ib}) \mathbf{x}_{ib} \right] \end{aligned}$$

$$= \sum_{i=1}^n \sum_{\ell=1}^{L+1} I(w_i = \ell) \mathbf{x}_{i\ell} - \sum_{i=1}^n \sum_{\ell=1}^{L+1} I(w_i \geq \ell) G(\boldsymbol{\eta}_{i\ell}) \mathbf{x}_{i\ell}.$$

For (b), let us first write

$$\mathbf{D}_w = \text{Diag}\{I(w_i \geq \ell) g(\mathbf{x}_{i\ell}^T \boldsymbol{\beta}) : (i, \ell) \in \mathcal{J}\},$$

$$\mathbf{D}_\beta = \text{Diag}\{P(w_i \geq \ell) g(\mathbf{x}_{i\ell}^T \boldsymbol{\beta}) : (i, \ell) \in \mathcal{J}\}$$

$$= \text{Diag}\left\{g(\mathbf{x}_{i\ell}^T \boldsymbol{\beta}) \prod_{b=1}^{\ell-1} [1 - G(\mathbf{x}_{i\ell}^T \boldsymbol{\beta})] : (i, \ell) \in \mathcal{J}\right\}. \quad (19)$$

so that $\mathbf{D}_\beta = E[\mathbf{D}_w]$. The last equality in Equation (19) can be justified by

$$p_{i\ell} = \frac{\pi_{i\ell}}{\pi_{i\ell} + \dots + \pi_{i,L+1}} = \frac{\pi_{i\ell}}{P(W_i \geq \ell)} = \frac{p_{i\ell} \prod_{b=1}^{\ell-1} (1 - p_{ib})}{P(W_i \geq \ell)} \Leftrightarrow (W_i \geq \ell)$$

$$= \prod_{b=1}^{\ell-1} (1 - p_{ib}).$$

Now the second derivative of the log-likelihood is

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log \mathcal{L}(\boldsymbol{\beta}) = - \sum_{i=1}^n \sum_{\ell=1}^L I(w_i \geq \ell) g(\mathbf{x}_{i\ell}^T \boldsymbol{\beta}) \mathbf{x}_{i\ell} \mathbf{x}_{i\ell}^T = -\mathbf{X}^T \mathbf{D}_w \mathbf{X}. \quad (20)$$

Taking the negative expectation of Equation (20) yields the desired information matrix.

8. References

- Agresti, A. 2013. *Categorical Data Analysis*. Wiley, 3rd edition.
- Albert, J.H., and S. Chib. 2001. "Sequential ordinal modeling with applications to survival data." *Biometrics* 57(3): 829–836,. DOI: <https://doi.org/10.1111/j.0006-341X.2001.00829.x>.
- Alho., J.M. 1990. "Adjusting for nonresponse bias using logistic regression." *Biometrika* 77(3): 617–624. DOI: <https://doi.org/10.1093/biomet/77.3.617>.
- Ashmead, R., E. Slud, and T. Hughes. 2017. "Adaptive intervention methodology for reduction of respondent contact burden in the American Community Survey." *Journal of Official Statistics* 33(4): 901–919. DOI: <https://doi.org/10.1515/jos-2017-0043>.
- Barboza, G.E., and S. Dominguez. 2016. "A sequential logit model of caretakers' decision to vaccinate children for the human papillomavirus virus in the general population." *Preventive Medicine* 85: 84–89. DOI: <https://doi.org/10.1016/j.ypmed.2016.01.010>.
- Blumenson, L.E. 1960. "A derivation of n -dimensional spherical coordinates." *The American Mathematical Monthly*, 67(1): 63–66. DOI: <https://doi.org/10.2307/2308932>.
- Boes, S., and R. Winkelmann. 2006. "Ordered response models." *Allgemeines Statistisches Archiv* 90: 167–181. DOI: <https://doi.org/10.1007/s10182-006-0228-y>.

- Brown, J.D., M.L. Heggeness, S.M. Dorinski, L. Warren, and M. Yi. 2018. *Understanding the quality of alternative citizenship data sources for the 2020 Census*. CES Working Paper Series: CES 18–38, Center for Economic Studies, U.S. Census Bureau. Available at: <https://www2.census.gov/ces/wp/2018/CES-WP-18-38.pdf> (accessed October 2022).
- Bush, S. 2015. “Sample size determination for logistic regression: A simulation study.” *Communications in Statistics – Simulation and Computation* 44(2): 360–373. DOI: [10.1080/03610918.2013.777458](https://doi.org/10.1080/03610918.2013.777458).
- Chow, S.-C., J. Shao, H. Wang, and Y. Lokhnygina. 2017. *Sample Size Calculations in Clinical Research*. Chapman and Hall/CRC, 3rd edition.
- Compton, E., and M. Bentley. 2012. *2010 Census nonresponse followup (NRFU) contact strategy experiment report*. 2010 Census Planning Memoranda Series: CPEX–174, Decennial Statistical Studies Division, U.S. Census Bureau. Available at: <https://www.census.gov/programs-surveys/decennial-census/decade/2010/program-management/cpex/2010-cpex-174.html>.
- Cordeiro, G.M., D.A. Botter, and S.L.D.P. Ferrari. 1994. “Nonnull asymptotic distributions of three classic criteria in generalised linear models.” *Biometrika* 81(4): 709–720. DOI: <https://doi.org/10.1093/biomet/81.4.709>.
- Cox, D.R. 1972. “Regression models and life-tables.” *Journal of the Royal Statistical Society. Series B* 34(2): 187–220. DOI: <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- Daas, P.J.H., M.J. Puts, B. Buelens, and P.A.M. van den Hurk. 2015. “Big Data as a Source for Official Statistics.” *Journal of Official Statistics* 31(2): 249–262. DOI: <https://doi.org/10.1515/jos-2015-0016>.
- Davern, M., M. Roemer, and W. Thomas. 2009. “Investing in a data quality research program for administrative data linked to survey data for policy research purposes is essential.” In Federal Committee on Statistical Methodology Research Conference, November 2–4, Washington, D.C. U.S.A., Available at: https://nces.ed.gov/FCSM/pdf/2009FCSM_Davern_IX-A.pdf (accessed October 2022).
- Demidenko, E. 2007. “Sample size determination for logistic regression revisited.” *Statistics in Medicine* 26(18): 3385–3397. DOI: <https://doi.org/10.1002/sim.2771>.
- Demidenko, E. 2008. “Sample size and optimal design for logistic regression with binary interaction.” *Statistics in Medicine* 27(1): 36–46. DOI: <https://doi.org/10.1002/sim.2980>.
- Deming, W.E. 1953. “On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response.” *Journal of the American Statistical Association* 48: 743–772. DOI: <https://doi.org/10.1080/01621459.1953.10501197>.
- Ellis, R., P. Goerman, K. Kephart, A.C. Fobia, A.S. Giron, M. Meyers, R. Terry, L. Fernandez, F. Lineback, M. Berger, A. Bruce, and E. Jensen. 2018. *Research on coverage of underrepresented populations in anticipation of a records-based census*. 2020 Census: Evaluation, Experiment, and Research and Testing Study. Internal report.
- Erdman, C., and N. Bates. 2016. “The low response score (LRS): A metric to locate, predict, and manage hard-to-survey populations.” *Public Opinion Quarterly* 81(1): 144–156. DOI: <https://doi.org/10.1093/poq/nfw040>.

- Erdman, C., and N. Bates. 2017. Available at: <https://www.census.gov/topics/research-guidance/planning-databases.html> (accessed October 2022).
- Fienberg, S.E. 2007. *The analysis of cross-classified categorical data*. Springer Science & Business Media, 2nd edition.
- Fullerton, A.S. 2009. "A conceptual framework for ordered logistic regression models." *Sociological Methods & Research* 38(2): 306–347. DOI: <https://doi.org/10.1177/0049124109346162>.
- Ghosal, S., and A. van der Vaart. 2017. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.
- Groves, R.M., and G.J. Schoeffel. "Use of administrative records in evidence-based policymaking." *The ANNALS of the American Academy of Political and Social Science* 678(1): 71–80. DOI: <https://doi.org/10.1177/0002716218766508>.
- Hansen, M.H., and W.N. Hurwitz. 1946. "The problem of non-response in sample surveys." *Journal of the American Statistical Association* 41(236): 517–529. DOI: <https://doi.org/10.1080/01621459.1946.10501894>.
- Lohr, S.L. 2010. *Sampling: Design and Analysis*. Brooks/Cole, Boston, MA, 2nd edition, 2010.
- Lyles, R.H., H.-M. Lin, and J.M. Williamson. 2007. "A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses." *Statistics in Medicine* 26(7): 1632–1648. DOI: <https://doi.org/10.1002/sim.2617>.
- Molfino, E., G. Korkmaz, S.A. Keller, A. Schroeder, S. Shipp, and D.H. Weinberg. 2017. "Can administrative housing data replace survey data?" *Cityscape* 19(1): 265–292.
- Morris, D.S., A. Keller, and B. Clark. 2016. "An approach for using administrative records to reduce contacts in the 2020 Decennial Census." *Statistical Journal of the IAOS* 32(2): 177–188. DOI: <https://doi.org/10.3233/SJI-161002>.
- Myers, R.H. 2000. *Classical and Modern Regression with Applications*. Duxbury Press, 2nd edition.
- National Research Council. 2010. In *Envisioning the 2020 Census*, edited by L.D. Brown, M.L. Cohen, D.L. Cork, and C.F. Citro., The National Academies Press, Washington, D.C. DOI:<https://doi.org/10.17226/12865>.
- Oehlert, G.W. 2000. *A first course in design and analysis of experiments*. W.H. Freeman. Available at: <http://users.stat.umn.edu/~gary/Book.html>.
- Pan, Y., and S. Lubkemann. 2013. *Observing census enumeration of non-English speaking households in the 2010 Census: Evaluation report*. Research Report Series: Survey Methodology 2013(2). Center for Survey Measurement, U.S. Census Bureau. Available at: <https://www.census.gov/library/working-papers/2013/adrm/ssm2013-02.html> (accessed October 2023).
- Peyhardi, J., C. Trottier, and Y. Guédon. 2015. "A new specification of generalized linear models for categorical responses." *Biometrika* 102(4): 889–906. DOI: <https://doi.org/10.1093/biomet/asv042>.
- Politz, A., and W. Simmons. 1949. "An attempt to get the "not at homes" into the sample without callbacks." *Journal of the American Statistical Association* 44(245): 9–16. DOI: <https://doi.org/10.2307/2280347>.

- Praestgaard, J. 2012. "A note on the power superiority of the restricted likelihood ratio test." *Journal of Multivariate Analysis* 104(1): 1–15. DOI: <https://doi.org/10.1016/j.jmva.2011.04.004>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/> (accessed October 2023).
- Rao, p. 1983 "Callbacks, follow-ups, and repeated telephone calls." In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, and D.B. Rubin 33–44. Academic Press, New York.
- Rigon, T., and D. Durante. 2021. "Tractable Bayesian density regression via logit stick-breaking priors." *Journal of Statistical Planning and Inference* 211: 131–142. DOI: <https://doi.org/10.1016/j.jspi.2020.05.009>.
- Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag New York, Inc., New York.
- SAS Institute Inc. 2018. *The GENMOD Procedure* 3407–3607. SAS Publishing. Available at: <http://support.sas.com/documentation/onlinedoc/stat/151/genmod.pdf> (accessed October 2023).
- Scheuren, F. 1999. "Administrative records and census taking." *Survey Methodology* 25(2): 151–160. Available at: <https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4878-eng.pdf> (accessed October 2022).
- Self, S.G., and R.H. Mauritsen. 1988. "Power/sample size calculations for generalized linear models." *Biometrics* 44(1): 79–86. DOI: <https://doi.org/10.2307/2531897>.
- Self, S.G., R.H. Mauritsen, and J. Ohara. 1992. "Power calculations for likelihood ratio tests in generalized linear models." *Biometrics* 48(1): 31–39. DOI: <https://doi.org/10.2307/2532736>.
- Shieh, G. 2000. "On power and sample size calculations for likelihood ratio tests in generalized linear models." *Biometrics* 56(4): 1192–1196. DOI: <https://doi.org/10.1111/j.0006-341X.2000.01192.x>.
- Shieh, G. 2005. "On power and sample size calculations for Wald tests in generalized linear models." *Journal of Statistical Planning and Inference* 128(1): 43–59. DOI: <https://doi.org/10.1016/j.jspi.2003.09.017>.
- Silvapulle, M.J., and P.K. Sen. 2004. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley-Interscience.
- Singer, E. 2006. "Introduction: Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5): 637–645. DOI: <https://doi.org/10.1093/poq/nfl034>.
- Slud, E., and B. Kedem. 1994. "Partial likelihood analysis of logistic regression and autoregression." *Statistica Sinica* 4(1): 89–106.
- Tutz, G. 1991. "Sequential models in categorical regression." *Computational Statistics & Data Analysis* 11(3): 275–295. DOI: [https://doi.org/10.1016/0167-9473\(91\)90086-H](https://doi.org/10.1016/0167-9473(91)90086-H).
- Tutz, G. 2022. "Ordinal regression: A review and a taxonomy of models." *WIREs Computational Statistics* 14(2): e1545. DOI: <https://doi.org/10.1002/wics.1545>.
- U.S. Census Bureau. 2013. *U.S. Census Bureau statistical quality standards*. Available at: https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf (accessed October 2022).

- U.S. Census Bureau. 2019. *2020 Census detailed operational plan for: 18. nonresponse followup operation (NRFU)*. Available at: <https://www.census.gov/programs-surveys/-decennial-census/2020-census/planning-management/planning-docs/NRFU-detailed-op-plan.html>.
- Walker, S., S. Winder, G. Jackson, and S. HeimeI. 2012. *2010 census nonresponse followup operations assessment*. Technical Report 190, 2010 Census Planning Memoranda Series. Available at: https://www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf (accessed October 2022).
- Wood, A.M., I.R. White, and M. Hotopf. 2006. "Using number of failed contact attempts to adjust for non-ignorable non-response." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(3): 525–542. DOI: <https://doi.org/10.1111/j.1467-985X.2006.00405.x>.

Received June 2021

Revised November 2021

Accepted October 2022

Estimating Intra-Regional Inequality with an Application to German Spatial Planning Regions

Marina Runge¹

Income inequality is a persistent topic of public and political debate. In this context, the focus often shifts from the national level to a more detailed geographical level. In particular, inequality between or within local communities can be assessed. In this article, the estimation of inequality within regions, that is, between households, is considered at a regionally disaggregated level. From a methodological point of view, a small area estimation of the Gini coefficient is carried out using an area-level model linking survey data with related administrative data. Specifically, the Fay-Herriot model is applied using a logit-transformation followed by a bias-corrected back-transformation. The uncertainty of the point estimate is assessed using a parametric bootstrap procedure to estimate the mean squared error. The validity of the methodology is shown in a model-based simulation for the point estimator as well as for the uncertainty measure. The proposed methodology is illustrated by estimating model-based Gini coefficients for spatial planning regions in Germany, using survey data from the Socio-Economic Panel and aggregate data from the 2011 Census. The results show that intra-regional inequality is more diverse than a consideration only between East and West suggests.

Key words: Fay-Herriot model; Gini coefficient; small area estimation; survey statistics.

1. Motivation

For some time now, and especially since the United Nations Sustainable Development Goals (SDGs) of 2015 ([UN General Assembly 2015](#)), the reduction of inequality within and among countries has increasingly become a focus of public debate. Regionally differentiated indicators to measure poverty and inequality are thereby receiving growing attention in the attempt to quantify inequality. In order to meet the demands and expand policies to reduce economic inequality, it is of great importance to provide reliable statistics that adequately capture regional differences in income inequality. In Germany, due to its division in 1949 and reunification in 1990, economic inequality, especially between East and West, has been a particular focus of political and public debate. At the latest since the financial crisis of 2008/2009 regional income and wealth disparities that go far beyond East and West have reached public awareness, and this is likely to be reinforced with the 2020/2021 pandemic. Therefore, [Goebel and Frick \(2005\)](#) already considered regional income stratification by dividing Germany into four parts. [Braml and Felbermayr](#)

¹Institute of Statistics and Econometrics, Freie Universität Berlin, Garystraße 21, 14195 Berlin, Germany. Email: marina.runge@fu-berlin.de

Acknowledgments: The author appreciates gratefully the support of the German Research Foundation within the TESAP project (281573942). The data used in this publication were made available by the German SocioEconomic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin. The author is grateful for the computation time provided by the HPC service (10.17169/refubium-26754) of the Freie Universität Berlin.

(2018) focus on inequality at the county level measured by gross domestic product per capita, just as [Kreutzmann et al. \(2022\)](#) consider regional heterogeneity in wealth. In both, the focus is on the difference between regions, while an additional aspect of inequality is income differences between households within a region. [Immel and Peichl \(2020\)](#) combine both perspectives and look at regional income inequality at the county level measured by the top 10% earners and the bottom 40% within regions. When examining the regional dimension of income distributions, a distinction must be made between intra- and inter-regional inequality, as noted before. When considering intra-regional inequality, an appropriate measure must be used to determine the level of income inequality. A popular indicator for this purpose is the Gini coefficient ([Gini 1912](#)), which is defined between zero and one, where zero means perfect equality and one maximum inequality. The presented methodology is illustrated by estimating Gini coefficients at a regionally disaggregated level for Germany, which additionally represents to best of knowledge the first attempt to estimate Gini coefficients for Germany at a regional level lower than the federal states.

When it comes to measuring regional differences, the level of observation can become very detailed and the unit sample sizes very small. A unit in this context can be a regional area, a sociodemographically defined domain or a combination of both. In either case it is referred to as a domain or an area and when sample sizes are small, as small area. For small sample sizes, common estimators that use only survey data (hereafter referred to as direct estimators) are often not accurate enough to provide reliable domain-specific estimates of an indicator of interest. In these cases, small area estimation (SAE) methods allow for an increase in accuracy. In particular, model-based SAE methods use related additional data sources and information from other areas for this purpose. Overviews of SAE methods can be found in [Pfeffermann \(2013\)](#), [Rao and Molina \(2015\)](#) and [Jiang and Rao \(2020\)](#). A general framework for the construction of small area statistics is presented by [Tzavidis et al. \(2018\)](#). In [Pratesi \(2016\)](#) SAE methods particularly for the analysis of poverty data are provided. The most common SAE methods to estimate poverty and inequality indicators, such as Gini coefficients, on a disaggregated level are the World Bank method proposed by [Elbers et al. \(2003\)](#) or the empirical best predictor (EBP) method proposed by [Molina and Rao \(2010\)](#). In practice, however, this is problematic for privacy reasons. Especially when it comes to population data on a micro/ individual level that are needed as auxiliary information. In these cases, area-level methods can help, where survey and related population data are only needed at the aggregated level. In addition, area-level models account for complex survey designs in the estimation of point and variance estimators. One of the most popular area-level SAE models is that proposed by [Fay and Herriot \(1979\)](#), known as the Fay-Herriot (FH) model, which is the underlying statistical model in this article. In addition, there are empirical and hierarchical Bayesian methods, see for a comprehensive overview for example, [Rao and Molina \(2015\)](#). In particular, the FH model can be estimated by an hierarchical Bayes model as well. [Liu et al. \(2014\)](#) use the hierarchical Bayes version of the FH model to compare it to a normal-logistic and a beta-logistic Bayes model for the use-case of estimating small area proportions. Also [Janicki \(2020\)](#) studies a hierarchical Bayesian model with a Beta distribution and a logit link to estimate poverty rates. The common property of proportions and Gini coefficients is that both are bounded in the interval $(0, 1)$. Therefore, some of the method can be used for both applications. [Fabrizi and Trivisano \(2016\)](#) propose a hierarchical Beta mixed

Bayesian regression area-level model with a logit link to estimate Gini coefficients for small areas and [Fabrizi et al. \(2016\)](#) apply this approach to jointly estimate at-risk-of-poverty rates and the Gini coefficients. The advantages of this and more general Bayesian approaches are that from the resulting posterior distribution, which is approximated by a Markov Chain Monte Carlo (MCMC) algorithm, the point estimates are directly given with an uncertainty measure as well as credible intervals. The possibility to specify different prior distributions of the model parameters also makes the model quite flexible. However, frequentist approaches probably predominate in the SAE literature and are widely accepted in National Statistical Institutes (NSI). From a frequentist perspective to the best of knowledge there is no SAE literature on the estimation of Gini coefficients at the regional level using area-level data, and specifically with application of the FH model. The possible advantages of using a frequentist approach are, that it is probably easier to follow for common users who are more used to frequentist regression models and the available software for SAE methods implements mostly frequentist methods. In addition, there are a number of elaborated results from a frequentist perspective for the FH model that can be adapted. As the FH model allows for the use of a transformation, it is a common approach to satisfy the normality assumptions of the error terms or to ensure that the estimated values are within a predefined range. [Slud and Maiti \(2006\)](#), for example, propose a log-transformed FH model for skewed data, and in the case of proportions, for example, [Casas-Cordero et al. \(2016\)](#) use an arcsine-transformed FH model to estimate poverty rates and [Schmid et al. \(2017\)](#) for literacy rates. To estimate Gini coefficients using the FH model, in this work the approach of [Fabrizi and Trivisano \(2016\)](#) is followed and a logit-transformation is used to link the response values to the related covariate information. This is also motivated by the condition that the estimated Gini coefficients must lie between zero and one, in addition to stabilizing the variance of the direct estimator and to promoting the normal distribution of the sampling errors and random effects of the model. The choice of a logit-normal rather than a beta likelihood as in [Fabrizi and Trivisano \(2016\)](#), is also driven by the possibility to use already existing results, such as those of [Sugasawa and Kubokawa \(2017\)](#) for the back-transformation. When using transformations the resulting point estimate is on the transformed scale and has to be back-transformed. An application of the inverse usually introduces a bias for nonlinear transformations, therefore [Sugasawa and Kubokawa \(2017\)](#) propose a bias-corrected back-transformation for general parametric transformations. This bias-corrected back-transformation is adopted to the logit-transformation in this article. Instead of the logit transformation, any other transformation could in principle also be used, as long as the inverse maps into a range between 0 and 1. For example, a complementary log-log or probit transformation could also be used if suitable transformations are available for the variance of the direct estimator and the back-transformation of the point estimator. In this article, however, the focus is on the logit transformation, since it is one of the most common. To evaluate the accuracy of model-based SAE estimators, uncertainty measures must be estimated. As a common practice, the MSE is considered for this purpose. If analytical solutions for its estimation cannot be derived, bootstrap methods are often implemented instead. Here, the uncertainty of the estimated Gini coefficients is assessed using a bootstrap procedure following [Gonzalez-Manteiga et al. \(2005\)](#) with an additional step of applying the bias-corrected back-transformation similar to [Hadam et al. \(2020\)](#).

The validity of the presented point estimator using a logit-transformed FH model with a bias-corrected back-transformation, as well as that of the uncertainty measure, is demonstrated in a simulation study.

The article is organized as follows. Section 2 describes the data used to illustrate the proposed methodology, in particular survey data from the Socio-Economic Panel (SOEP) and auxiliary data from administrative sources, such as the Census 2011 in Germany. The statistical methodology is introduced in Section 3. The validity of the proposed methodology is assessed in a simulation study in Section 4. Section 5 presents the application of the model-based small area method to estimate Gini coefficients for German regions. Section 6 completes the article with some concluding remarks and discusses further potential research.

2. Sources of Data and Initial Analysis

In this section, the data sources used for the analysis in Section 5 are described. Specifically, data from the German SOEP ([Socio-Economic Panel 2019](#)) are used to form the target indicator, and data from the 2011 Census ([Statistische Ämter des Bundes und der Länder 2011a](#)) and the regional data base from the National Statistical Office ([Statistische Ämter des Bundes und der Länder 2011b](#)) are taken as auxiliary information. To have both data sources from the same year, the SOEP data collected in 2011 are used. Furthermore, a preliminary calculation of the Gini coefficients at a regional level is presented.

2.1. German Socio-Economic Panel

The German SOEP is a longitudinal study that has been running since 1984 and is conducted annually. It currently covers about 15,000 private households in Germany and aims to represent German society. Information is collected on various areas of life, such as demography, employment, taxes, income, education, health and satisfaction. The SOEP-team at the German Institute for Economic Research (DIW Berlin) prepares and provides the survey data. The main data set SOEP Core currently consists of 12 sub samples. The initial sample, sample A, was first surveyed in 1984 and represents the West German population of the Federal Republic of Germany ([Kara et al. 2019](#)). In 1990, the initial sample East after the reunification was included, representative of the East German population of the German Democratic Republic. Over the years (1998, 2000, 2002, 2011), four refreshment samples were added, further enlarging the total sample. In addition to the refreshment samples, other special samples to increase statistical power were included, such as the migration samples in 1984 and 1994/95, which oversamples foreigners or the high income sample in 2002 to represent households at the top of the income distribution. Sampled households are surveyed every year. The SOEP questionnaires are constructed in such a way that individuals in a SOEP household can be studied from birth to adulthood and over the rest of their lives. The SOEP aims to measure stability and identify changes across time, so the survey methodology remains almost identical over time ([Kara et al. 2019](#)). In the analysis in Section 5 data from the available refreshment sample in survey year 2011 is used. The sample aimed to cover a cross-section of private German households and is based on a clustered sampling strategy. Households were drawn at random from 307 primary sampling units (PSU) stratified by federal states, administrative regions and a classification of municipalities by number of inhabitants ([Siegers et al. 2020](#)). A random walk procedure was

applied to select the addresses within each PSU. The provided household weights account for sampling design, non-response, and panel attrition and are further post-stratified to known population distributions based on the German microcensus.

The Gini coefficients calculated in this article are computed with household-level data. The variable to form the target indicator in this section and for the application in Section 5 is the equivalised disposable household income, which is calculated using total net household income divided by equivalised household size. The equivalised household size is derived using the Organisation for Economic Co-operation and Development (OECD) modified scale first proposed by [Hagenaars et al. \(1994\)](#). The distribution of the variable in the sample is reported in [Table 1](#). The Gini coefficient for the equivalised disposable household income reported in 2011 for Germany by [OECD \(2011\)](#) is 0.29. [Goebel and Frick \(2005\)](#) investigate regional income inequality by estimating Gini coefficients for East and West Germany and for a further regional stratification by dividing Germany into northern, eastern, western and southern states. This analysis indicates that there is regional heterogeneity in income inequality. In addition, the OECD reports Gini coefficients for the German federal states ([OECD 2013](#)), which reveals further regional differences in inequality ranging from 0.23 in Saxony to 0.32 in Hesse. Another spatial dis-aggregation that enables the examination of inequality in rural and urban regions is the consideration of 96 spatial planning regions (SPRs) of the Federal Office for Building and Regional Planning. SPRs are composed of several administrative districts and form an intermediate regional level between these districts and the federal states. A map showing the assignment of the SPRs and associated labels can be found in the appendix in [Figure 7](#) and [Table 7](#). The information to which SPR the residence of a SOEP household is assigned to can be found in the SOEP geocodes ([Goebel 2017](#)). The investigation of regional differences in income inequality in Germany is therefore done for the 96 SPRs. [Figure 1](#) shows estimated Gini coefficients from left to right for East and West Germany, a fourfold division of Germany into East, North, South and Central, the federal states and the SPRs.

Table 1. Distribution of equivalised disposable household income (EUR), sample sizes for SPRs and number of SPRs without observations.

	Min	1stQ	Median	Mean	3rdQ	Max	No obs.
Equal. disp. income	0	12363	17805	20579	25270	322508	
SPR sample size	4	17	27	35	47	153	7

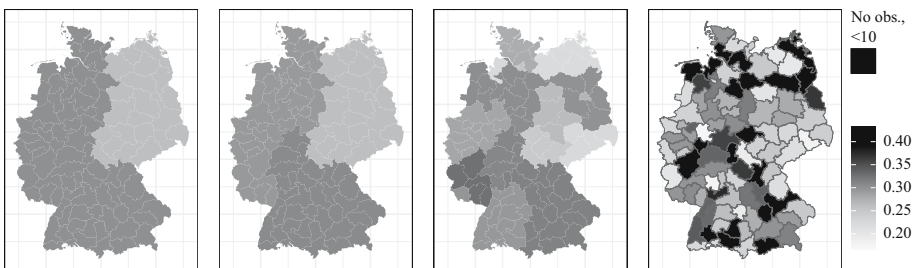


Fig. 1. Gini coefficients for equivalised disposable income for East and West Germany (left), a fourfold division of Germany into East, North, South and Central, the federal states and SPRs (right). SPRs with no or less than ten observations are colored in black.

The first two maps already show that there are regional differences, as illustrated by [Goebel and Frick \(2005\)](#). The map of the federal states underlines this heterogeneity. Looking at the fourth map, the regional differences in income inequality become even more obvious. At the same time looking at [Table 1](#), for some SPRs, these estimates are based on a very small sample size, so that the reliability of the estimates cannot be guaranteed. To improve the accuracy of estimated Gini coefficients for SPRs with small sample sizes, model-based SAE methods combine direct estimates with auxiliary information from registers by statistical models. Furthermore, those methods allow to provide estimates for regions that have no observations in the survey, usually referred to as out-of-sample (OOS) regions. This is the case for seven SPRs. According to the privacy agreement with the data provider, direct estimates of SPRs with less than ten observations cannot be reported. This applies to eleven SPRs. In the map for the SPRs ([Figure 1](#)), these and the OOS SPRs are colored in black.

2.2. Auxiliary Information

To improve the accuracy of the target indicator the model described in Subsection 3.1 makes use of auxiliary information from administrative data sources as registers or census data on an aggregated level. For the application in this work German Census data from 2011 ([Statistische Ämter des Bundes und der Länder 2011a](#)) is used, which is publicly available on an administrative district level. Furthermore data on taxes, gross domestic product (GDP), mortality and birth numbers available from the National Statistical Offices are used ([Statistische Ämter des Bundes und der Länder 2011b](#)). A detailed explanation of the calculation of the GDP on district level can be found in [Statistische Ämter der Länder \(2021\)](#). To obtain the data at the same level as the survey data, they are aggregated to SPR level. The assignment of counties and districts to SPRs is provided by the Federal Office for Building and Regional Planning ([Bundesinstitut für Bau-, Stadt-, und Raumforschung 2021](#)). The objective is to find variables in the data that are related to income inequality and could serve as possible predictors. [Furceri and Ostry \(2019\)](#) examine robust drivers of income inequality and identify, among other factors, the level of development and demographics as key determinants, as well as the extent of unemployment. [Perugini and Martino \(2008\)](#) examine the factors that drive inequality within European regions. Both divide the factors into groups of demographic, institutional and economic condition variables, among others. The possible covariates that were able to be extracted and aggregated from the data sources available are presented in [Table 2](#) with summary statistics. Although [Furceri and Ostry \(2019\)](#) consider inequality determinants between countries, this could be transferred to within country inequality and development. When considering economic conditions, in addition to GDP, which is a measure of a region's development, the shares of the agricultural, industrial and social service sectors in the labor market are also an indicator of economic development. Since the industrial sector is generally expected to generate higher income, this could lead to a better distribution of income than a high share in the agricultural sector. In line with [Fabrizi and Trivisano \(2016\)](#) and [Perugini and Martino \(2008\)](#) taxable income and the share of income taxpayers can be an indirect measure of labor performance and, moreover, an indicator of the resources that local governments could use to fund education, child care, health, and so on, to foster future growth and thus reduce inequality.

Table 2. Distributions of possible auxiliary information.

	Min	1stQ	Median	Mean	3rdQ	Max
<i>Economic/Institutional conditions</i>						
GDP per resident [EUR]	22159	72625	121001	132171	174988	381263
log(GDP per resident)	10.010	11.190	11.700	11.630	12.070	12.850
Avg. taxable income per person in Tsd. [EUR]	2.972	4.192	4.933	4.886	5.553	8.840
Share income tax payer	0.399	0.457	0.481	0.480	0.499	0.683
Share agricultural employment sector	0.000	0.002	0.008	0.010	0.015	0.039
Share industrial employment sector	0.100	0.181	0.216	0.219	0.253	0.359
Share service sector	0.452	0.512	0.542	0.546	0.572	0.664
Unemployment ratio	0.002	0.021	0.035	0.038	0.053	0.087
High education ratio	0.153	0.247	0.293	0.297	0.340	0.488
<i>Demographics</i>						
Population density	44.0	117.5	178.0	330.7	274.5	3927.0
log(Population density)	3.784	4.766	5.182	5.347	5.615	8.276
Foreign residents ratio	0.009	0.033	0.054	0.060	0.085	0.153
Child dependency ratio	0.162	0.194	0.206	0.204	0.220	0.241
Elderly dependency ratio	0.263	0.291	0.316	0.318	0.338	0.415
Births rate	6.734	7.455	7.814	7.922	8.280	11.837
Mortality rate	8.000	9.701	10.691	10.778	11.677	14.358

The level of unemployment naturally measures the economic situation of a region, just as the level of education is a proxy for development. An approach similar to [Fabrizi and Trivisano \(2016\)](#) is used to calculate a high education ratio. Therefore the number of people aged between 18 and 64 with at least high school diploma are divided by the number of all people aged between 18 and 64. Following [Furceri and Ostry \(2019\)](#) demographic data such as dependency ratios, birth, and death rates are also among the possible covariates, as they indirectly approximate economic development. This is also true for the foreigner rate, as immigration could lead to an increasing wage gap ([Furceri and Ostry 2019](#)).

3. Small Area Estimation Method

In this section, the statistical methodology is presented. The underlying model for estimating small area means was proposed by [Fay and Herriot \(1979\)](#), which combines aggregate population auxiliary variables with direct estimators based on survey data. In this work, the target indicators are area-specific Gini coefficients. Since it is a nonlinear indicator within a specified range, a logit-transformation is applied to promote the normality assumption of the model and to ensure that the estimates are between zero and one. To measure the uncertainty of the point estimator, a parametric bootstrap procedure is presented.

3.1. Logit-Transformed Fay-Herriot Model

Let N be the size of a finite population divided into $d = 1, \dots, D$ domains and n the sample size with $i = 1, \dots, n_d$ units per domain so that $n = \sum_{d=1}^D n_d$. The FH model is a two-level

model that includes a sampling model at the first level, assuming that the direct estimator consists of the true domain-specific population indicator θ_d and sampling errors e_d .

$$\hat{\theta}_d^{Dir} = \theta_d + e_d, \quad e_d \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{e_d}^2). \quad (1)$$

The sampling errors e_d are assumed to be independently normally distributed with known variance $\sigma_{e_d}^2$. However, although the sample variance $\sigma_{e_d}^2$ is taken as known, in many applications it has to be estimated itself, what can be done on the basis of unit-level sample data (Rivest and Vandal 2002; Wang and Fuller 2003; You and Chapman 2006) or by bootstrap algorithms proposed in Alfons and Templ (2013). There are several proposed direct estimators for the Gini coefficient in the literature. A common estimator is the one proposed by Alfons and Templ (2013). Fabrizi and Trivisano (2016) show in a simulation experiment, that this estimator can have a negative bias when sample sizes are small and propose a corrected version with a bias reduction. The direct estimator proposed by Fabrizi and Trivisano (2016) is defined as

$$\hat{\theta}_d^{Dir} = \frac{1}{2\hat{Y}_d} \frac{\sum_{i=1}^{n_d} \sum_{j=1}^{n_d} w_{di} w_{dj} |y_{di} - y_{dj}|}{\hat{N}_d^2 = \sum_{i=1}^{n_d} w_{di}^2}, \quad (2)$$

with $\hat{N}_d = \sum_{i=1}^{n_d} w_{di}$ and $\hat{Y}_d = \hat{N}_d^{-1} \sum_{i=1}^{n_d} w_{di} y_{di}$, where y_{di} is the income or wealth variable, in this article the equivalised disposable household income and w_{di} denote the sampling weights. By including the sample weights in the associated variance estimate, the direct estimator incorporates the complex design information. The variances $\sigma_{e_d}^2$ of $\hat{\theta}_d^{Dir}$ for $d = 1, \dots, D$ can be estimated via a naive or calibrated bootstrap procedure described in Alfons and Templ (2013). Since the direct variance estimates are based on small sample sizes a variance smoothing model analogous to that in Fabrizi and Trivisano (2016) is used for stabilization. The model assumes a beta distribution for the Gini coefficient and uses the relationship between the expected value and the variance of the beta distribution. It is defined as follows:

$$\frac{\hat{\theta}_d^{Dir^2} (1 - \hat{\theta}_d^{Dir^2})}{2\sigma_{e_d}^2} = \lambda n_d + \epsilon_d \quad (3)$$

where the error term is assumed to be normally distributed $\epsilon_d \sim \mathcal{N}(0, \tau^2)$ and λ is estimated using least squares.

The second level of the FH model is a linking model that links covariate information to the population indicator. x_d is a $p \times 1$ vector of domain-specific population covariates and β is the corresponding $p \times 1$ vector of regression coefficients. v_d are domain-specific random effects, which are normally distributed:

$$\theta_d = x_d^T \beta + v_d, \quad v_d \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_v^2). \quad (4)$$

To ensure that the estimated Gini coefficients lie within $(0, 1)$, to further stabilize the variance and following Fabrizi and Trivisano (2016), the logit function is applied to the direct estimator from Equation (2):

$$\hat{\theta}_d^{Dir*} = \text{logit}(\hat{\theta}_d^{Dir}) = \log\left(\frac{\hat{\theta}_d^{Dir}}{(1 - \hat{\theta}_d^{Dir})}\right).$$

In the following, * always refers to the logit-scale. To obtain the variances of the direct estimator on the transformed scale, one can transfer the smoothed bootstrap variances to the logit scale using Taylor expansion for moments, which leads to:

$$\sigma_{e_d}^{2*} = \frac{\sigma_{e_d}^2}{\left[\hat{\theta}_d^{Dir} (1 - \hat{\theta}_d^{Dir})\right]^2}. \tag{5}$$

Using a Taylor expansion for moments to transform variances from the original scale to the transformed scale is a common procedure in SAE as in [Neves et al. \(2013\)](#) and [Citro and Kalton \(2000\)](#).

The combination of the sampling model in Equation (1) and the linking model in Equation (4) with the logit-transformed direct estimator results in:

$$\text{logit}(\hat{\theta}_d^{Dir}) = x_d^T \beta + v_d + e_d^*, \quad v_d \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_v^2), \quad e_d^* \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{e_d}^{2*}). \tag{6}$$

The unknown parameters of the model (4) to be estimated are the model variance σ_v^2 and the regression coefficients β . Methods to estimate σ_v^2 are for example Restricted Maximum Likelihood (REML), Maximum Likelihood (ML) and the FH method-of-moments. Details on model variance estimation methods can be found, for example, in [Rao and Molina \(2015\)](#). In this article, the REML method is used, which has the advantage over the ML method of taking into account the loss of degrees of freedom in the estimation of the regression coefficients β ([Rao and Molina 2015](#)). Let $\hat{\sigma}_v^2$ be an unbiased estimator for σ_v^2 . Then the best linear unbiased estimator (BLUE) under model (6) for the regression coefficients β is given by:

$$\hat{\beta} = \hat{\beta}(\hat{\sigma}_v^2) = \left(\sum_{d=1}^D \frac{x_d x_d^T}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\right)^{-1} \left(\sum_{d=1}^D \frac{x_d \hat{\theta}_d^{Dir*}}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}\right).$$

Since the model inputs are on the logit scale, the estimated regression coefficients $\hat{\beta}$ as well. Therefore, only the direction of the effect on the estimated model-based Gini coefficient on the original scale can be interpreted.

The FH estimator on the logit scale is obtained by:

$$\hat{\theta}_d^{FH*} = x_d^T \hat{\beta} + \hat{v}_d = \hat{\gamma}_d \hat{\theta}_d^{Dir*} + (1 - \hat{\gamma}_d) x_d^T \hat{\beta} \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\sigma_{e_d}^{2*} + \hat{\sigma}_v^2}. \tag{7}$$

$\hat{\gamma}_d$ is the shrinkage factor which determines an optimal balance between the direct estimator and the synthetic component. If the variance of the direct estimator is large, more weight is given to the synthetic component. The estimated model variance, that is, the variance of the random effects σ_v^2 , is also on the logit scale, as are the sampling variances. Therefore, the weighting factor can also be interpreted as the proportion of the variation explained by the hierarchical structure of the data. For highly skewed data, the transformation helps to better fit the linear relationship in the model, so using a

transformation on skewed data can often give more weight to the synthetic part. Since the direct estimators and their variances of the Gini coefficients were transformed to the logit scale as model input for the FH model, the resulting FH estimator $\hat{\theta}_d^{FH*}$ of the Gini coefficients is also still on the logit scale. To obtain the estimates on the original scale, a back transformation is required. As naive inverse back-transformations (in this case the logistic function) usually introduce a bias for nonlinear functions, [Sugasawa and Kubokawa \(2017\)](#) present an asymptotically unbiased back-transformation for a general parametric transformation. [Hadam et al. \(2020\)](#) applies this to the arcsine transformation, for example. Following [Sugasawa and Kubokawa \(2017\)](#) to obtain a bias-corrected back-transformation for $\hat{\theta}_d^{FH}$, the normal distribution of the transformed FH estimator on the logit-scale and the expected value (E) of a transformation (here the inverse logit) are used. The bias-corrected back-transformation applied to obtain the final FH estimates of the Gini coefficients $\hat{\theta}_d^{FH}$ at the original scale is as follows:

$$\begin{aligned} \hat{\theta}_d^{FH} &= E \left[\text{logit}^{-1} \left(\hat{\theta}_d^{FH*} \right) \right] = E \left[\frac{\exp \left(\hat{\theta}_d^{FH*} \right)}{1 + \exp \left(\hat{\theta}_d^{FH*} \right)} \right] = \int_{-\infty}^{\infty} \frac{\exp(t)}{1 + \exp(t)} f_{\hat{\theta}_d^{FH*}}(t) dt \\ &= \int_{-\infty}^{\infty} \frac{\exp(t)}{1 + \exp(t)} \frac{1}{\sqrt{2\pi \frac{\hat{\sigma}_v^2 \hat{\sigma}_{e_d}^{2*}}{\hat{\sigma}_v^2 + \hat{\sigma}_{e_d}^{2*}}}} \exp \left(- \frac{\left(t - \hat{\theta}_d^{FH*} \right)^2}{2 \frac{\hat{\sigma}_v^2 \hat{\sigma}_{e_d}^{2*}}{\hat{\sigma}_v^2 + \hat{\sigma}_{e_d}^{2*}}} \right) dt. \end{aligned} \quad (8)$$

In Equation (8) the integral has to be solved by numerical integration methods. The advantage of the bias-corrected back-transformation over the naive inverse is illustrated in the simulation experiment in Section 4.

3.2. Uncertainty Measure

In order to evaluate the accuracy of the FH estimator with a logit-transformation and to demonstrate the benefit of model-based estimators over direct ones, it is necessary to determine the degree of uncertainty. In the case of the FH estimator without a transformation, analytical solutions exist to estimate the MSE, such as the MSE estimator according to [Prasad and Rao \(1990\)](#). In the log-transformed FH model, [Slud and Maiti \(2006\)](#) also derived an analytical MSE estimator. There, the relationships between the log-normal distribution and the normal distribution and their expected values are used. This approach cannot be straightforwardly applied to the logit-transformation and the relationship between the logitnormal and the normal distribution, as there are no analytical solutions for the moments of the former. A common approach to estimating the MSE if no analytical estimator can be derived is to use a bootstrap algorithm. In line with [Gonzalez-Manteiga et al. \(2005\)](#), the MSE of $\hat{\theta}_d^{FH}$ is approximated with the following parametric bootstrap procedure:

1. Estimate the regression synthetic components $\hat{\beta}$ and $\hat{\sigma}_v^2$ using the direct components $\hat{\theta}_d^{Dir*}$ and $\hat{\sigma}_{e_d}^{2*}$ on the logit-scale.
2. For $b = 1, \dots, B$
 - (a) Generate sampling errors $e_d^{*(b)} \stackrel{\text{ind}}{\sim} N(0, \hat{\sigma}_{e_d}^{2*})$ and random effects $v_d^{(b)} \stackrel{\text{ind}}{\sim} N(0, \hat{\sigma}_v^2)$.

- (b) Simulate a bootstrap sample $\hat{\theta}_d^{Dir^*(b)} = x_d^T \hat{\beta} + v_d^{(b)} + e_d^{*(b)}$.
- (c) Calculate the true bootstrap population parameter $\theta_d^{(b)} = x_d^T \hat{\beta} + v_d^{(b)}$ on the transformed scale and back-transform with $\theta_d^{(b)} = \frac{\exp(\theta_d^{(b)})}{1 + \exp(\theta_d^{(b)})}$.
- (d) Estimate the bootstrap estimator of the model variance $\hat{\sigma}_v^{2(b)}$ using $\hat{\theta}_d^{Dir^*(b)}$ and $\hat{\sigma}_{e_d}^{2*}$.
- (e) Using $\hat{\sigma}_v^{2(b)}$ and $\hat{\theta}_d^{Dir^*(b)}$, estimate bootstrap estimators of the regression coefficients $\hat{\beta}^{(b)}$ and update the random effects $v_d^{(b)}$.
- (f) Determine the bootstrap estimator $\hat{\theta}_d^{FH^*(b)}$ with Equation (7) by using the estimates from the previous step and back-transform to the original scale by applying the bias-corrected back-transformation from Equation (8) to obtain $\hat{\theta}_d^{FH(b)}$.

3. Estimate the MSE:

$$\widehat{\text{MSE}}(\hat{\theta}_d^{FH}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\theta}_d^{FH(b)} - \theta_d^{(b)} \right)^2. \tag{9}$$

The performance of the presented bootstrap MSE estimator is evaluated in the simulation experiment in Section 4.

3.3. An Alternative Estimator from a Bayesian Perspective

As an alternative to the proposed methodology from a frequentist perspective [Fabrizi and Trivisano \(2016\)](#) presented a Bayesian Beta-regression model to get model-based estimators for the Gini concentration coefficients for small regions. This estimator is used in the simulation experiment in Section 4 as a comparative estimator. For a better understanding it is shortly introduced in the following. The sampling model with a Beta distribution as the underlying distribution for the direct estimator from Equation (2) for $d = 1, \dots, D$ domains is defined as follows:

$$\hat{\theta}_d^{Dir} \sim \text{Beta} \left(\frac{2\phi_d}{1 + \theta_d} - \theta_d, \frac{2\phi_d - \theta_d(1 + \theta_d)}{1 + \theta_d} \frac{1 - \theta_d}{\theta_d} \right),$$

with expected value $E(\hat{\theta}_d^{Dir} | \theta_d) = \theta_d$ and variance $V(\hat{\theta}_d^{Dir} | \theta_d) = 2\hat{\phi}_d^{-1} \theta_d^2 (1 + \theta_d^2)$, where ϕ_d is the precision parameter of the Beta distribution and can be estimated from the survey data and the variances of the direct estimator $\hat{\theta}_d^{Dir}$, which are assumed to be known here as well, inline with SAE literature. Using the variance smoothing model from Equation (3) ϕ_d can be estimated by $\hat{\phi}_d = \hat{\lambda} n_d$. For further details it is referred to [Fabrizi and Trivisano \(2016\)](#). The linking model with a logit link is defined as follows:

$$\log(\theta_d) = x_d^T \beta + v_d, \tag{10}$$

where x_d is a $p \times 1$ vector of domain-specific population covariates, β the corresponding $p \times 1$ vector of regression coefficients and v_d are the domain-specific random effects. To estimate the model in Equation (10) the specification of prior distributions for the random effects v_d , their variance σ_v^2 and the regression coefficients β are necessary. For β a normal prior with zero mean and large variances can be suggested: $\beta \sim N(0, kI)$, with $k = 100$ and I is the $p \times p$ identity matrix. For the random effects and their variance various prior

specifications are possible. In the simulation experiment in Section 4 the following prior distribution is assumed because it proved to be preferable to other prior distributions according to [Fabrizi and Trivisano \(2016\)](#): $v_d \sim N(0, \sigma_v^2)$ with $\sigma_v^2 \sim \text{half-}t(\nu = 3, A = 1)$, where ν are the degrees of freedom and A is the scale parameter. For the other possible specifications it is referred to [Fabrizi and Trivisano \(2016\)](#). The posterior distributions of the Gini coefficients are approximated by a MCMC algorithm, from which one directly obtains the point estimate for θ_d and a corresponding uncertainty measure, usually the expected value and variance of the posterior distribution given the data.

4. Simulation Study

To evaluate the performance of the proposed estimators in Section 3 in terms of bias and accuracy a model-based simulation experiment is conducted. In particular, the performance of the point estimator compared to three alternative estimators is of interest, as well as the presented uncertainty measure. The simulation setup is based on the estimated parameters from Section 5 and was chosen to mimic real data. The data are created for $D = 89$ domains. For the data generation process of the true parameter of interest and its direct estimator, the model variance and sampling variances from the SOEP data from Section 5 are used. The true parameters of interest θ_d for $d = 1, \dots, 89$ domains are derived via $\text{logit}(\theta_d) = \beta_0 + \beta_1 x + v_d$ with $\beta_0 = -1.5$, $\beta_1 = 1$ and covariate $x \sim \mathcal{LN}(-0.5, 0.04)$ generated so that the true values lie in a range of realistic Gini coefficients. The random effects v_d follow a normal distribution $\mathcal{N}(0, 0.029)$, where the variance parameter equals the estimated model variance in Subsection 5.1. The direct estimates are generated as $\text{logit}(\hat{\theta}_d^{\text{Dir}}) = \beta_0 + \beta_1 x + v_d + e_d$, with $e_d \sim \mathcal{N}(0, \sigma_{e_d}^2)$ where $\sigma_{e_d}^2$ are the direct variances on the logit-scale of the 89 observed SPRs from Section 5. They are listed in Table 8 in the Appendix (Section 7). The distributions of the given and resulting parameters in the simulation are reported in [Table 3](#). The data scenario was generated for $R = 1,000$ simulation runs.

The performance of the proposed bias-corrected estimator from Equation (8), denoted by logit FH.bc, is evaluated in comparison to three estimators: To a logit-transformed FH estimator with a naive back transformation using the inverse of the logit function (logit FH.naive), to the usual FH estimator (FH), and to the estimator proposed by [Fabrizi and Trivisano \(2016\)](#) and shortly introduced in Subsection 3.3. In the MCMC algorithm for the latter, a sample of 10,000 draws, with a preceding burn-in phase of 20,000 draws was used and the code provided by [Fabrizi et al. \(2016\)](#) was utilized to implement the estimator. The performance of the estimators is assessed by the distribution over the domains of the domain-specific absolute bias (ABias) and root mean squared error (RMSE), given as follows:

Table 3. Summary of parameters in the simulation setting.

	Min	1stQ	Median	Mean	3rdQ	Max
θ_d	0.206	0.264	0.298	0.294	0.319	0.396
$\hat{\theta}_d^{\text{Dir}}$	0.141	0.250	0.295	0.293	0.332	0.448
$\hat{\theta}_{e_d}^{\text{Dir}}$	0.141	0.250	0.295	0.293	0.332	0.448
γ_d	0.077	0.356	0.454	0.453	0.567	0.813
x	0.547	0.583	0.598	0.601	0.616	0.670

$$\text{ABias}(\hat{\theta}_d) = \left| \frac{1}{R} \sum_{r=1}^R (\hat{\theta}_{d,r} - \theta_{d,r}) \right|, \quad \text{RMSE}(\hat{\theta}_d) = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\theta}_{d,r} - \theta_{d,r})^2}, \quad (11)$$

where $\hat{\theta}_{d,r}$ denotes the estimator of the target indicator in domain d and replication r and $\theta_{d,r}$ the true value. Table 4 reports the distributions of the domain-specific ABias and RMSE over domains for the evaluated estimators. Starting with the bias it can be noted that the estimators, which use a logit transformation (Bayesian Beta, logit FH.bc and logit FH.naive) outperform the FH estimator (FH) without a transformation, which is a natural result due to the data generating process. Looking specifically at logit FH.bc and logit FH.naive, the reduction in bias due to the bias-corrected back-transformation is noticeable across the entire range of the distribution. Comparing the two median values, the use of logit FH.bc resulted in a 30% reduction in the median value of logit FH.naive. Further the results of the proposed bias-corrected estimator are comparable to those of the Bayesian estimator. In terms of efficiency, the four estimators provide very similar results with negligible differences. It is worth mentioning here that the bias-corrected back-transformation does not lead to a loss of efficiency and that the performance is similar to that of the Bayesian estimator proposed by Fabrizio and Trivisano (2016). Since in the data generating process the logit transformation is used, the comparison of the three estimators which use a logit-link is in that sense fair, that this refers to their use-case. Further the simulated direct estimators lie within a range of realistic values for Gini coefficients, and are not at the edges of the distribution, where a higher gain of the bias-corrected back-transformation compared to the naive can be expected. Only the comparison to the standard FH estimator is somewhat unfair, since the data scenario does not fit the untransformed FH model. Nevertheless, the comparison is of interest, since this approach corresponds to the simplest and is mainly used in practice. To investigate whether the differences between the methods are a result of the SAE estimators themselves or may be within a simulation-induced margin of error, the Monte Carlo error (MCE) is estimated with a Jackknife estimator following Koehler et al. (2009). The distributions of MCEs of the quantities of interest presented in Equation (11) are given in Table 9 in the appendix. Since the distributions across the domains of each method per quantity are very similar, it can be concluded that the differences from Table 4 are effective and not attributable to a MCE.

Next, the bootstrap MSE estimator from Equation (9) is examined for the estimator defined in Equation (8). It is denoted by $\widehat{\text{MSE}}_{d,r}$ for domain d of simulation run r . The

Table 4. Summary over domains of absolute bias and RMSE.

	Estimator	Min	1stQ	Median	Mean	3rdQ	Max
10 ³ x ABias	Bayesian Beta	0.005	0.241	0.528	0.69	1.126	2.398
	FH	0.099	0.686	1.408	1.457	2.033	4.024
	logit FH.bc	0.031	0.238	0.548	0.589	0.840	1.790
	logit FH.naive	0.042	0.338	0.778	0.827	1.188	2.447
10 ³ x RMSE	Bayesian Beta	15.320	23.260	26.290	26.330	28.970	34.680
	FH	15.220	23.200	26.940	26.630	29.320	36.730
	logit FH.bc	15.270	23.150	26.240	26.210	28.830	34.250
	logit FH.naive	15.260	23.160	26.260	26.220	28.830	34.270

estimator was calculated with $B = 500$ bootstrap replications in each simulation run. Its performance is evaluated comparing the estimated and the RMSE defined in Equation (11), which is treated as the true RMSE. As a measure of bias the relative bias (RB RMSE) is chosen, which is defined as follows:

$$\text{RB RMSE}(\hat{\theta}_d) = \frac{\sqrt{\frac{1}{R} \sum_{r=1}^R \widehat{\text{MSE}}_{dr} - \text{RMSE}(\hat{\theta}_d)}}{\text{RMSE}(\hat{\theta}_d)}.$$

Table 5 reports the distributions of the domain-specific RB RMSE over domains. It can already be seen that the percentage values are within an acceptable and common range for MSE estimators with a median relative bias of -1.1%. To have a closer look on the performance of the bootstrap MSE estimator with a bias-corrected back-transformation the estimated and true RMSE values per domain are plotted in Figure 2. The domains are ordered by decreasing sampling variances, which were used to construct the direct estimators. First, it can be observed that as the sampling variance decreases, the true RMSE also decreases, since a lower sampling variance is usually associated with a higher sample size and thus a lower RMSE. Second, the estimated RMSE tracks this behavior very well and thus captures the true uncertainty of the estimate in this setting. In summary, the bias-correction in the back-transformation is advantageous over the naive back-transformation in the given setting based on real data. Furthermore, the bootstrap MSE estimator leads to good results and provides a good estimate for the uncertainty.

5. Application to German Spatial Planning Regions

In this section, the methodology presented in Section 3 is illustrated using the data described in Section 2. In particular, the logit-transformed FH model with a bias-corrected back-transformation is used to estimate Gini coefficients for German SPRs, which are the domains in this application. At the same time, the advantage of using model-based small area methods in terms of increased accuracy is demonstrated. The SOEP sample used here contains data for 89 out of 96 SPRs in Germany with a total sample size of about 3,100

Table 5. Summary over domains of relative bias of estimated RMSE of logit FH.bc.

	Min	1stQ	Median	Mean	3rdQ	Max
RB RMSE [%]	-8.746	-3.434	-1.132	-0.836	1.450	8.283

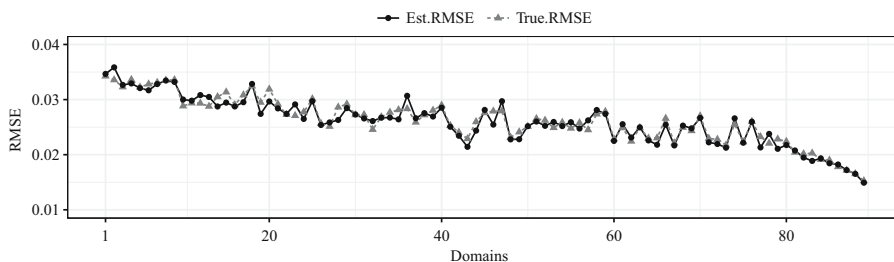


Fig. 2. Estimated and true RMSE of logit FH.bc. Domains are ordered by decreasing sampling variances.

households. In this application the Gini coefficients for the equivalised disposable household income are estimated. Since income distributions often have a heavy right-hand tail, the sensitivity of inequality measures to outliers based on those variables is discussed in [Alfons et al. \(2013\)](#) and [Cowell and Flachaire \(2007\)](#). The Gini coefficient is especially affected by extreme outliers and [Alfons et al. \(2013\)](#) therefore propose a Pareto tail modeling, which is also applied here. In this case, observations in the income distribution that are above a threshold, that is, the scale parameter of the Pareto distribution determined according to [Van Kerm \(2007\)](#), and are additionally extreme for the Pareto distribution are identified as outliers. These outliers are replaced by values of the underlying theoretical Pareto distribution. This approach was implemented by [Alfons and Templ \(2013\)](#) in the `laeken` R-package. In the whole sample, 65 households lie in the upper tail of the distribution of which in total two households from one SPR each (Cologne, Southern Upper Rhine) are identified as outliers and are replaced. The Gini coefficients for the SPRs are estimated using the direct estimator $\hat{\theta}_d^{Dir}$ from Equation (2) proposed by [Fabrizi and Trivisano \(2016\)](#). The sampling variances $\theta_{e_d}^2$ are estimated with the naive bootstrap procedure according to [Alfons and Templ \(2013\)](#) and implemented in the R package `laeken`. Following [Fabrizi and Trivisano \(2016\)](#) the variance smoothing model from Equation (3) was estimated to further smooth and stabilize the variances. Afterwards the smoothed sampling variances are brought to the logit scale with Equation (5).

5.1. Model Selection and Validation

Before moving to the discussion of model-based estimates of Gini coefficients obtained with Equations (7) and (8), the variable selection and testing of model assumptions using diagnostics is reviewed. From the set of possible covariates for predicting Gini coefficients and improving accuracy given in [Table 2](#), reasonable covariates are selected using an approach developed especially for FH models. [Marhuenda et al. \(2014\)](#) discuss various methods for FH model selection which are variants of common criteria like the Akaike Information criterion (AIC) and Kullback symmetric divergence criterion (KIC) and argue that common AIC over-parameterize FH models. They conclude, that a KIC bootstrap variant (KICb2) is the best selection criterion for FH models. Therefore a step-wise selection procedure with KICb2 criterion proposed by [Marhuenda et al. \(2014\)](#) with $B = 300$ bootstrap replications was applied, which is implemented in the R-package `emd.i` ([Kreutzmann et al. 2019](#)). The model selection was done with $\text{logit} \left(\hat{\theta}_d^{Dir} \right)$ as dependent variable and the transformed direct variances $\theta_{e_d}^2$. The final model includes only the variable $\log(\text{GDP per resident})$, which has an estimated positive effect. This is consistent with the hypothesis of [Perugini and Martino \(2008\)](#) that an increase in the regional level of development, with GDP serving as a proxy for economic development, promotes income inequality. The predictive power of the model is evaluated using an adjusted R^2 specifically for FH models proposed by [Lahiri and Suntornchost \(2015\)](#), which incorporates the variability of the sampling error. The model yields only a value of 16%, which is comparatively low, nevertheless the main goal of model-based small area methods, namely the gain in accuracy for small sample sizes, can be achieved, as can be seen in the next section. The model assumptions of normally distributed residuals and random effects are tested with the Shapiro-Wilk test and yield p-values of 0.854 and 0.147,

respectively, thus normality cannot be rejected at a significance level of 5%. The model variance estimated using the REML method is $\hat{\sigma}_v^2 = 0.029$ and is used in Section 4 as part of the data generating process.

5.2. Gain in Accuracy

Before looking at the model-based estimates of the Gini coefficients the gain in accuracy compared to the direct estimator is examined. The coefficients of variation (CV) per SPR for the proposed model-based estimator (logit FH.bc) and the direct estimator (Direct) are reported in Figure 3, where the SPRs are ordered by increasing sample sizes, starting with the OOS SPRs. The uncertainty of the bias-corrected logit-transformed FH estimator from Equation (8) is measured using the bootstrap algorithm presented in Section 3.2 with $B = 500$ bootstrap replications. The gain in efficiency is achieved for all SPRs as the CVs of the model-based estimators are always smaller than of the direct ones with a decreasing difference with higher sample sizes. This behavior is to be expected, as direct estimates become more reliable with higher sample sizes thus more weight is put on the direct component. For 13 of the 89 observed SPRs, the CV can be moved from above 20% to below this threshold using the model-based estimator. The threshold of 20% is a common value up to which estimates are considered reliable.

Table 6 shows the distribution of the estimated Gini coefficients and the corresponding CVs. The first observation is that the distribution of the direct estimator across SPRs is wider than that of the modelbased estimator, while the mean and median values of the distribution correspond to each other. This is in line with the expectation that the model-based estimates should be consistent with the direct estimates but more precise. The expected shrinkage to the mean effect can additionally be seen in Figure 4, where the direct estimates are plotted against the model-based estimates. It can be observed that the

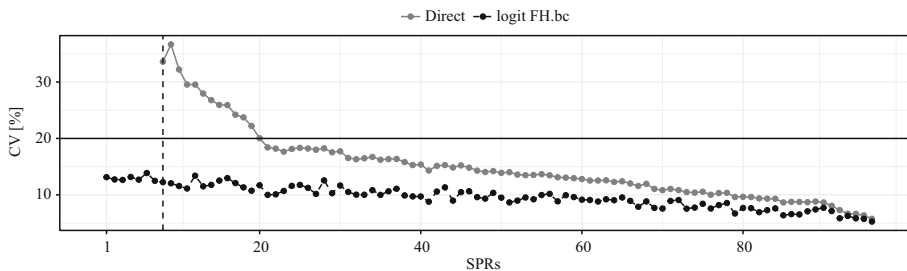


Fig. 3. CVs of Direct and logit FH.bc. SPRs are ordered by increasing sample sizes, OOS SPRs first.

Table 6. Summary of point estimators and corresponding CVs [%] over SPRs, OOS SPRs in separate lines.

	Min	1stQ	Median	Mean	3rdQ	Max
Direct	0.1674	0.2313	0.2631	0.2706	0.3031	0.4321
logit FH.bc	0.2112	0.2484	0.2657	0.2691	0.2884	0.3568
logit FH.bc OOS	0.2428	0.2493	0.2503	0.2543	0.2614	0.2656
CV Direct	5.75	10.48	13.66	14.99	17.54	36.66
CV logit FH.bc	5.26	7.74	9.53	9.38	10.63	13.39
CV logit FH.bc OOS	12.48	12.68	12.73	12.96	13.15	13.86

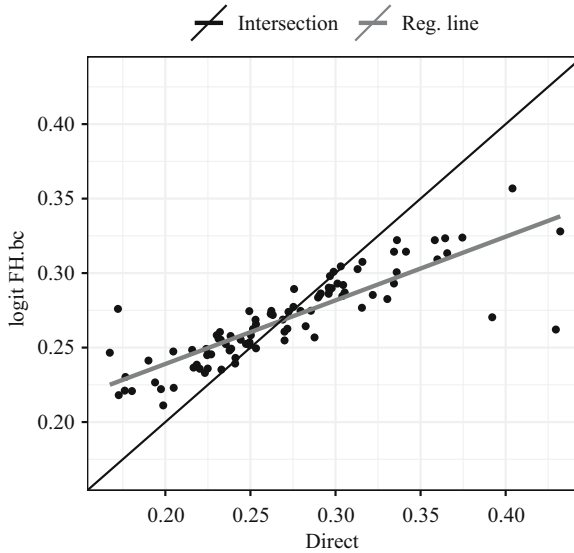


Fig. 4. Direct vs. model-based estimated Gini coefficients.

SPRs with a low direct estimate correspond to a higher model-based estimate and vice versa, indicating the regression to the mean. Examination of the OOS SPRs in Table 6 shows that the point estimates lie in the middle of the distribution of model-based estimates for observed SPRs. The CVs are instead at the high end of the distribution, which makes sense considering that these observations were not used to estimate the model.

To further investigate the quality of the model-based estimator, a closer look can be taken at Figure 5. There, the shrinkage factor $\hat{\gamma}_d$ from Equation (7), which indicates how much the direct component is weighted, is presented for each SPR with the corresponding sample size. On the x-axis are the SPRs ordered by decreasing sample sizes. It can be observed that in SPRs with higher sample sizes, the direct component is weighted more heavily, so that direct estimates and model-based estimates are very similar for SPRs with larger sample sizes. While the model-based estimator is more synthetic at smaller sample sizes.

5.3. Small Area Estimates

The regional distribution of the Gini coefficients estimated using the presented methodology for the 96 SPRs is mapped in Figure 6. The regional heterogeneity of income

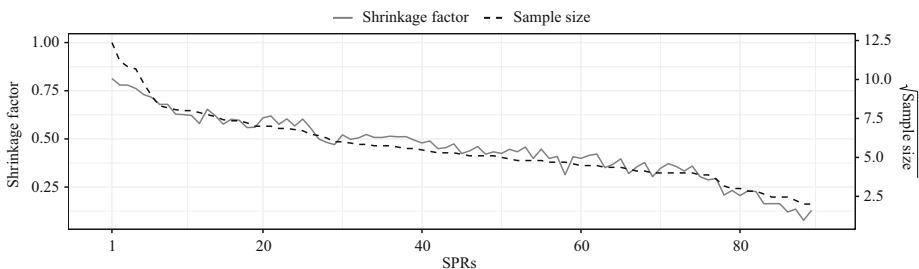


Fig. 5. Shrinkage factor $\hat{\gamma}_d$ and sample sizes per SPR. SPRs are ordered by decreasing sample sizes.

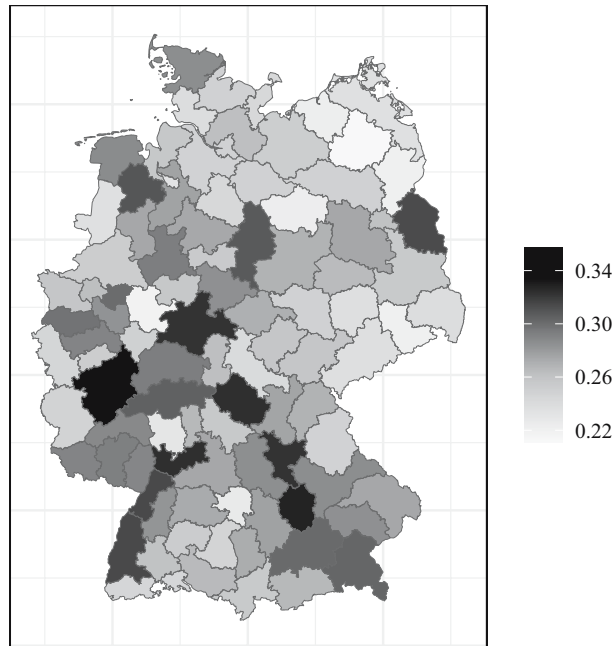


Fig. 6. Model-based estimates of the Gini coefficients for SPRs.

inequality within a region can be observed similar to the map in [Figure 1](#). [Figure 6](#) shows a similar pattern to [Goebel and Frick \(2005\)](#) in that income inequality is still lower in eastern Germany than in the west, although different levels of inequality are estimated within the eastern regions. In the rural SPRs of the Northeast, inequality is lower than in the Baltic region. The estimated Gini coefficient of the SPR east of Berlin (Oderland-Spree) is relatively high compared to neighboring SPRs. This maybe due to a mixture of rural and urban SPRs next to Berlin and, according to [Perugini and Martino \(2008\)](#), to the coexistence of specific and mobile labor segments. Furthermore, taking into account the results of [Immel and Peichl \(2020\)](#) that in these regions the share of the lowest-income 40% of households is relatively high compared to the rest of Germany. Likewise, the share of the highest-income top 10% is not exceptionally low, probably due to proximity to Berlin. This mix could lead to higher income inequality. The estimated Gini coefficient for Berlin is 0.26, which is similar to the value of 0.28 reported by [OECD \(2013\)](#) for 2013. A more general result, that the northern regions of West Germany tend to have lower Gini coefficients than the regions in the south and center, could be explained by [Immel and Peichl \(2020\)](#)'s findings that disproportionately few of the top 10% income earners live in the north of West Germany and disproportionately many in the south. The highest estimated Gini coefficient is for the SPR Central Rhine-Westerwald, with the city of Koblenz at its center, surrounded by more suburban SPRs. According to [Immel and Peichl \(2020\)](#), the city of Koblenz has a relatively high share of top 10% highest income households, which could be the driver of income inequality in this region. In general, it can be noted that where [Immel and Peichl \(2020\)](#) identify a high share of the highest-income 10%, income inequality also tends to be rather high.

6. Concluding Remarks

Measuring inequality at a regionally detailed level within counties and municipalities can provide deep insight into the income and wealth structures of these entities and can serve policymakers to target policies, taxation and funding to address inequality. A common indicator for measuring inequality is the Gini coefficient, which can be applied equally to income before and after taxes or to the value of wealth. The approach presented provides model-based estimates of the Gini coefficients at a regionally detailed level, which entails a gain in precision for small sample sizes compared to direct estimates based only on survey data. To achieve this, additional data sources and information from other domains are used in addition to the survey data. As an alternative when micro-data is not available, an area-level model, namely a logit-transformed FH model, is applied to the nonlinear indicator of interest. To avoid a bias when transforming back from the logit scale to the original, a bias-corrected back-transformation is used, which is also incorporated into the parametric bootstrap to measure the uncertainty of the estimate. The methodology presented is a straightforward extension of elaborated results for the transformed FH-model, can be easily integrated into existing SAE software, such as the R-package `emdi` (Kreutzmann et al. 2019), and poses no computational challenges. The validity of the approach is demonstrated in a model-based simulation, where the point estimator also performs similarly well to the Bayesian approach of Fabrizi and Trivisano (2016) chosen for comparison. The methodology is illustrated by means of an example for German SPRs using survey data from the SOEP and data from the 2011 Census. The analysis shows that there are intra-regional differences in income inequality and the proposed model-based methodology has achieved the desired gain in precision. The approach can be readily applied to estimate Gini coefficients for other regions, sub-populations, or survey data.

For future research, the methodology could be extended to the use of survey data where the data have been imputed multiple times by the data provider due to item non-response. The approach of Kreutzmann et al. (2022), which uses multiply imputed data from the Household Finance and Consumption Survey to estimate wealth averages, could therefore be extended to nonlinear indicators and appropriate transformations to allow Rubin's pooling rules (Rubin 1987) for multiply imputed data to be applied. Esteban et al. (2012) study area-level time models for nonlinear indicators such as poverty incidence and poverty gap. This approach could be transferred to also obtain time-stable estimates of inequality measures such as the Gini coefficient. Furthermore, the multivariate FH model proposed by Benavent and Morales (2016) could be extended for nonlinear indicators to jointly estimate Gini coefficients for multiple panel waves. Moreover, as mentioned in the introduction, other transformations could be used instead of the logit transformation as long as the estimated Gini coefficients are between zero and one. In any case, the variances of the direct estimator on the transformed scale are needed, and a suitable backtransformation for the estimated model-based Gini coefficients is required. Derivation of methodologies for for example probit or complementary log-log transformation could be part of further research.

7. Appendix

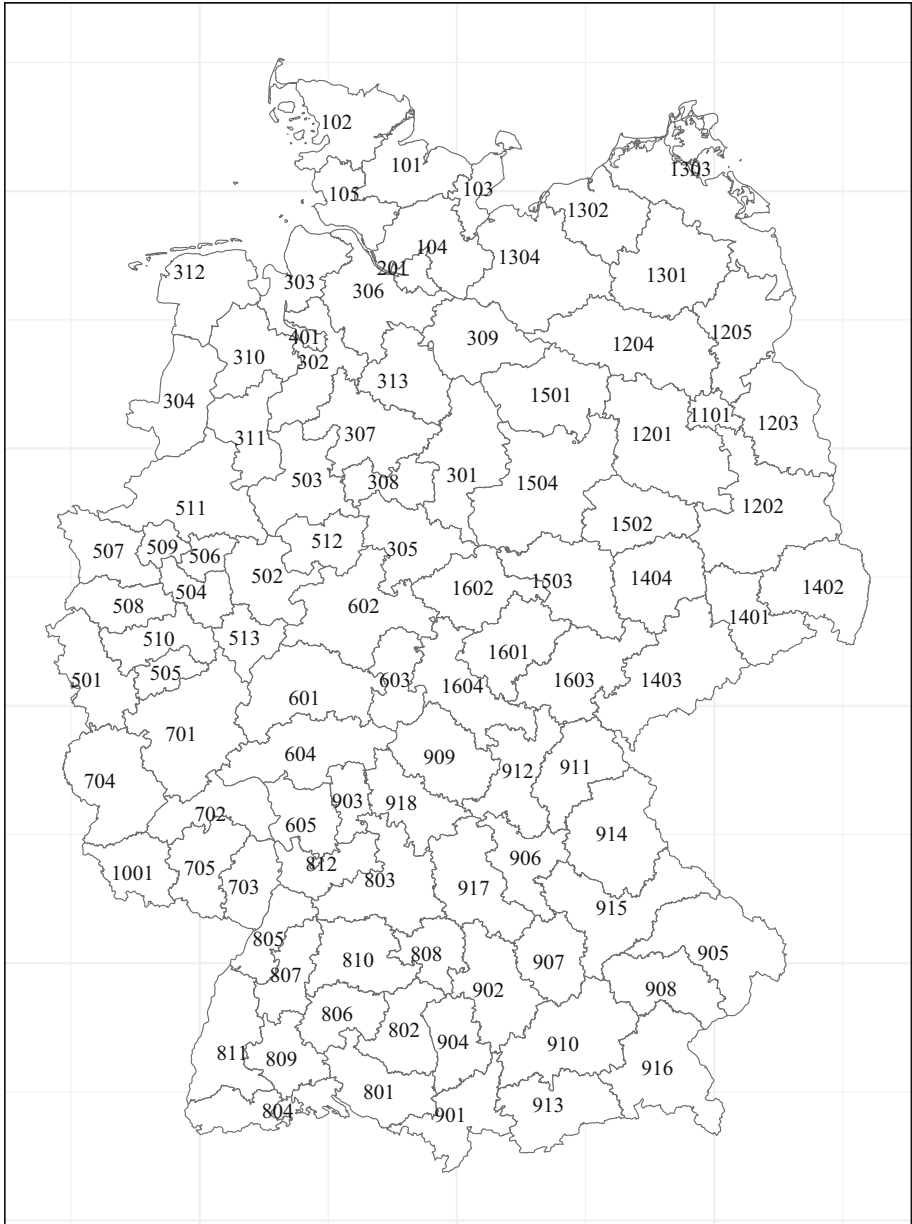


Fig. 7. SPR labels

Source: Bundesinstitut für Bau-, Stadt-, und Raumforschung (2021).

Table 7. Official SPR names and labels.

SPR	Name	SPR	Name
101	Schleswig-Holstein Mitte	806	Neckar-Alb
102	Schleswig-Holstein Nord	807	Nordschwarzwald
103	Schleswig-Holstein Ost	808	Ostwürttemberg
104	Schleswig-Holstein Süd	809	Schwarzwald-Baar-Heuberg
105	Schleswig-Holstein Süd-West	810	Stuttgart
201	Hamburg	811	Südlicher Oberrhein
301	Braunschweig	812	Unterer Neckar
302	Bremen-Umland	901	Allgäu
303	Bremerhaven	902	Augsburg
304	Emsland	903	Bayerischer Untermain
305	Göttingen	904	Donau-Iller (BY)
306	Hamburg-Umland-Süd	905	Donau-Wald
307	Hannover	906	Industrieregion Mittelfranken
308	Hildesheim	907	Ingolstadt
309	Lüneburg	908	Landshut
310	Oldenburg	909	Main-Rhön
311	Osnabrück	910	München
312	Ost-Friesland	911	Oberfranken-Ost
313	Südheide	912	Oberfranken-West
401	Bremen	913	Oberland
501	Aachen	914	Oberpfalz-Nord
502	Arnsberg	915	Regensburg
503	Bielefeld	916	Südostoberbayern
504	Bochum/Hagen	917	Westmittelfranken
505	Bonn	918	Würzburg
506	Dortmund	1001	Saar
507	Duisburg/Essen	1101	Berlin
508	Düsseldorf	1201	Havelland-Fläming
509	Emscher-Lippe	1202	Lausitz-Spreewald
510	Köln	1203	Oderland-Spree
511	Münster	1204	Prignitz-Oberhavel
512	Paderborn	1205	Uckermark-Barnim
513	Siegen	1301	Mecklenburgische Seenplatte
601	Mittelhessen	1302	Mittleres Mecklenburg/Rostock
602	Nordhessen	1303	Vorpommern
603	Osthessen	1304	Westmecklenburg
604	Rhein-Main	1401	Oberes Elbtal/Osterzgebirge
605	Starkenburger	1402	Oberlausitz-Niederschlesien
701	Mittelrhein-Westerwald	1403	Südsachsen
702	Rheinhessen-Nahe	1404	Westsachsen
703	Rheinpfalz	1501	Altmark
704	Trier	1502	Anhalt-Bitterfeld-Wittenberg
705	Westpfalz	1503	Halle/S.
801	Bodensee-Oberschwaben	1504	Magdeburg
802	Donau-Iller (BW)	1601	Mittelthüringen
803	Franken	1602	Nordthüringen
804	Hochrhein-Bodensee	1603	Ostthüringen
805	Mittlerer Oberrhein	1604	Südthüringen

Table 8. Direct variances on logit-scale of 89 observed SPRs.

SPR	$\sigma_{e_d}^2$	SPR	$\sigma_{e_d}^2$	SPR	$\sigma_{e_d}^2$	SPR	$\sigma_{e_d}^2$	SPR	$\sigma_{e_d}^2$
101	0.1949	504	0.1463	803	0.1778	911	0.1678	1404	0.1337
102	0.1036	505	0.1171	804	0.2092	912	0.4429	1501	0.2102
103	0.2641	506	0.1508	805	0.1317	914	0.1665	1502	0.1952
104	0.4305	507	0.1846	806	0.124	915	0.3350	1503	0.1628
105	0.3098	508	0.0817	807	0.3318	916	0.1400	1504	0.1635
201	0.1679	509	0.1687	808	0.2026	917	0.2684	1601	0.1794
301	0.1987	510	0.1489	809	0.3849	918	0.2056	1602	0.4599
302	0.258	511	0.1380	810	0.1172	1001	0.1711	1603	0.1075
304	0.1901	512	0.2584	811	0.1328	1101	0.0908	1604	0.2188
305	0.1387	601	0.2484	812	0.1767	1201	0.1984		
306	0.2301	602	0.1460	901	0.1668	1202	0.1365		
307	0.1931	603	0.5892	902	0.1524	1203	0.1515		
308	0.2093	604	0.0953	903	0.2411	1205	0.3145		
310	0.1808	605	0.1855	904	0.2236	1301	0.2276		
311	0.1887	701	0.1451	905	0.2047	1302	0.1896		
312	0.3857	702	0.2318	906	0.1310	1303	0.3853		
401	0.2335	703	0.1867	907	0.2520	1304	0.1725		
501	0.1741	704	0.2290	908	0.2093	1401	0.1997		
502	0.2215	705	0.1999	909	0.1703	1402	0.1657		
503	0.1334	802	0.3116	910	0.0906	1403	0.1383		

Table 9. Distributions of MCEs of the ABias and RMSE values.

	Estimator	Min	1stQ	Median	Mean	3rdQ	Max
$10^3 \times \widehat{\text{MCE}}(\text{ABias})$	Bayesian Beta	0.456	0.726	0.82	0.815	0.916	1.097
	FH	0.480	0.734	0.849	0.841	0.926	1.155
	logit FH.bc	0.483	0.732	0.823	0.826	0.910	1.084
	logit FH.naive	0.483	0.733	0.830	0.827	0.910	1.083
$10^3 \times \widehat{\text{MCE}}(\text{RMSE})$	Bayesian Beta	0.329	0.526	0.598	0.591	0.664	0.793
	FH	0.326	0.523	0.603	0.600	0.664	0.932
	logit FH.bc	0.328	0.524	0.593	0.590	0.659	0.791
	logit FH.naive	0.329	0.525	0.595	0.592	0.662	0.794

8. References

- Alfons, A., and M. Templ. 2013. "Estimation of social exclusion indicators from complex surveys: The R package laeken." *Journal of Statistical Software* 54: 1–25. DOI: <https://doi.org/10.18637/jss.v054.i15>.
- Alfons, A., M. Templ, and P. Filzmoser. 2013. "Robust estimation of economic indicators from survey samples based on pareto tail modeling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62: 271–286. DOI: <https://doi.org/10.1111/j.1467-9876.2012.01063.x>.
- Benavent, R., and D. Morales. 2016. "Multivariate Fay Herriot models for small area estimation." *Computational Statistics and Data Analysis* 94: 372–390. DOI: <https://doi.org/10.1016/j.csda.2015.07.013>.
- Braml, M., and G. Felbermayr. 2018. *Regionale Ungleichheit in Deutschland und der EU: Was sagen die Daten?* ifo Schnelldienst 71: 37–49. Available at: <https://www.ifo.de/>

- publikationen/2018/aufsatz-zeitschrift/regionale-ungleichheit-deutschland-und-der-eu-was-sagen-die (accessed August 2021).
- Bundesinstitut für Bau-, Stadt-, und Raumforschung. 2021. *Indikatoren und Karten zur Raum- und Stadtentwicklung*. Datenlizenz Deutschland – Namensnennung – Version 2.0. Available at: <http://www.inkar.de/> (accessed May 2022).
- Casas-Cordero, C., J. Encina, and P. Lahiri. 2016. "Poverty mapping for the chilean comunas. In *Analysis of Poverty Data by Small Area Estimation*, edited by M. Pratesi: 379–404. John Wiley and Sons. DOI: <https://doi.org/10.1002/9781118814963.ch20>.
- Citro, C.F., and G. Kalton. 2000. *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. Washington, DC: The National Academies Press. DOI: <https://doi.org/10.17226/10046>.
- Cowell, F., and E. Flachaire 2007. "Income distribution and inequality measurement: The problem of extreme values." *Journal of Econometrics* 141: 1044–1072. DOI: <https://doi.org/10.1016/j.jeconom.2007.01.001>.
- Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro level estimation of poverty and inequality." *Econometrica* 71: 355–364. DOI: <https://doi.org/10.1111/1468-0262.00399>.
- Esteban, M., D. Morales, A. Perez, and L. Santamaria. 2012. "Small area estimation of poverty proportions under area-level time models." *Computational Statistics and Data Analysis* 56: 2840–2855. DOI: <https://doi.org/10.1007/s11749-019-00688-w>.
- Fabrizi, E., M.R. Ferrante, and C. Trivisano. 2016. "Bayesian beta regression models for the estimation of poverty and inequality parameters in small areas." In *Analysis of Poverty Data by Small Area Estimation*,. 299–314. John Wiley and Sons, Ltd, Hoboken, NJ, USA.
- Fabrizi, E., and C. Trivisano. 2016. "Small area estimation of the Gini concentration coefficient." *Computational Statistics and Data Analysis* 99: 223–234. DOI: <https://doi.org/10.1016/j.csda.2016.01.010>.
- Fay, R.E., and R. A Herriot. 1979. "Estimates of income for small places: An application of James Stein procedures to census data." *Journal of the American Statistical Association* 74: 269–277. DOI: <https://doi.org/10.2307/2286322>.
- Furceri, D., and J.D. Ostry. 2019. "Robust determinants of income inequality." *Oxford Review of Economic Policy* 35(3): 490–517. DOI: <https://doi.org/10.1093/oxrep/grz014>.
- Gini, C. 1912. *Variabilita emutabilita: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Bologna: C. Cuppini. 3–159.
- Goebel, J. 2017. *Informationen zu den SOEP-Geocodes (SOEP.V35)*. SOEP Survey Papers 407: Series D. Berlin: DIW/SOEP. Available at: https://www.diw.de/documents/publikationen/73/diw_01.c.571219.de/diw_ssp0407.pdf (accessed January 2021).
- Goebel, J., and J. Frick. 2005. "Regional income stratification in unified Germany using a Gini decomposition approach." *Regional Studies* 42: 555–577. DOI: <https://doi.org/10.1080/00343400701543181>.
- Gonzalez-Manteiga, W., M. J. Lombardia, I. Molina, D. Morales, and L. Santamaria. 2005. "Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model." *Computational Statistics and Data Analysis* 52: 5242–5252. DOI: <https://doi.org/10.1016/j.csda.2008.04.031>.

- Hadam, S., N. Wuerz, and A.-K. Kreutzmann. 2020. *Estimating regional unemployment with mobile network data for functional urban areas in Germany*. DOI: <https://doi.org/10.17169/refubium-26791>.
- Hagenaars, A., K. de Vos, and M.A. Zaidi. 1994. *Poverty Statistics in the Late 1980s: Research Based on Micro-data*. Office for Official Publications of the European Communities. Luxembourg.
- Immel, L., and A. Peichl 2020. "Regionale Ungleichheit in Deutschland: Wo leben die Reichen und wo die Armen." *ifo Schnelldienst* 73: 43–47. Available at: <https://www.ifo.de/DocDL/sd-2020-05-immel-peichl-regionale-ungleichheit.pdf> (accessed August 2021).
- Janicki, R. 2020. "Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates." *Communications in Statistics-Theory and Methods* 49: 2264–2284. DOI: <https://doi.org/10.1080/03610926.2019.1570266>.
- Jiang, J., and J. Rao. 2020. "Robust small area estimation: An overview." *Annual Review of Statistics and Its Application* 7: 337–360. DOI: <https://doi.org/10.1146/annurev-statistics-031219-041212>.
- Kara, S., S. Zimmermann, and SOEP Group. 2019. *SOEPcompanion (v34)*, v.2. SOEP Survey Papers 743: SeriesG. Berlin: DIW/SOEP. Available at: https://www.diw.de/-documents/publikationen/73/diw_01.c.611577.de/diw_ssp0588.pdf (accessed December 2020).
- Koehler, E., E. Brown, and S.J.P.A. Haneuse. 2009. "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses." *The American Statistician* 63(2), 155–162. DOI: <https://doi.org/10.1198/tast.2009.0030>.
- Kreutzmann, A.-K., P. Marek, M. Runge, N. Salvati, and T. Schmid. 2022. "The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany." *Journal of Applied Statistics* 49(13): 3278–3299. DOI: <https://doi.org/10.1080/02664763.2021.1941805>.
- Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. 2019. "The R package emdi for estimating and mapping regionally disaggregated indicators." *Journal of Statistical Software* 91: 1–33. DOI: <https://doi.org/10.18637/jss.v091.i07>.
- Lahiri, P., and J.B. Suntornc host. 2015. "Variable selection for linear mixed models with applications in small area estimation." *The Indian Journal of Statistics* 77: 312–320. DOI: <https://doi.org/10.1007/s13571-015-0096-0>.
- Liu, B., P. Lahiri, and G. Kalton. 2014. "Hierarchical bayes modeling of survey-weighted small area proportions." *Survey methodology* 40: 1–13. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14030-eng.pdf?st=0WhqkvP4> (accessed June 2022).
- Marhuenda, Y., D. Morales, and M. Pardo. 2014. "Information criteria for Fay-Herriot model selection." *Computational Statistics and Data Analysis* 70: 268–280. DOI: <https://doi.org/10.1016/j.csda.2013.09.016>.
- Molina, I. and J.N.K. Rao. 2010. "Small area estimation of poverty indicators." *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 38: 369–385. Available at: <http://www.jstor.org/stable/27896031> (accessed July 2020).

- Neves, A., D. Silva, and S. Correa. 2013. "Small domain estimation for the Brazilian service sector survey." *Estatística* 65: 13–37. Available at: <https://www.statistics.gov.hk/wsc/CPS003-P7-S.pdf> (accessed November 2019).
- OECD. 2011. *Income distribution database: by country - inequality*. Available at: <https://stats.oecd.org>. (accessed June 2021).
- OECD. 2013. *Regional well-being: Regional income distribution and poverty*. Available at: <https://stats.oecd.org> (accessed June 2021).
- Perugini, C., and G. Martino. 2008. "Income inequality within european regions: Determinants and effects on growth." *Review of Income and Wealth* 54: 373–406. DOI: <https://doi.org/10.1111/j.14754991.2008.00280.x>.
- Pfeffermann, D. 2013. "New important developments in small area estimation." *Statistical Science* 28(1): 40–68. DOI: <https://doi.org/10.1214/12-STS395>.
- Prasad, N., and J. Rao. 1990. "The estimation of the mean squared error of small-area estimators." *Journal of the American Statistical Association* 85(409): 163–171. DOI: <https://doi.org/10.2307/2289539>.
- Pratesi, M. 2016. *Analysis of Poverty Data by Small Area Estimation*. John Wiley and Sons, Inc, Hoboken, NJ, USA. DOI: <https://doi.org/10.1002/9781118814963>.
- Rao, J., and I. Molina. 2015. *Small Area Estimation*. John Wiley and Sons, Inc, Hoboken, NJ, USA.
- Rivest, L.-P., and N. Vandal. 2002. "Mean squared error estimation for small areas when the small area variances are estimated. In Proceedings of International Conference of Recent Advanced Survey Sampling, edited by J. Rao: 197–206. Ottawa, July 10-13, 2002. Available at: <https://www.mat.ulaval.ca/fileadmin/mat/documents/lrivest/Publications/64-RivestVandal2003.pdf>.
- Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, Hoboken.
- Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(4), 1163–1190. DOI: <https://doi.org/10.1111/rssa.12305>.
- Siegers, R., V. Belcheva, and T. Silbermann. 2020. *SOEP-core v35 documentation of sample sizes and panel attrition in the German socio-economic panel (SOEP), 1984 until 2018*. SOEP Survey Papers 826: Series C. Berlin: DIW/SOEP. Available at: https://www.diw.de/documents/publikationen/73/diw_01.c.745900.de/diw_ssp0826.pdf (accessed December 2020).
- Slud, E., and T. Maiti. 2006. "Mean-squared error estimation in transformed Fay-Herriot models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68: 239–257. DOI: <https://doi.org/10.1111/j.1467-9868.2006.00542.x>.
- Socio-Economic Panel. 2019. *Data for years 1984–2017, version 34, SOEP*. Socio-Economic Panel, Berlin. DOI: <https://doi.org/10.5684/soep.v34>.
- Statistische Ämter der Länder. 2021. *Volkswirtschaftliche gesamtrechnungen der länder – zusammenhänge, bedeutung, ergebnisse*. Available at: https://www.statistikportal.de/sites/default/files/2020-11/vgrdl_brochure_2020.pdf (accessed June 2022).

- Statistische Ämter des Bundes und der Länder. 2011a. Available at: <https://ergebnisse2011.zensus2022.de/datenbank/online/>. (accessed: January 2021).
- Statistische Ämter des Bundes und der Länder. 2011b. Regionaldatenbank Deutschland. Available at: <https://www.regionalstatistik.de/genesis/online> (accessed January 2021).
- Sugasawa, S., and T. Kubokawa. 2017. “Transforming response values in small area prediction.” *Computational Statistics and Data Analysis* 114: 47–60. DOI: <https://doi.org/10.1016/j.csda.2017.03.017>.
- Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla. 2018. “From start to finish: a framework for the production of small area official statistics.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(4), 927–979. DOI: <https://doi.org/10.1111/rssa.12364>.
- UN General Assembly. 2015. *Transforming our world: the 2030 agenda for sustainable development*, Available at: <https://www.refworld.org/docid/57b6e3e44.html> (accessed April 2023).
- Van Kerm, P. 2007. *Extreme incomes and the estimation of poverty and inequality indicators from EUSILC*. Working Paper Series 2007-01. Centre 39 d'Etudes de Populations, de Pauvrete et de Politiques Socio-Economiques International Network for Studies in Technology, Environment, Alternatives, 40 Development. Available at: <https://liser.elsevierpure.com/en/publications/extreme-incomes-and-the-estimation-of-poverty-and-inequality-indi> (accessed December 2020).
- Wang, J., and W.A. Fuller. 2003. “The mean squared error of small area predictors constructed with estimated area variances.” *Journal of the American Statistical Association* 98(463): 716–723. DOI: <https://doi.org/10.1198/016214503000000620>.
- You, Y., and B. Chapman. 2006. “Small area estimation using area level models and estimated sampling variances.” *Survey Methodology* 32(3): 97–103. Available at: <https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019263> (accessed March 2020).

Received February 2022

Revised July 2022

Accepted January 2023

Constructing Building Price Index Using Administrative Data

Masahiro Higo¹, Yumi Saita², Chihiro Shimizu³, and Yuta Tachi⁴

Improving the accuracy of deflators is crucial for measuring real GDP and growth rates. However, construction prices are often difficult to measure. This study uses the stratification and hedonic methods to estimate price indices. The estimated indices are based on the actual transaction prices of buildings (contract prices) obtained from the Statistics on Building Starts survey information from the administrative sector in Japan. Compared with the construction cost deflator (CCD), calculated by compounding input costs, the estimated output price indices show higher rates of increase during the economic expansion phase after 2013. This suggests that the profit surge in the construction sector observed in that period is not fully reflected in the CCD. Furthermore, the difference between the two “output-type” indices obtained by stratification and hedonic methods shrinks when the estimation methods are precisely configured.

Key words: Building price index; stratification method; hedonic method; Japan; administrative data.

1. Introduction

Improving the accuracy of deflators (price indices) is crucial for correctly measuring real GDP and growth rates. However, price indices often deviate from deflators that meet the System of National Accounts (SNA) international standards (Eurostat et al. 2009) due to difficult access to price information and technical limitations in indexing. This issue is particularly evident in the construction industry, as the contribution of construction investment in a country’s GDP is typically large. Hence, improving the accuracy of construction price indices is fundamental for statistical departments in various countries. For example, the Japanese Statistics Commission, the command post for statistics, together with the ministries and agencies that prepare statistics, has been working on reforms required by the government’s Council on Economic and Fiscal Policy and the Council for the Promotion of Statistical Reform. The “Basic Plan for the Development of Official Statistics” (the 3rd Basic Plan), approved by the Cabinet in March 2018, states that “from the perspective of improving the accuracy of real values (omission), research and empirical studies for the use of market-based prices for

¹ The University of Tokyo. 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan. Email: masahiro.higo@e.u-tokyo.ac.jp

² Kwansai Gakuin University, 1-155, Uegahara Ichiban-cho, Nishinomiya, Hyogo, 662-8501, Japan. Email: saita@kwansai.ac.jp

³ Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo, 186-8601, Japan. Email: c.shimizu@r.hit-u.ac.jp

⁴ Reitaku University, 2-1-1, Hikarigaoka, Kashiwa-shi, Chiba, 277-8686, Japan. Email: tachiyuta@gmail.com
Acknowledgments: We thank W. Erwin Diewert, Paul Schreyer, Marshall Reinsdorf, Niall O’Hanlon, Jens Mehrhoff, Alicia Rambaldi, Kate Burnett-Isaacs, Naohito Abe, Koji Nomura and Rhys Lewis for their advice and materials. This study was conducted in the office of the Statistics Committee. The authors also thank the Ministry of Land, Infrastructure, Transport, and Tourism for providing the data. This work gratefully acknowledges the support received from JSPS 20H00082.

construction and retail services (margins) will be promoted.” The report recognizes that improving the accuracy of the deflator (price index) in realizing nominal values will substantially improve the accuracy of GDP estimates.

In Japan, the Statistics Commission, the command post for statistics, together with the ministries and agencies that prepare statistics, has been working on reforms required by the government’s Council on Economic and Fiscal Policy and the Council for the Promotion of Statistical Reform. The “Basic Plan for the Development of Official Statistics” (the 3rd Basic Plan), approved by the Cabinet in March 2018, states that “from the perspective of improving the accuracy of real values (omission), research and empirical studies for the use of market-based prices for construction and retail services (margins) will be promoted.” The report recognizes that improving the accuracy of the deflator (price index) in realizing nominal values will substantially improve the accuracy of GDP estimates.

In Japan, there is no output-type construction price index based on actual transaction prices (contract prices) of buildings and civil engineering structures. The SNA’s deflator uses “input-cost” prices as an alternative price index based on the intermediate input and labor costs required for construction activities. However, the resulting construction deflator does not reflect changes in the profit margins of the construction sector, possibly causing errors in the real value of construction investment.

In this context, there is a need to develop an output-type construction price index. In our survey of previous studies, we could not find any that focused on estimation methods for output-based construction price indexes. We believe that this is due to the following reasons: (1) buildings are made-to-order and thus highly heterogeneous, making quality adjustment extremely difficult; and (2) strong data constraints make empirical research difficult as seen in the papers such as [Diewert and Shimizu \(2015, 2017, 2022\)](#), where they apply the Builders Model to the construction price index by separating land and buildings from property prices.

Some countries have adopted output-type construction price indices that directly measure output prices (contract prices), namely actual transaction prices of buildings and civil engineering structures, instead of the conventional input-cost price indices. Statistical authorities in Canada, Germany, the U.S., and the UK have already developed output price indexes, some of which are used as price indexes for the construction industry in GDP statistics (see, for example, [Office for National Statistics 2018](#)).

In this study, we attempt to create an output-type building price index directly measuring output prices by using large-scale administrative data from the Statistics on Building Starts, based on the Notification of Building Construction that must be submitted in Japan when constructing a building. To the best of our knowledge, this is the first study to construct a price index using these data. We propose a new approach that utilizes large-scale information collected by administrative agencies. This approach may be considered entirely novel, not only in Japan but also in other countries.

The remainder of this article is organized as follows. Section 2 summarizes the limits of the current input-cost construction price indices and provides an overview of three common approaches to creating output-type construction price indices. Section 3 provides an overview of the administrative data obtained from the Statistics of Construction Starts, explaining how the stratification and hedonic methods are used to create the proposed price index. Section 4 reports the estimation results of the output-type building price indices obtained through the stratification and hedonic methods, comparing them with the input-cost construction cost

deflator (CCD). Additional analyses and results are then proposed and discussed. Section 5 summarizes the study's findings and outlines the remaining issues for future research.

2. Output-Type Construction Price Indices

2.1. Factors of the Use of Input-Cost Price Indices in the Construction Sector

The SNA calculates real amounts by dividing nominal amounts, such as the production value, by a deflator (see [Cabinet Office \(2022\)](#) for detailed instructions on how to create an index in the case of Japan). A deflator is a price index calculated by continuously surveying the prices of goods and services with constant quality through repeated transactions, indexing the prices of each product, so that the base point is equal to 100, and weighing the price indices of individual products using weights corresponding to the transaction amounts. In Japan's SNA, the Consumer Price Index (CPI) produced by the Statistics Bureau of the Ministry of Internal Affairs and Communications; the Corporate Goods Price Index (CGPI) and the Services Producer Price Index (SPPI) produced by the Bank of Japan; and the Agricultural Price Index produced by the Ministry of Agriculture, Forestry and Fisheries are typically used. See [Cabinet Office \(2022\)](#) for detailed instructions on how to create an index.

However, in the construction sector, there is no market transaction price-based construction price index. In the case of goods and services covered by the CPI and the CGPI and SPPI, goods and services of the same quality are exchanged repeatedly; hence, a constant-quality price index may be created by continuously examining price trends. In contrast, buildings and civil engineering structures are custom-made products, and those of the same quality are not traded repeatedly. Therefore, it is not possible to continuously survey the market transaction prices (contract prices) of buildings and civil engineering structures of the same quality.

The SNA has developed an alternative, input-cost construction price index based on the input costs of buildings and civil engineering structures and used it as a deflator. The Bank of Japan's CGPI and the SPPI are employed for the materials used for assessing construction activities (intermediate inputs), and the Ministry of Health, Labor, and Welfare's (MHLW) Monthly Labor Survey is used to derive per capita wages in the construction industry, used for assessing labor costs (compensation of employees). In addition, the Ministry of Land, Infrastructure, Transport, and Tourism (MLIT) produces the CCD, one of Japan's most used input-cost construction price indices. Although we omit the details here and want readers to see [MLIT \(2021\)](#) for detailed information in Japanese and [MLIT \(n.d.a.\)](#) for general information in English, CCDs are made in a similar way to the SNA construction deflator except that they are based on Laspeyres formula. They are made of piling up input costs, material costs from CGPI, SPPI and wages from MHLW's Monthly Labor Survey, but profits of construction firms are not incorporated.

2.2. Factors Causing Bias in the Input-cost Construction Price Index

The current input-cost construction price indices, such as the SNA construction deflator, have two major limitations.

First, these indices only cover the intermediate inputs and compensation of employees. Intermediate inputs and compensation of employees accounted for 90% of the output of

construction in 2019. The profit generated by construction activities (operating surplus and mixed income), fixed capital depletion, and taxes (taxes imposed on “produced and imported goods” minus “subsidies”), which accounted for the remaining 10%, were not covered. The profit share of the construction sector, which is not covered, often causes errors in the “input cost” price index as it fluctuates significantly compared to input costs due to changes in the environment in which construction companies receive orders. Price changes due to a variation in the profit of the construction industry, not considered by the current input-cost construction price index, are assumed to be equal to those of intermediate inputs and compensation of employees. If these price changes are different, the resulting price index is substantially biased, as the coverage of the input-cost price index changes over time.

Second, the Monthly Labor Survey only reports the average per capita wages of construction workers, used as price data to assess labor costs; hence, it does not consider changes in the quality of labor, such as age, length of service, education, and employment status. Compensation of employees’ accounts for more than 30% of the construction industry’s output (35% in 2019). As [Fukao et al. \(2017\)](#) point out, its impact is significant and likely to influence changes in building quality. For instance, according to the Japan Industrial Productivity (JIP) Database 2021, the quality of labor in the construction industry (the average of building and civil engineering) has improved by 12% from 1994 to 2018. Moreover, [JILPT \(2020\)](#) estimates that the rate of increase in simple average wages and average wages with fixed attributes in the construction industry from the MHLW’s “Basic Survey on Wage Structure” shows a 13% increase from 1994 to 2019 due to improvements in the quality of work (The gap is 13% over 25 years).

2.3. How to Create An Output-Type Construction Price Index

To overcome insufficient coverage of the price indices and wage data with no fixed quality, it is necessary to create a price index based on market transaction prices, namely, an output-type index reflecting the output prices of buildings and civil engineering structures. [OECD \(1997\)](#) introduces the following six methods to construct a price index: (1) Model price method, (2) Quoted prices method, (3) List prices method, (4) Matched models method, (5) Stratification method, and (6) Hedonic method. Among these, the methods typically adopted in most countries are (1) the model price approach, (5) the per square meter (stratified) approach, and (6) the econometric (hedonic) approach. In this section, we summarize their characteristics. Furthermore, for house price indexes, [Hill \(2013\)](#) and [Silver \(2011\)](#) provide a comprehensive overview of quality-fixing methods, and [Hill et al. \(2018\)](#) compared the differences to price indexes by several methods, including the Stratification method (Mixed adjustment approach), the hedonic approach, and others.

2.3.1. Model Price Method

The model price method is a compiling method in which “models” of typical buildings and civil engineering structures are obtained by adding up the hypothetical prices of construction materials, labor, machinery, and equipment for each component, and adding the assumed profit of the construction company. This method has been used in the U.S., Canada and Germany, among others. In Japan, this method has been adopted for some items in the CGPI and the SPPI; however, it has not yet been adopted for the CCD and the SNA deflator.

For example, in the U.S., the U.S. Bureau of Labor Statistics purchases data on past construction projects from construction cost estimating companies, selecting representative construction projects for each region (Northeast, Midwest, South, and West) to construct a “building model.” The construction contract price is calculated by adding all the “assemblies” required to obtain the building model and all the “components” required for the assemblies. Furthermore, the U.S. BLS, together with experts (construction cost estimating companies), periodically reviews the building models to ensure that they are representative. Every month, when each contractor responds to the survey, they are asked to check in advance the factors affecting the determination of bid prices so that realistic prices are investigated. In addition, the collected prices are verified to ensure that they do not deviate from the actual prices, and cases in which the margin ratio fluctuates significantly due to changes in the content of the work (price changes due to quality changes) are discarded. In Canada, Statistics Canada designates a representative newly constructed building and asks construction companies the price change of each production factor from the previous quarter. In Germany, the Federal Statistical Office of Germany directly obtains the unit price of constructions, equivalent to the output price, from the surveyed companies.

However, this method has some limitations. First, the price indices may not reflect the transaction prices traded in actual markets if some of the production factors in the building or construction models are not representative. Second, as the modeled prices are not actual transaction prices, they may not accurately reflect the market situation. Third, the cost of producing these statistics is generally high since it is often necessary to hire experts for designing a standard model. Fourth, the burden on construction companies reporting hypothetical estimated prices is also high, which can lead to difficulty of Statistics bureaus to collect markup rate data. The last and not the least, quality adjustment of labor costs and estimation of user costs also involve difficulties in terms of data collection and analysis. In this regard, the U.S. BLS (U.S. Bureau of Labor Statistics), together with experts (construction cost estimating companies), periodically reviews the building models to ensure that they are representative. Every month, when each contractor responds to the survey, they are asked to check in advance the factors affecting the determination of bid prices so that realistic prices are investigated. In addition, the collected prices are verified to ensure that they do not deviate from the actual prices, and cases in which the margin ratio fluctuates significantly due to changes in the content of the work (price changes due to quality changes) are discarded.

2.3.2. Stratification Method

This method stratifies price data into multiple attributes, which have a large impact on prices (use, structure, construction method, building method, and region, among others), creating a price index based on the average price for each subdivided stratum. In this method, buildings and civil engineering structures with different qualities are regarded as having the same quality for price data belonging to the same stratum. This approach is commonly employed in existing price indices (See [Eurostat et al. \(2013\)](#) for further information).

In the stratification method, once the subdivision rules have been determined, the burden of producing price indices remains constant as that of tabulating ordinary statistical surveys, and price indices may be produced with a small workload. The degree of homogeneity of the price index increases with the availability of attribute data. In addition, unlike the model pricing approach, this method does not require extensive expertise in

building and civil engineering. However, increasing the number of attributes used for subdivision reduces the number of observations in the same stratum, often resulting in no price data in many strata (empty strata), generating bias and noise.

2.3.3. Hedonic Method

This approach estimates the price of a product by considering it as an aggregate of the values of various performances and functions (a bundle of attributes) and using regression analysis. The commodity price is expressed as an equation consisting of a bundle of attributes, and this equation is called a “hedonic function.” The regression analysis employs the collected price and attribute data, controlling for the effects of various attributes, and then creating a price index based on the estimated time-series dummies.

Among the construction-related deflators currently used for GDP statistics, only the U.S. uses the hedonic method. In the U.S., this approach is based on data from the Survey of Construction, a survey conducted by the U.S. Bureau of the Census on housing, including actual construction costs and data on the location, layout, and construction method of the housing sector.

In the hedonic method, it is easier to increase the degree of homogeneity compared to the stratification method because empty data can be handled by specifying dummy variables precisely. However, this approach requires knowledge of econometrics to estimate the hedonic function, and the estimation requires many attribute variables. When attributes are not available or are measured in an incorrect way, the model may be subject to omitted variable or other misspecification issues, which could lead to endogeneity problems. Furthermore, it is necessary to periodically re-estimate the hedonic function, increasing the burden of producing price indices. We outline the advantages and disadvantages of the hedonic method in the estimation of price indexes here. Advantages include: (1) As well as having a basis in economic theory and index theory, the theoretical biases of the hedonic method are clear, and (2) Since it makes it possible to control for the many characteristics of building, it enables the sorting of data into specialized indexes by purpose. Disadvantages include: (1) Since it is necessary to collect many characteristics, information-gathering costs are high, (2) In cases where it is not possible to collect important characteristics for determining property prices, one faces the problem of omitted variable bias, (3) In cases of strong heterogeneity, it may not be possible to control for quality. For further information on estimating price indices using the hedonic method, see [Eurostat et al. \(2013\)](#), [Benedetti et al. \(2022\)](#), for example.

3. Estimation of Output-Type Construction Price Indices Using the Administrative Data

3.1. Data

We construct an output-type construction price index for housing and non-housing buildings using the stratification and hedonic methods. Individual data from the Statistics of Building Starts are used as the source material (see [Table 1](#) for attribute items available in the Statistics).

In Japan, the Building Standard Law requires the owner to submit a Notification of Construction Work to the prefectural governor when constructing a building. The MLIT compiles the Notification of Construction Work and publishes the Statistics on Building Starts. These statistics comprise all building investments. The administrative data used in this study

Table 1. List of attribute items in the statistics of the building starts questionnaire.

Number	Item name	Sign and description
1	Year of survey	2005–2020
2	Survey month	01–12
3	Prefectural Number	01 to 47, Hokkaido to Okinawa
4	City, Town and Village Code	XXX
5	Intra-municipal serial number	XXXX
6	Scheduled construction period	01–99 (months)
7	The builder	1: country, 2: prefecture, 3: municipality, 4: company, 5: non-company organization, 6: individual
8	Structure	1: Wooden structure, 2: Steel-framed reinforced concrete structure, 3: Reinforced concrete structure, 4: Steel structure, 5: Concrete block structure, 6: Others
9	Building use	In addition to the classification based on the Standard Industrial Classification, classification is based on the use of offices and stores.
10	Construction type	1: New construction, 2: Extension, 3: Reconstruction
11	Capital stock classification	1: 10 million yen or less, 2: Over 10 million yen to 30 million yen or less, 3: Over 30 million yen to 100 million yen or less, 4: Over 100 million yen to 1 billion yen or less, 5: Over 1 billion yen (Only when the architect is “4: Company”)
12	City planning classification	1:Urbanized area, 2:Urbanized control area, 3:Undefined urban planning area, 4:Quasi-urban planning area, 5:Under urban planning area and quasi-urban planning area
13	Building classification	1 to 9: Enter a series of numbers when there are two or more buildings in one construction report. The same number should be entered for the same building, and “9” should be entered for all buildings above “9”
14	Small number	If there are two or more houses in one building with different use relationships, enter the series number.
15	Number of floors above ground in new construction	01-99 (only when the construction type is “1: New Construction”)
16	Number of basement floors in new construction	1 to 9 (only if the construction type is “1: New Construction”)
17	Site area of new construction	m ² (only if the construction type is “1: New Construction”)
18	Total floor area	(As there is no obligation to report building work if the area is less than 10 m ²) it takes a value of 11 m ² or more
19	Construction contract expenses	10,000 yen
20	Versatile or not	If 1, it indicates a multi-use building
21	Number of housing units to be removed	XXX
22	Relationship of use of retired housing	1: Owner-occupied house, 2: Rental house, 3: Salary housing
23	Building method	1: Conventional construction method, 2: Prefabricated construction method, 3: Frame wall construction method

Table 1. Continued

Number	Item name	Sign and description
24	Construction type	1: Newly established, 2: Other
25	Funds for new housing	1:Privately financed housing, 2:Public housing, 3:JHF housing, 4:Urban Renaissance Agency housing, 5:Others (only if the construction type is “1: New construction”)
26	Type of housing	1: Dedicated housing, 2: Conjoined housing, 3: Other housing
27	Building (e.g., house)	1: Single-family houses, 2: Row houses, 3: Apartment houses
28	Usage restrictions, limitations	1: Owner-occupied house, 2: Rental house, 3: Salaried house, 4: Condominium house
29	Number of housing units	XXXX
30	Total floor area of the house	m ²

Note: Variables 1 to 20 are used to describe both housing and non-housing buildings while those from 21 to 30 are only used for housing buildings.

Source: Compiled based on survey sheets from the MLIT’s Building Starts statistics.

cover 7.92 million cases from January 2005 to December 2020 (an average of approximately 500,000 cases per year). A large amount of information on building prices and attributes may be obtained from the administrative data (see Online Supplemental Data, Appendix 1 for details). We use the unit price per floor area of the construction contract expenses (construction contract expenses/total floor area) as the price information of the buildings. The unit price per floor area (#19 / #18 in the Table 2) is one of the most important indicators in the dataset. It shows wide range of variations based on the type of structures and characteristics of the buildings. Wood structure and Non-Housing S buildings show relatively cheap average price of 14.10–16.26 ten k. yen compared with the high price of 26 ten k. of Non-housing RC, 22.10 of Housing S and 20.99 of Housing RC. The standard deviation of prices stands at the highest in Non-Housing RC which implies large heterogeneity among buildings.

It should be noted that the construction contract expenses in the Statistics on Building Starts is the value at the time the notification of construction work is submitted, not the actual amount of construction expenses at the time of completion, nor is it converted to a progress basis. However, contracting practice observed in construction companies in Japan is rather different from those in other countries. In Japan, so called general construction companies give one stop services to customers, and changes in construction schedule or plan have less impact compared with those in other countries. According to the statistics by the MLIT, deviations in unit price per floor area of completed buildings from contracted prices remain at 2–3%, which means construction expenses at the time of completion don’t substantially deviate from those at the contract stage unlike other countries. That information is available at MLIT (n.d.b) but only published in Japanese.

3.2. Overview of the Stratification Method

The stratification method involves the following four steps:

1. The data are divided into strata based on the attributes expected to affect the quality of a building, such as the structure of the building, construction method, and venue. Hence,

Table 2. Descriptive statistics on the numeric data from building starts.

	Housing-W					Non-Housing-W				
	Min	Max	Avg	Med	Sd	Min	Max	Avg	Med	Sd
#19 (ten k. yen)	5	0.9 bn.	1,997	1,800	1,044.03	2	2.8 bn.	2,524	1,432.5	4,922.85
#18 (m ²)	11	2,993	124	111	56.49	11	14,001	174	104	246.96
#19 / #18	0.10	88.48	16.26	15.71	4.16	0.03	206.68	14.38	14.18	6.72
#15 (N. floor)	1	7	2	2	0.39	1	5	1	1	0.49
#16 (N. floor)	1	7	1	1	0.11	1	3	1	1	0.11
#6 (month)	1	98	4	4	1.24	1	72	3	3	1.98
#17 (m ²)	12	1.49 mm.	227	180	907.68	11	3.49 mm.	992	355	18,694
#29 (unit)	1	64	1	1	1.38	-	-	-	-	-

	Housing-RC					Non-Housing-RC				
	Min	Max	Avg	Med	Sd	Min	Max	Avg	Med	Sd
#19 (ten k. yen)	26	104 bn.	20,206	8,000	85,146	2	65.8 bn.	37,873	9,000	136,694
#18 (m ²)	11	0.20 mm.	1,016	403	3,214.03	11	0.25 mm.	1,672	408	5,951.73
#19 / #18	0.10	355.63	20.99	19.53	8.43	0.11	365.59	25.91	22.86	14.87
#15 (N. floor)	1	60	4	3	3.50	1	42	3	2	2.21
#16 (N. floor)	1	8	1	1	0.30	1	8	1	1	0.47
#6 (month)	1	93	8	7	3.82	1	99	8	7	5.29
#17 (m ²)	15	1.75 mm.	551	306	5,057.25	9	9.95 mm.	4,984	849	82,317
#29 (unit)	1	1,634	15	7	30.63	-	-	-	-	-

Table 2. (Continued)

	Housing-S					Non-Housing-S				
	Min	Max	Avg	Med	Sd	Min	Max	Avg	Med	Sd
#19 (ten k. yen)	5	80 bn.	3,784	3,000	13,072	1	300 bn.	11,605	2,500	112,314
#18 (m ²)	11	0.24 mm.	182	132	425.10	11	0.58 mm.	765	194	4,001.90
#19 / #18	0.12	240.04	22.10	21.37	5.81	0.001	459.39	14.10	11.90	10.37
#15 (N. floor)	1	54	2	2	0.62	1	54	2	1	1.21
#16 (N. floor)	1	4	1	1	0.18	1	8	1	1	0.52
#6 (month)	1	73	3	3	1.26	1	99	4	3	2.98
#17 (m ²)	12	39,338	273	210	241.82	4	9.24 mm.	2,441	626	36,612
#29 (unit)	1	1,493	2	1	4.36	-	-	-	-	-

Note 1: Min = minimum, max = maximum, ave = average, med = median, sd = standard deviation, k. = thousand, mm. = million, bn. = billion

Note 2: Variable # in each row corresponds to that of Table 1.

Note 3: #19 / #18 means the unit price per floor area of the construction expenses (construction contract expenses divided by total floor area).

Note 4: The descriptive statistics of #16 is calculated only for buildings that have basement floors.

Source: Compiled based on microdata of the MLIT's Building Starts statistics.

- the quality of the buildings within the same group is differentiated even though group members have similar characteristics.
2. The total amount of contract construction expenses, as well as the total floor area of a property, are calculated for the buildings included in the same group. Then, the unit price per floor area is obtained by dividing total expenses by the total floor area. This unit price represents the building price of the group.
 3. The value of the unit price is transformed so that the index considers the average in FY 2011 as 100 following the base year of the CCD. In this article, we denote Japanese Fiscal Year (from April to next March) as FY and Calendar Year as CY unless otherwise noted. The index is constructed for every stratum. In this article, we denote Japanese Fiscal Year (from April to next March) as FY and Calendar Year as CY unless otherwise noted.
 4. Finally, the indexed unit prices are aggregated by taking a weighted average based on the FY 2011 value of the total contract construction costs of each group. That is, the price index calculated in this method is a fixed-standard Laspeyres index. As described later, the output-type indices estimated in this study are compared with the input-cost indices, the MLIT's CCD. CCD undergoes major revisions about every five years, so indices in the figures are constructed based on the link coefficients for FY2005–FY2011 and FY2011–FY2015. All the indices are rebased so that FY 2011 = 100 just for comparison. We estimated geometric mean-based indices as well as transaction numbers-based indices in the following sections.

The index obtained after the above four steps, becomes the final output building price index. In this study, we classify buildings into two categories: housing and non-housing buildings, which greatly differ in quality, and adopt quarterly aggregates. We do this because the available attribute items are different for housing and non-housing properties, and the percentage of empty strata can be greatly improved by using quarterly aggregation instead of monthly aggregation. In addition, for empty strata, we examine the following five typically used imputation methods: (1) the last observed price data, (2) the price data of the same period in the previous year, (3) the average price of non-empty strata, (4) the average quarter-on-quarter growth rate of non-empty strata, and (5) the average year-on-year growth rate of non-empty strata. The results are omitted here for the sake of brevity. For details, please contact the authors. Regarding “private rents” in the CPI, the index is created by supplementing the rents for the most recent month horizontally for rented houses that became vacant due to renters moving out, thus becoming missing values. We replaced empty strata with the last observed value because this approach results in the smallest fluctuations in the estimated price index. (If you want to know the details, please contact the authors). This is the same method adopted in “private rents” in the Japanese CPI, the index is created by supplementing the rents for the most recent month horizontally for rented houses that became vacant due to renters moving out, thus becoming missing values.

For stratification, items that greatly affect the quality of the building and have an impact on the unit price of the building should be selected. Specifically, for housing buildings, the “construction method” (prefabricated and two-by-four, among others), “construction method” (single-family house and apartment house, among others), “structure” (wooden and reinforced concrete, among others), “prefecture (region),” and “use relationship” (owner-occupied and rental house, among others) are selected. In the case of non-housing

buildings, “building use” (e.g., industry of the company that built the building), “structure,” and “prefecture” (region) are used (Table 3).

3.3. Overview of the Hedonic Method

The dependent variable of the hedonic function is the unit price per floor area of the estimated construction expenses (construction contract expenses/total floor area). In this study, we use the logarithmically transformed value, in line with Diewert (2003). We adopt a rolling estimation method, in which the estimation is performed with the window length of 12 months. This is a widely accepted method for estimating housing price indices by the Eurostat et al. (2013), Hill et al. (2022) and Shimizu et al. (2010), for example. The method has the advantage of reflecting time-series changes in parameters and has been adopted in the official statistics in many countries including Japan. The rolling window hedonic method is used in the “Official Property Price Index” published by Statistics Office in many countries. All available attribute items are used as explanatory variables, and a one-sided log-linear hedonic function is employed, as follows: The estimation period ranges from January 2005 to December 2020, 181 times the rolling estimation in a 12-month window frame; all of the variables are found to be statistically significant at the 10% level of significance throughout the whole period.

$$\log p_i = \alpha + \sum_{j=1}^n \beta_j x_{i,j} + \sum_{k=2}^m \delta_k d_{i,k} + \sum_{t=2}^{\tau} \gamma_t TD_{i,t} + u_i, \quad (1)$$

Table 3. Attribute items in the stratification method.

	Attribute (specification) item	Contents
Housing	Building method	Prefabricated, two-by-four, other (conventional construction method)
	Building (e.g., house)	Single-family houses, row houses, and apartment buildings
	Structure	Wooden construction, steel-framed reinforced concrete construction, reinforced concrete construction, steel construction, and concrete block construction, among others
	Administrative divisions of Japan	47 Prefectures
	Usage restrictions, limitations	Owner-occupied houses, rental houses, salaried housing, and condominiums
Non-housing	Versatile or not	Multi-purpose buildings, among others
	Purpose	Standard Industrial Classification (Middle Classification)
	Structure	Wooden construction, steel-framed reinforced concrete construction, reinforced concrete construction, steel construction, and concrete block construction, among others
	Administrative divisions of Japan	47 Prefectures

p_i : unit price per square meter of building i ;
 α : constant term;
 $x_{i,j}$: j th attribute of building i (numerical value);
 β_j : parameter of numerical data;
 $d_{i,k}$: k th attribute of building i (dummy);
 δ_k : parameter of the dummy variable;
 $TD_{i,t}$: survey month t of building i (time dummy);
 γ_t : time dummy parameters (representing quality-adjusted prices);
 u_i : error term.

The estimation period ranges from January 2005 to December 2020, 181 times the rolling estimation in a 12-month window frame; all of the variables are found to be statistically significant at the 10% level of significance throughout the whole period, to say the estimation results first.

It is difficult to assume a linear relationship between the qualities of buildings due to the wide distribution of the number of stories. Therefore, in addition to setting multiple dummy variables in a nonparametric manner, a piecewise linear function is also introduced like did in [Diewert and Shimizu \(2015, 2016, 2017\)](#). This function divides the number of floors into multiple categories and assumes linear relationships within each category. The use of piecewise linear functions allows us to consider cases in which the impact of increasing the number of floors from a one-story building to a two-story building is different from the impact of increasing the number of floors from a 20-story building to a 21-story building. For example, in the case in which the number of floors is divided into three categories, we obtain:

$$\begin{aligned}
 f_S(S_i) \equiv & D_{S,i1} \lambda_1 S_i + D_{S,i2} [\lambda_1 S_1 + \lambda_2 (S_i - S_1)] + D_{S,i3} [\lambda_1 S_1 + \lambda_2 (S_2 - S_1) \\
 & + \lambda_3 (S_i - S_2)], \quad (2)
 \end{aligned}$$

where S_i is the building i 's number of stories above ground (S_1, S_2 are the maximum number of stories in Category 1 and Category 2, respectively), $D_{S,i}$ is a dummy variable that takes a value of 1 if the number of stories above ground falls into each category, and λ is the coefficient through which each story category affects the unit price per total floor area.

In cases in which a dummy variable is not continuously observed throughout the rolling window, an adjustment is made to redefine it into a "wider range of dummy variables" integrated with other items. This adjustment is made since the inclusion or exclusion of these explanatory variables (depending on the point in time of the estimation) may lead to significant changes in the estimated values of the coefficients. For example, the dummy variable representing Aomori Prefecture is redefined into a regional dummy for the Tohoku region, and the usage dummies are redefined by industry instead of dividing them by use, such as warehouses and stores. Although indices are estimated based on the monthly basis, we use quarterly series by taking simple average to them. For further details of the Hedonic estimation, please see Online Supplemental Data, Appendix 2.

4. Estimation Results of the Output-Type Building Price Index

4.1. Output-type Building Price Index and CCD

This section compares the output-type building price indices estimated by the stratification and hedonic methods with the CCD, an input-cost index. We evaluate the estimation results focusing on six categories, namely wood-frame, reinforced concrete (RC), and steel-frame (S) construction for housing and non-housing buildings, respectively, which account for a large share of the total building stock.

The estimation period ranges from the first quarter of 2005 to the fourth quarter of 2020. It includes an expansionary phase, approximately until 2007, a sharp recessionary phase triggered by the Lehman shock in 2008–2009, and a long expansionary phase from the fall of 2012 to the fall of 2018. From autumn 2012 onward, construction investment has raised substantially, and the construction sector's profits have increased markedly due to an improvement in the order environment. In addition to these factors, the recession caused by the spread of COVID-19 and the subsequent rise in housing demand have caused lumber prices to start rising from approximately July 2020, leading to a "wood shock." The Bank of Japan's CGPI confirms that the impact of rising import prices for materials has become more significant since the beginning of 2021, but the impact has not been as pronounced during the analysis period (until the end of 2020). In terms of prices, the sharp rise in resource prices up to approximately 2008 has increased the prices of construction materials, followed by a decline in material prices in 2009, and a gradual rise in prices since 2013. In addition, the recession caused by the spread of COVID-19 and the subsequent rise in housing demand have caused lumber prices to start rising from approximately July 2020, leading to a "wood shock." The Bank of Japan's CGPI confirms that the impact of rising import prices for materials has become more significant since the beginning of 2021, but the impact has not been as pronounced during the analysis period (until the end of 2020). These phenomena reflect substantial price changes, such as increases in labor costs due to the persistent labor shortage in the construction sector and rises in the profits of construction companies. Hence, we evaluate changes in the construction price index that reflect such variations in the economic conditions. The estimated indices are compared with the CCD (Figure 1).

In comparing those indices, we need to be careful about possible influences caused by differences in the methodologies. For the CCD, since it is impossible for us to reconstruct the indicator, we referred to the Paasche-check results published by the data source. CCD is a Laspeyres index based on the weights derived from the Input-Output Table and it undergoes major changes in the methodological standards almost every five years. In the sample period of this paper, it has three different standard years, that is, FY2005, FY2011 and FY2015. We estimated the Fisher indices from the Paasche-check results conducted at the time of standard revisions and checked their deviations from the Laspeyres indices. The estimated biases are about 0.2 point to 0.5 point, so differences in calculation methods can be negligible when evaluating the movements of CCD.

Considering the discussion in De Haan (2004), the hedonic price indices (the quality-adjusted Jevons index) are interpreted as the stratification method based on geometric mean stratum. We conducted the Stratification estimation based on the geometric mean with three

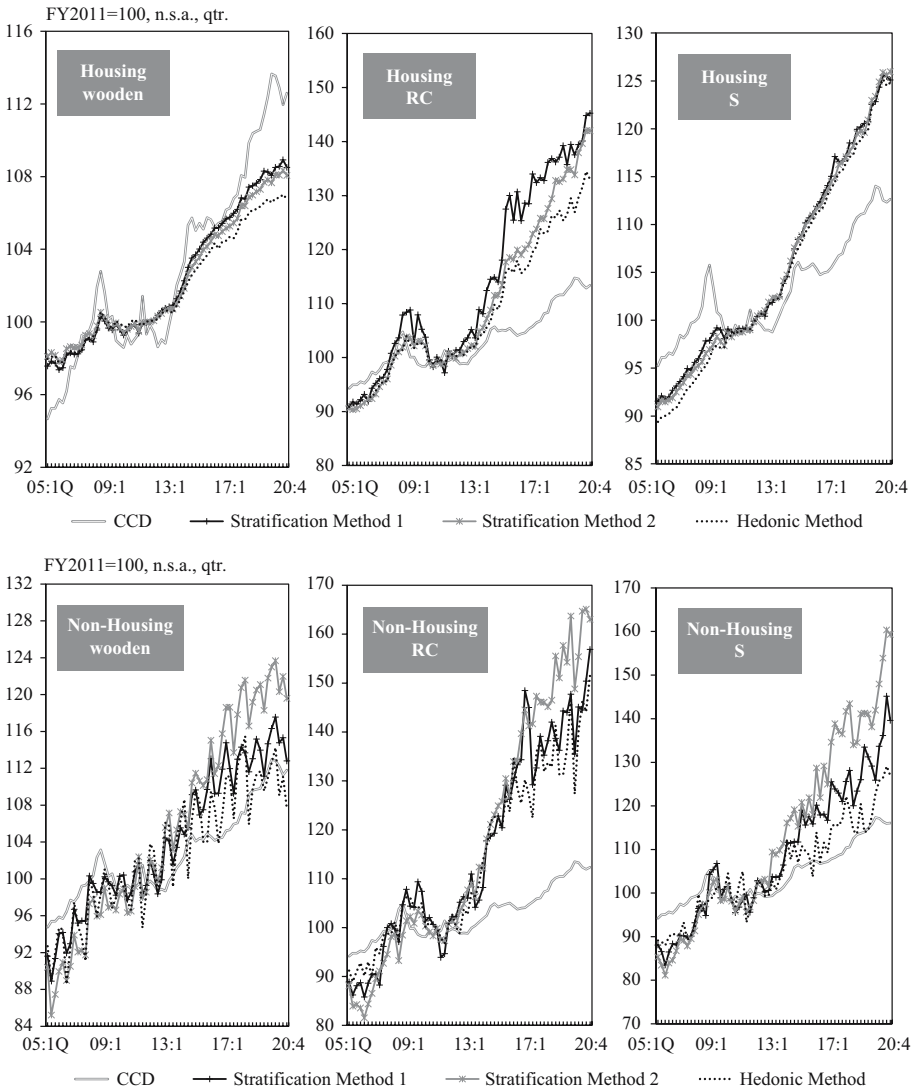


Fig. 1. Output-type building price index.

Note 1: Quarterly indices for Hedonic are obtained by simple average of monthly indices.

Note 2: Stratification Method 1 is based on the transaction value weights, and Method 2 is on transaction numbers weights.

Source: MLIT, CCD; calculations by the authors.

base years of FY2005, FY2011, and FY2015. These indices are then connected by using link coefficients like we did in the CCD. Moreover, we considered two different types of weights to aggregate each stratum; the transaction value weights (Stratification Method 1) and transaction numbers weights (Stratification Method 2). The former is compatible with the CCD and the latter is to make comparison with the Hedonic estimation. Eurostat et al. (2013) recommend the following for the choice of weights, in the case of the house price index: “A price index which is required to measure the wealth associated with the ownership of residential property should be stock-weighted. A price index which is required for

measuring the real output of the residential real estate industry should be sales-weighted". [Diewert \(2003\)](#) mentions weights for price indexes using the hedonic method.

The output-type construction price indices obtained by the stratification and hedonic methods show a larger increase than the input-cost CCD for all series except for housing and wood construction (housing and RC construction, housing and S construction, non-housing and wood construction, non-housing and RC construction, and non-housing and S construction). In many cases, the gap between the CCD and the output-type construction price index was almost negligible until 2012, but from 2013 onward, the gap has gradually increased. In the economic expansion phase, since 2013, the true construction prices captured by the output-type construction price index are higher than those captured by the existing input-cost price indices (the CCD and SNA's index). The results suggest that the estimated output-type construction price index may effectively overcome the measurement difficulties outlined in this study.

In the late 2000s, the timing of the increase in the indices based on both the stratification and hedonic methods, especially for housing and RC construction, lags behind the increase in the CCD, and the fluctuation is modest. This result indicates that construction companies cannot immediately pass on changes in materials and labor costs reflected in the CCD to building owners in a competitive environment; hence, they negotiate prices over time and reflect them in their contract prices.

Finally, comparing the indices obtained by the stratification method with those generated by the hedonic method, we find that in some cases, they show approximately the same increase (housing/S and non-housing/RC); however, a certain gap is observed between the price index obtained using the stratification method (non-housing/wooden) and that obtained by the hedonic method. The former increases more than the latter, and the gap between the two is, at times, large (housing/RC, non-housing/wooden, and non-housing/S) ([Table 4](#)).

As described above, among the six series of output-type indices, the gap between the two indices is large for three series: housing/RC construction, non-housing/wood construction, and non-housing/S construction. This discrepancy may be due to inappropriate selection of attributes for stratification and subdivision in the stratification

Table 4. Discrepancies between the stratification method, hedonic method, and CCD – housing/non-housing and structure.

Housing/ Non-housing	Structure	Stratification versus CCD	Hedonic versus CCD	Stratification versus Hedonic
Housing	wooden reinforced concrete construction	medium gap large gap	medium gap large gap	medium gap large gap
Non-housing	steel construction	medium gap	medium gap	small gap
	wooden reinforced concrete construction	large gap large gap	medium gap large gap	large gap small gap
	steel construction	large gap	medium gap	large gap

method and insufficient quality adjustment, among others. In addition, the hedonic method may suffer from misspecification bias, such as omitted variable bias since RC high rise condominium and non-housing buildings are likely to be insufficient quality adjustments due to a lack of attributes because of their heterogeneity. As for misspecification bias, please see Ekeland et al. (2004) and Heckman et al. (2010) for further information. In the next subsection, we further discuss this point.

4.2. Evaluation of Output-Type Building Price Index: Increasing Variability

Among the three series in which substantial gaps are observed, for housing/RC and non-housing/S constructions, the unit price per total floor area of the construction contract expenses has been increasing in recent years (Figure 2). Although only the graph for housing/RC is shown, the same trend is confirmed for non-housing/S.

This result indicates that the variation in the quality of buildings has been large in housing/RC and non-housing/S. This phenomenon may be due to an increase in the number of high-rise condominiums such as tower apartments and the larger variation in buildings constructed with steel frames owing to the recent improvement in construction technology.

To adjust for the effects of these increased variations, in the stratification method, the elements detailed in Table 5 are added as items for subdivision and stratification, and the stratification items are further subdivided to improve the degree of homogeneity of attributes.

In addition, in the hedonic method, we divide the data by setting a total floor area as a threshold to estimate the function for each data point. Referring to the distribution of the number of observations, we estimate the cases divided into three categories for housing

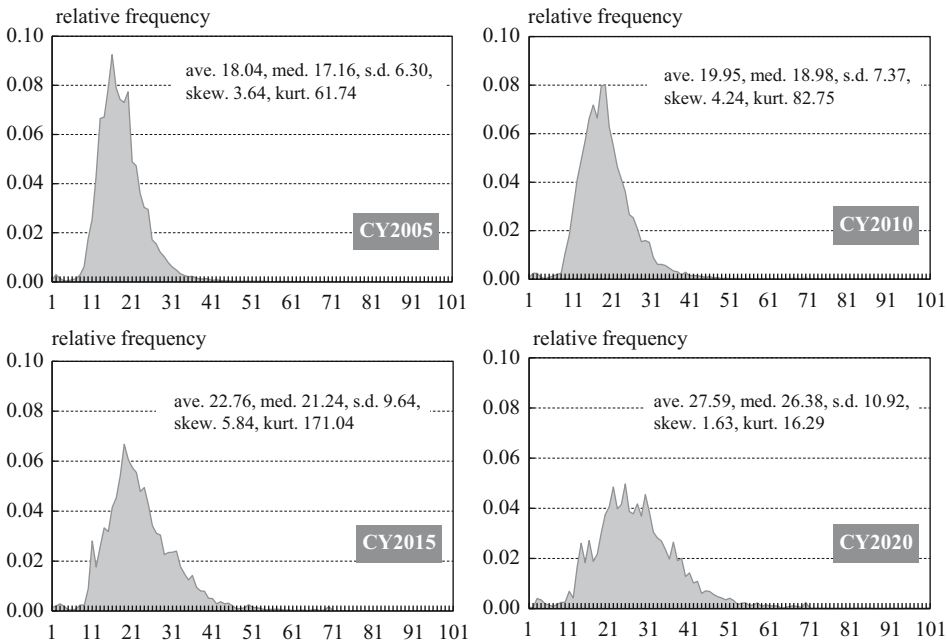


Fig. 2. Distribution of unit price per square meter in housing RC construction.

Note: The unit of x-axis is 10,000 yen except the last bin which shows over one million.

Source: MLIT and calculations by the authors.

Table 5. Additional items in the stratification method.

Housing/non-housing	Baseline	Addition of stratification and subdivision items
Housing	Building method, construction method, structure, prefecture, and use relationship	Baseline; Builder, capitalization category, city planning category, basement, retired housing, funds for new housing, use (building with housing industrial use)
Non-housing	Multi-use or not, use, structure, and prefecture	Baseline; Architect, capitalization category, city planning category, underground

and RC structures: total floor area of 200 m² or less, over 200 m² to 800 m², and over 800 m²; for non-housing and S structures: total floor area of 100 m² or less, over 100 m² to 300 m², and over 300 m². By dividing the sample by the size of the buildings, differences in the coefficients of the explanatory variables in the hedonic estimation may be considered, preventing small buildings, which have a small share in the value of the construction contract expenses but a high share in the sample size, from having excessive influence.

Figure 3 compares the indices obtained by the stratification method with the addition of subdivision and stratification items and the indices obtained by the hedonic method with sample division by total floor area. Comparing Figure 3 with Figure 1, we show that the difference between stratification and hedonic indices is much smaller than that obtained before dividing the sample for both housing/RC and non-housing/S.

The degree of homogeneity in the stratification method is improved by increasing the number of items to be subdivided and stratified, enhancing the accuracy of the index. In addition, by dividing the sample addressed by the hedonic method, the price trends of

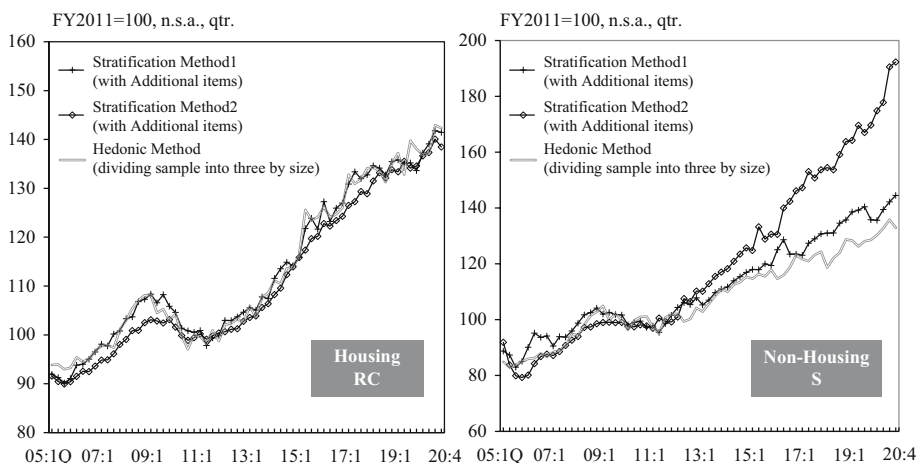


Fig. 3. Stratification, hedonic method: Additional estimation index 1.

Note 1: Quarterly indices for Hedonic Method are obtained by simple average of monthly indices.

Note 2: Stratification Method 1 is based on the transaction value weights, and Method 2 is on transaction numbers weights.

Source: MLIT and calculations by the authors.

buildings with a high rate of price increase and a large scale, such as tower condominiums, are reflected by a more appropriate weight, narrowing the gap between the two methods. However, it should be noted that the stratification method undergoes large fluctuations depending on the choice of aggregation weights. It implies immanent instabilities in the stratification method, and the Hedonic method is superior in terms of quality adjustment.

4.3. Effect of Seasonality on Non-Housing and Wooden Structures

Among the three series in which index gaps are observed, the non-housing and wood-frame series show large quarterly fluctuations in both indices, suggesting that seasonality is likely observed (Figure 1). By addressing the number of observations for each month for warehouses in agriculture, forestry and fisheries industry (corresponds to 314 classification number of types of dwellings, industries, and use) and stores in food services industry (classification number: 532), which have a large share in the non-housing/wooden category (Figure 4, right), we observe seasonal fluctuations with a peak in June. The seasonal changes in the composition of non-housing/wooden buildings may affect the estimated indices because the unit price per square meter for seasonally built constructions such as beachside houses and small cottages used for storing harvests is significantly lower than for other buildings.

Based on these characteristics, in the stratification method, the index is created by adding stratification items in the same way as in the two series described in the previous subsection. In the hedonic function, “use of buildings” dummy variables, such as “office,” “store” and so on are added in addition to the industry of the building owner company, and the interaction term between the industrial/usage dummy and the time dummy is employed to capture seasonal changes in the composition of buildings within each industry and use category (Figure 4, left). The results show that the gap between the index obtained by the stratification method with the addition of stratification items and the hedonic index with

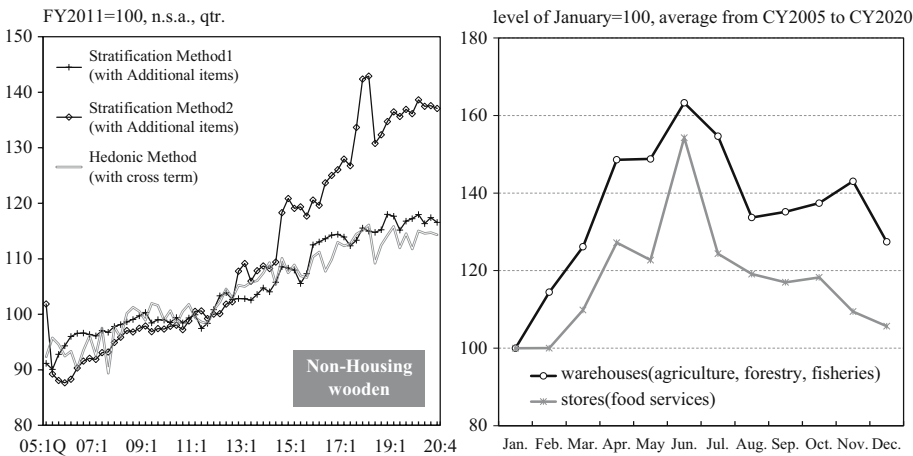


Fig. 4. Left: stratification, hedonic method; Additional estimation index 2. Right: number of observations for use/application categories applicable to non-housing/wooden.

Note 1: Quarterly indices for Hedonic Method are obtained by simple average of monthly indices.

Note 2: Stratification Method 1 is based on the transaction value weights, and Method 2 is on transaction numbers weights.

Source: MLIT and calculations by the authors.

the addition of the cross term is reduced. In addition to the increase in the degree of quality adjustment in the stratification method, the effect of seasonal variation is removed by adding the cross term in the hedonic method; thus, the fluctuation in the index is smaller than that shown in [Figure 1](#). The index is thus considered more appropriate.

5. Conclusion

This study builds an output-type building price index using large-scale administrative data from the Statistics of Building Starts, using the stratification and hedonic methods.

By considering changes in the six series of indices for housing and non-housing buildings categorized by the structure type, we obtain an output-type building price index with a reasonable level of accuracy. Except for housing/wooden, during the sample period, all series (housing/RC, housing/S, non-housing/wooden, non-housing/RC, and non-housing/S) have increased at a higher rate than the CCD, the input-cost price index, and currently official statistics. This trend has been particularly eminent recently, especially since 2013. The output-type building price indices obtained by the stratification and hedonic methods reflect that profits in the construction sector have been increasing due to the recent improvements in the environment for construction orders. As such, the output-type building price index provides valuable information.

We confirm that the current input-cost price indices may not be sufficient to improve the accuracy of real construction investment in Japan and accurately reflect the activities of the construction sector. A new, highly accurate output-type construction price index may better serve this scope.

The hedonic method allows a higher degree of homogeneity or quality adjustment. Although the stratification method requires relatively small compiling costs, the estimated indices have shown some fluctuations especially in non-housing structures where the number of samples is limited, and those buildings are supposed to have larger heterogeneity than housing buildings. Chain linked stratification indices have shown chain drift while the hedonic method indices have been relatively stable for the sample periods used in this research, which supports the findings of [Ivancic et al. \(2011\)](#).

This study's results indicate that the indices obtained by the stratification and hedonic methods for the three series of housing/wooden, housing/S, and non-housing/RC show approximately the same level of increase. However, for the three series of housing/RC construction, non-housing/wooden construction, and non-housing/S construction, a certain discrepancy is observed between the two indices. These gaps may be reduced by refining the estimation. The stratification method may be used to create an output-type building price index at a low cost. However, as seen in the estimated indices, it is also true that the degree of quality adjustment in the stratification method might be limited, and we must carefully check its specifications when adopting this method.

Despite its methodological contribution, the current study has some limitations. First, the proposed price index does not reflect improvements in the quality of labor, which are expected to have a significant impact in the medium to long term. The attribute items of the Statistics of Building Starts used do not reflect improvements in building quality associated with increased quality of labor due to data limitations. Second, the construction contract expenses of a building provided by the Statistics of Building Starts are estimated at the time when

construction began and not at the time of completion; hence, they do not reflect design changes after the construction start. The method for converting the price index into an accrual-based index to match the SNA standards for estimating construction investment is also worth further investigation. While our research has shown benefits of using the administrative data, most of them are not generally available in the electronic format in Japan. Construction documents, which are abundant in a building's attribute information, are to be submitted to the local authorities as papers while the Statistics of Building Starts are available in the digital form since they are surveyed in the different questionnaire which is usually to be filled in by extracting only small fraction of the Construction documents. In addition to further verification of the estimation method, the above issues should be examined in the future.

6. References

- Benedetti, I., L. Biggeri, and T. Laureti. 2022. "Sub-National Spatial Price Indexes for Housing: Methodological Issues and Computation for Italy." *Journal of Official Statistics* 38(1): 57–82. DOI: <http://dx.doi.org/10.2478/JOS-2022-0004>.
- Cabinet Office (Government of Japan). 2022. *Explanatory Notes on the Methodology for Estimating National Accounts (Annual Estimates), 2015 Standard Version*. Available at: https://www.esri.cao.go.jp/jp/sna/data/reference1/h27benchmark/pdf/kaisetsu_20220726.pdf (accessed December 2022).
- De Haan, J. 2004. "Direct and indirect time dummy approaches to hedonic price measurement." *Journal of Economic and Social Measurement* 29: 427–443. DOI: <http://doi.org/10.3233/JEM-2004-0235>.
- Diewert, W.E. 2003. "Hedonic Regressions: A Review of Some Unresolved Issues," In Proceedings of the Seventh Meeting of the International Working Group on Price Indices: the United Nations Statistical Commission, May 27–29, 2003. Available at: https://econ.sites.olt.ubc.ca/files/2013/06/pdf_paper_erwin-diewert-hedonic-regressions.pdf (accessed December 2022).
- Diewert, W.E., and C. Shimizu. 2015. "Residential Property Price Indices for Tokyo." *Macroeconomic Dynamics* 19: 1659–1714. DOI: <https://doi.org/10.1017/S1365100514000042>.
- Diewert, W.E., and C. Shimizu. 2016. "Hedonic Regression Models for Tokyo Condominium Sales." *Regional Science and Urban Economics* 60: 300–315. DOI: <https://doi.org/10.1016/j.regsciurbeco.2016.08.002>.
- Diewert, W.E., and C. Shimizu. 2017. "Alternative Approaches to Commercial Property Price Indexes for Tokyo." *Review of Income and Wealth* 63: 492–519. DOI: <https://doi.org/10.1111/roiw.12229>.
- Diewert, W.E., and C. Shimizu. 2022. "Residential Property Price Indexes: Spatial Coordinates versus Neighborhood Dummy Variables." *Review of Income and Wealth* 68: 770–796. DOI: <https://doi.org/10.1111/roiw.12534>.
- Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Scanner data, time aggregation and the construction of price indexes." *Journal of Econometrics* 161: 24–35. DOI: <https://doi.org/10.1016/j.jeconom.2010.09.003>.
- Eurostat, OECD, and World Bank. 2013. *Handbook on Residential Property Price indices (RPPIs)*. Luxembourg: Publications Office of the European Union. Available at:

- <https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF> (accessed December 2022).
- Eurostat, IMF, OECD, UN, and World Bank. 2009. *System of National Accounts 2008*. Washington, D.C. World Bank Group. Available at: <http://documents.worldbank.org/curated/en/417501468164641001/System-of-national-accounts-2008>. (accessed December 2022).
- Ekeland, I., J.J. Heckman, and L. Nesheim. 2004. "Identification and Estimation of Hedonic Models." *Journal of Political Economy* 112: 60–109. DOI: <https://doi.org/10.1086/379947>.
- Fukao, K., T. Kameda, K. Nakamura, R. Namba, and M. Sato. 2017. "Measurement of Deflators and Real Value Added in the Service Sector." *Economic Analysis* 194: 9–44. Available at: <https://www.esri.cao.go.jp/jp/esri/archive/bun/bun194/bun194.html>.
- Heckman, J.J., R.L. Matzkin, and L. Nesheim. 2010. "Nonparametric Identification and Estimation of Nonadditive Hedonic Models." *Econometrica* 78: 1569–1591. DOI: <https://doi.org/10.3982/ECTA6388>.
- Hill, R.J. 2013. "Hedonic price indexes for residential housing: A survey, evaluation and taxonomy." *Journal of economic surveys* 27: 879–914. DOI: <https://doi.org/10.1111/j.1467-6419.2012.00731.x>.
- Hill, R.J., M. Scholz, C. Shimizu, and M. Steurer. 2018. "An Evaluation of the Methods Used by European Countries to Compute their Official House Price Indices." *Economie et Statistique* 500–501–502: 221–238. DOI: <https://doi.org/10.24187/ecostat.2018.500t.1953>.
- Hill, R.J., M. Scholz, C. Shimizu, and M. Steurer. 2022. "Rolling-Time-Dummy House Price Indexes: Window Length, Linking and Options for Dealing with Low Transaction Volume." *Journal of Official Statistics* 38: 127–151. DOI: <https://doi.org/10.2478/jos-2022-0007>.
- JILPT, Japan Institute for Labour Policy and Training 2020. *Youthful Labour Statistics 2020*. Available at: <https://www.jil.go.jp/kokunai/statistics/kako/2020/documents/useful2020.pdf> (accessed May 2023).
- MLIT, Ministry of Land, Infrastructure, Transport, and Tourism. 2021. *Outline and Revision of the Construction Cost Deflator (FY2015 Base)*. Construction Economics and Statistics Office, Policy Bureau, Ministry of Land, Infrastructure, Transport and Tourism. Available at: <https://www.mlit.go.jp/sogoseisaku/jouhouka/content/001411495.pdf> (accessed December 2022).
- MLIT, Ministry of Land, Infrastructure, Transport, and Tourism. n.d.a. *Construction Statistics Guidebook*. Construction Research and Statistics Office, Policy Bureau, Ministry of Land, Infrastructure, Transport. Available at: https://www.mlit.go.jp/tou-keijouhou/chojou/csg/csg_f.htm (accessed December 2022).
- MLIT, Ministry of Land, Infrastructure, Transport, and Tourism. n.d.b. *Construction Economics Statistics Survey*. Construction Economics and Statistics Office, Policy Bureau, Ministry of Land, Infrastructure, Transport and Tourism. Available at: https://www.mlit.go.jp/statistics/sosei_jouhouka_fr1_000004.html. (In Japanese, accessed December 2022).
- OECD. 1997. *Construction Price Indices: Sources and Methods*. Paris: Statistics Directorate, Organisation for Economic Co-operation and Development. Luxembourg:

- Statistical Office of the European Community. Available at: <https://www.oecd.org/industry/business-stats/2372435.pdf> (accessed December 2022).
- Office for National Statistics. 2018. *Construction Output Price Indices (OPIs) QMI*. Available at: <https://www.ons.gov.uk/businessindustryandtrade/constructionindustry/methodologies/constructionoutputpriceindicesopisqmi> (accessed December 2022).
- Shimizu, C., H. Takatsuji, H. Ono, and K.G. Nishimura. 2010. “Structural and Temporal Changes in the Housing Market and Hedonic Housing Price Indices: A case of the Previously Owned Condominium Market in the Tokyo Metropolitan Area.” *International Journal of Housing Markets and Analysis* 3: 351–368. DOI: <https://doi.org/10.1108/17538271011080655>.
- Silver, M. 2011. “House Price Indices: Does Measurement Matter?” *World Economics* 12: 69–86. Available at: <https://econpapers.repec.org/article/wejwldecn/482.htm>.

Received October 2021

Revised June 2022

Accepted December 2022

From Quarterly to Monthly Turnover Figures Using Nowcasting Methods

Daan Zult¹, Sabine Krieg², Bernd Schouten¹, Pim Ouweland¹, and Jan van den Brakel²

Short-term business statistics at Statistics Netherlands are largely based on Value Added Tax (VAT) administrations. Companies may decide to file their tax return on a monthly, quarterly, or annual basis. Most companies file their tax return quarterly. So far, these VAT based short-term business statistics are published with a quarterly frequency as well. In this article we compare different methods to compile monthly figures, even though a major part of these data is observed quarterly. The methods considered to produce a monthly indicator must address two issues. The first issue is to combine a high- and low-frequency series into a single high-frequency series, while both series measure the same phenomenon of the target population. The appropriate method that is designed for this purpose is usually referred to as “benchmarking”. The second issue is a missing data problem, because the first and second month of a quarter are published before the corresponding quarterly data is available. A “nowcast” method can be used to estimate these months. The literature on mixed frequency models provides solutions for both problems, sometimes by dealing with them simultaneously. In this article we combine different benchmarking and nowcasting models and evaluate combinations. Our evaluation distinguishes between relatively stable periods and periods during and after a crisis because different approaches might be optimal under these two conditions. We find that during stable periods the so-called Bridge models perform slightly better than the alternatives considered. Until about fifteen months after a crisis, the models that rely heavier on historic patterns such as the Bridge, MIDAS and structural time series models are outperformed by more straightforward (S)ARIMA approaches.

Key words: Benchmarking; nowcasting, register statistics; mixed frequency models.

1. Introduction

The purpose of national statistical institutes (NSIs) is to publish relevant, accurate and timely official statistics. However, the production of high-frequency timely statistics generally compromises the accuracy of these figures. This trade-off is even increased if a major part of the data is observed on a frequency that is lower than the required output. For instance, in The Netherlands, short-term business statistics rely for the most part on turnover obtained from value added tax (VAT) administrations, where most companies declare turnover either quarterly or monthly. The current approach is to wait for the quarter to be finished and produce a sector specific quarterly statistic, based on turnover from both monthly and quarterly declarants that is available at the publication date. The

¹ Statistics Netherlands – Methodology PO Box 4000, Den Haag 2270 JM Netherlands Emails: db.zult@cbs.nl, bp.schouten@cbs.nl and p.ouweland@cbs.nl

² Statistics Netherlands – Department of Statistical Methods PO Box 4481, Heerlen 6401 CZ Netherlands. Emails: s.krieg@cbs.nl and ja.vandenbrakel@cbs.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors are grateful to the Associate Editor and the anonymous referees for careful reading a former draft of the manuscript and providing useful comments.

level of publication is referred to as primary publication cell (PPC) and can be quite detailed (e.g., “Restaurants” or “Publishers”). The question is whether the same data can also be used to produce an earlier, more frequent, and sufficiently accurate monthly estimate on this detailed level. This question has two distinctive elements.

The first element is an increase in frequency, which can be achieved with a method referred to as benchmarking (BM). BM models are extensively discussed in literature, such as in the “ESS Guidelines on temporal disaggregation, benchmarking, and reconciliation” (Eurostat 2018). In case of sufficiently long time series, which is what we assume in this article, these guidelines recommend BM models that are based on *movement preservation* as in Denton (1971), Dagum and Cholette (2006) and Chow and Lin (1971). We will briefly discuss these BM models in Subsection 2.2.

The second element is an increase in timeliness, which can be achieved with a method referred to as nowcasting (NC). The most straightforward NC approach would be to extrapolate the monthly series obtained with BM by change in turnover of monthly declarants over the previous and current month. This would closely follow the current production process of the quarterly series and is therefore attractive. However, because the monthly declarants may constitute a selective and/or small sample of the population, this nowcast probably needs to be adjusted as soon as the quarterly data is available. To minimize this adjustment, it is worthwhile to investigate some more sophisticated NC models.

NC models are extensively discussed in literature, such as in the “Handbook of rapid estimates” (Eurostat 2017). This includes mixed frequency models that combine the BM and NC problem in one model. We discuss NC models in more detail in Subsection 2.3. Simultaneous (multivariate) estimation of series for different PPCs might improve the accuracy of the estimates. We leave this for further research.

Another sample selection and sample size bias correction approach would be to weight the monthly declarants to the entire population with the help of background characteristics and trends. This method is called pseudo design-based estimation (Baker et al. 2010). In our case we deem this approach less fruitful, because the background characteristics that are available are limited (i.e., number of employees). However, maybe by using historic turnovers on the company level in an imputation model, it might be possible to improve our auxiliary monthly series and hereby our nowcasting results. This is outside the scope of this article but might be worthwhile further research.

It is important to note that the combination of BM and NC introduces an evaluation problem. Normally NC models can be evaluated by simply waiting for the future to unfold. In our case, this refers to the monthly series that is obtained with a BM model. This comparison is less trivial because the true monthly values remain unknown for the respondents that report on a quarterly basis. Which (type of) NC model is best suited for this problem is not trivial and is an important topic of this article. Furthermore, this question requires elaboration on evaluation criteria, which are therefore discussed in Subsection 2.4.

A final element that deserves special attention, especially in the light of the recent COVID-19 pandemic and resulting lockdowns, is the impact of extreme (economic) developments. Therefore, in Section 3, we apply and evaluate the BM and NC models discussed in Section 2 both in stable and extreme conditions. Both conditions are present in several economic sectors in the Netherlands over the period January 2010 to June 2020, which can be characterized as a long, economically stable period that ends with the

COVID-19 pandemic. To further investigate the long term impact of a crisis on the different models, we also simulate different types of crises and evaluate how the accuracy of the nowcasts is affected during and after a crisis. In Section 4 we conclude. The online Supplemental Material file contains some additional technical details about the models discussed in this article.

2. Notation Benchmarking Models and Nowcasting Models

This section introduces some notation and assumptions (Subsection 2.1) and discusses BM (Subsection 2.2) and NC models (Subsection 2.3). In the main text we discuss the models primarily on the conceptual level with a modest level of technical detail because they are already well described in the literature. We first introduce some notation that allows us to describe the problem and the BM and NC models.

2.1. Notation

Let y_t^Q be an observed quarterly time series with $t = 1, 2, \dots, T_q$ where $t = 1$ is the first month of the index series (e.g., January 2010) and T_q is the third month of the last available quarter. Furthermore, $Q(t)$ is a function that transforms t to the first month of its corresponding quarter. This implies, for example, that when $t = 11$, then $Q(t) = 10$, or when $t = 25$, then $Q(t) = 25$. It is understood that y_t^Q is written as a monthly series, where the months within a quarter are equal, i.e., $y_{Q(t)}^Q = y_{Q(t)+1}^Q = y_{Q(t)+2}^Q$. The three equal elements $y_{Q(t)}^Q, y_{Q(t)+1}^Q$ and $y_{Q(t)+2}^Q$ are all observed at time $Q(t) + 3$ (i.e., the first month of the next quarter). In the intermediate months, y_t^Q does not change. Furthermore, there is a monthly auxiliary time series x_t with $t = 1, 2, \dots, T_m$ where x_t is observed in month t and T_m is the last available month (i.e., $T_q = Q(T_m) - 1$). In other words, the monthly series x_t always extends 1, 2 or 3 months beyond the quarterly series y_t^Q . Next, we define an unobserved monthly time series y_t^M with $t = 1, 2, \dots, T_{\max}$, ($T_{\max} \geq T_m$) which is simply the quarterly series y_t^Q , disaggregated to a monthly series that goes into a yet unobserved future (until T_{\max}). The series y_t^Q and y_t^M are related within each quarter by either their mean (e.g., in case of an index) or their sum (e.g., in case of turnover). Therefore, we can write either $y_{Q(t)}^Q = \frac{y_{Q(t)}^M + y_{Q(t)+1}^M + y_{Q(t)+2}^M}{3}$ or $y_{Q(t)}^Q = y_{Q(t)}^M + y_{Q(t)+1}^M + y_{Q(t)+2}^M$ with $t = 1, 2, \dots, T_q$. Ideally, we would observe y_t^M , but instead we only observe the quarterly aggregate y_t^Q and the related variable x_t . Therefore, we also define \hat{y}_t^M with $t = 1, \dots, T_m$, which is an estimate of y_t^M based on y_t^Q and x_t . Note that \hat{y}_t^M is the target series of this article. Because \hat{y}_t^M depends on the available information at the time of estimation, we also define $\tilde{y}_{t|T}^M$, which is \hat{y}_t^M estimated given the information available at time T . The appropriate method to estimate $\tilde{y}_{t|T}^M$ depends on the data available about time t at time T . When both y_t^Q and x_t are available for t (i.e., for $t \leq T_q$), $\tilde{y}_{t|T}^M$ can be estimated with a BM model. We denote this BM estimate by $\tilde{y}_{t|T}^{M,BM}$ where BM indicates the type of BM model. When only x_t is available for time t (i.e., for $T_q < t \leq T_m$), $\tilde{y}_{t|T}^M$ should be estimated with a NC model, denoted as $\tilde{y}_{t|T}^{M,BM,NC}$ where NC indicates the NC model. Note that $\tilde{y}_{t|T}^{M,BM,NC}$ also contains BM in the superscript, because a nowcast changes when the target series is the result of a different BM model. The most interesting element in \hat{y}_t^M from methodological perspective is $\hat{y}_{T_m|T_m}^M$. At this month T_m, x_t with $t = 1, \dots, T_m$ is known, but $\hat{y}_{T_m}^Q$ is not. The full series $\hat{y}_{t|T_m}^M$ can be written as:

$$\hat{y}_{i|T_m}^M = \begin{cases} \hat{y}_{i|T_m}^{M,BM} & \text{for } t = 1, \dots, T_q \\ \hat{y}_{i|T_m}^{M,BM,NC} & \text{for } t = Q(T_m), \dots, T_m \end{cases} \quad (1)$$

The series $\hat{y}_{i|T_m}^{M,BM}$ is based on a BM model that uses y_t^Q and x_t with $t = 1, \dots, T_q$ as input. This implies that each element in $\hat{y}_{i|T_m}^{M,BM}$ changes each time a new quarter becomes available. The series $\hat{y}_{i|T_m}^{M,BM,NC}$ is based on both a BM and NC model and uses y_t^Q with $t = 1, \dots, T_q$ and x_t with $t = 1, \dots, T_m$ as input, so $\hat{y}_{i|T_m}^{M,BM,NC}$ changes each time a new month in x_t becomes available.

Finally, some nowcasting models provide a quarterly estimate $\hat{y}_{i|T_m}^{Q,NC}$ for $t = Q(T_m), \dots, Q(T_m) + 2$. Therefore we define:

$$\hat{y}_{i|T_m}^Q = \begin{cases} y_t^Q & \text{for } t = 1, \dots, T_q \\ y_{i|T_m}^{Q,NC} & \text{for } t = Q(T_m), \dots, Q(T_m) + 2 \end{cases} \text{ and}$$

$$\hat{x}_t = \begin{cases} x_t & \text{for } t = 1, \dots, T_m \\ \hat{x}_{i|T_m}^{NC} & \text{for } t = T_m + 1, \dots, Q(T_m) + 2, \end{cases}$$

which are simply the series y_t^Q and x_t extended with nowcasts for missing values in the current quarter. In this article, x_t is based on the monthly VAT declarants and y_t^Q is based on the combination of all declarants (monthly and quarterly). Furthermore, both series are assumed to be index series, as this is the publication format of these short-term business statistics.

2.2. Benchmarking Models

BM models have the aim of temporally disaggregating a low frequency series into a high-frequency series, with the help of (an) auxiliary high-frequent series. BM can be considered a specific case of temporal disaggregation (Eurostat 2018), where the high-frequency indicator series and the low frequency benchmark series describe the same phenomenon, as is the case in our problem. Extensive literature is available on BM models, see Eurostat (2018), for an overview. The most basic BM model is developed by Denton (Denton 1971; Dagum and Cholette 2006) (DC) and a slightly more advanced BM model is developed by Chow and Lin (Chow and Lin 1971) (CL). The DC and CL BM models are widely used in the production of official statistics and are implemented in standard software (Barcellan and Buono 2002). Both models are also discussed in the Eurostat (2017), because they also might be considered as mixed frequency nowcasting models, which will be discussed in the next section. Both models can disaggregate a quarterly series y_t^Q into a monthly series with the help of a monthly auxiliary series x_t for $t = 1, \dots, T_q$, such that both series are consistent in each quarter. Furthermore, both models aim at movement preservation of the high-frequency series x_t . There are also other BM models available, such as by Fernández (1981) and Litterman (1983), but they are better suited for non-stationary residual models, which in our case is less likely because we use two series that measure the same phenomenon. Both the DC and CL model require high and low-frequency data over the same period. This implies that a new monthly

benchmarked estimate $\hat{y}_{t|T_m}^{M,BM}$ can be obtained only each time new quarterly data becomes available. The difference between CL and DC is that DC aims to preserve the movement by mimicking the month-on-month growth in x_t as close as possible (minimising either the proportional or absolute deviations), while CL is a regression approach that controls for the estimated relation between y_t^Q and x_t . Furthermore, the CL model can deal with more than one auxiliary time series. When the pattern in x_t is representative for the pattern in y_t^Q , DC and CL produce similar results. However, because CL also estimates the relation (coefficient and statistical significance) between y_t^Q and x_t , it may produce a more accurate result, so CL is usually preferred. Nonetheless, because the DC model is widely used and intuitively attractive, we apply and test both models. The BM of both CL and DC is computed using R (R Core Team 2018), using the R package “tempdisagg” (Sax and Steiner 2013) and the function “td”. In this function we set method = “chow-lin-maxlog” for CL and method = “denton-cholette” for DC. The technical details of both BM models are further discussed in Sax and Steiner (2013).

2.3. Nowcasting Models

In literature a wide variety of nowcasting models is discussed. An extensive literature overview is provided by the Eurostat (2017). This Handbook also discusses mixed frequency models, which combine both temporal disaggregation/BM and nowcasting models by dealing with both issues simultaneously as a single missing data problem. For instance, the CL and DC BM models from the previous section can also be considered a mixed frequency NC model. Many of the advanced mixed frequency models are designed to deal with larger sets of auxiliary series and apply multivariate estimation. For example, recently Antolín-Díaz et al. (2021) propose a Bayesian Dynamic Factor model that allows for time series with different frequencies to estimate daily GDP growth. Their model allows the use of a large set of time series and takes things like movements in long-run growth, time-varying uncertainty, and fat tails into account, by utilizing lag-lead properties of, and correlations between, auxiliary macroeconomic series with different frequencies. Frequentist versions of dynamic factor time series models are proposed by Giannone et al. (2008) and Doz et al. (2012). Another option is to use a vector autoregression (VAR) model, which estimates different PPC series simultaneously, see for example Sims (1980) or Stock and Watson (2001). These complex and data intensive models are beyond the scope of this article, as we only consider the case where for each nowcast only one high-frequent auxiliary series and one low-frequent target series, which both measure the same phenomenon, are used. This simple approach has the advantage that resulting estimates have a relatively straightforward interpretation, because the estimates do not depend on a large set of auxiliary series and no mutual dependences between different PPC series are introduced.

We separate the NC models discussed in this article into two groups. The first group we refer to as “NC after BM” models, which are the models that nowcast the high-frequent BM series directly with the help of the auxiliary series. We will discuss them in Subsection 2.3.1. The second group we refer to as “NC before BM” models. These models first nowcast the quarterly and monthly series for the current quarter, and then apply BM to obtain a nowcast for the current month. We will discuss them in Subsection 2.3.2.

2.3.1. NC after BM

The most basic nowcasting model we consider is a simple extrapolation (SE) nowcast model. This can be written as:

$$\hat{y}_{t|T_m}^{M,BM,SE} = \hat{y}_{T_q|T_m}^{M,BM} * \left(\frac{x_t}{x_{T_q}} \right) \quad \text{for } t = Q(T_m), \dots, T_m. \quad (2)$$

This straightforward nowcasting approach is equivalent to the mixed frequency nowcasting model that results from extrapolating the DC BM model. However, Equation (2) is slightly more general in the sense that $\hat{y}_{T_q|T_m}^{M,BM}$ can also be the result of a CL (or any other) BM model.

The second nowcasting model follows directly from the CL BM model, which can also be considered a mixed frequency NC model. CL performs linear regression on the quarterly level with y_t^Q and x_t^Q (x_t^Q is x_t aggregated to the quarterly level) and the estimated linear relation with x_t can be used to extrapolate over $t = Q(T_m), \dots, T_m$. This can be written as:

$$\hat{y}_{t|T_m}^{M,BM,CL} = \beta^{CL} x_t \quad \text{for } t = Q(T_m), \dots, T_m. \quad (3)$$

with β^{CL} the CL regression coefficient.

A third type of nowcasting model is the well-known (seasonal) autoregressive-integrated moving average ((S)ARIMA) model (Box and Jenkins 1976). (S)ARIMA can also incorporate auxiliary variables to obtain a nowcast of a target series. To select an appropriate (S)ARIMA model, a standardized stepwise procedure explained in Hyndman and Khandakar (2008) is used. This method is implemented in their R-package 'forecast' and is used in this article to obtain the nowcast $\hat{y}_{t|T_m}^{M,BM,ARIMA}$ and $\hat{y}_{t|T_m}^{M,BM,SARIMA}$. We apply both models because it is unclear whether the auxiliary variable can cover the seasonal pattern, which is present in most economic time series, as well.

Another method which also applies SARIMA modeling is known as the Benchmark-to-Indicator-ratio (BIR) model (Bloem et al. 2001; Daalmans 2018). Its first step is to estimate a SARIMA model of the ratio $\frac{\hat{y}_{t|T_m}^{M,BM}}{x_t}$ for $t = 1, \dots, T_q$, then obtain a SARIMA nowcast of the ratio series $\frac{\hat{y}_{t|T_m}^{M,BM,BIR}}{x_t}$ for $t = Q(T_m), \dots, T_m$ and finally obtain a nowcast $\hat{y}_{t|T_m}^{M,BM,BIR}$ by multiplying by x_t for $t = Q(T_m), \dots, T_m$. The BIR model might give better results when the ratio between $\hat{y}_{t|T_m}^{M,BM,BIR}$ and x_t is fixed but should be used with care when x_t can have values that are close to zero.

2.3.2. NC before BM

NC before BM models first perform a NC model on x_t and y_t^Q to obtain \hat{x}_t and \hat{y}_t^Q and then use these to obtain $\hat{y}_{t|T_m}^{M,BM,NC}$ with a BM model. This approach might be advantageous when the relation between the series y_t^Q and x_t is stronger on the quarterly level. To obtain $\hat{y}_{t|T_m}^Q$, we consider the Bridge, Mixed Data Sampling (MIDAS), and Structural Time Series (STS) models. In the Bridge and MIDAS model, \hat{x}_t is estimated with a univariate SARIMA model with x_t as input, while in the STS model it is estimated simultaneously with \hat{y}_t^Q .

The Bridge model (see Baffigi et al. 2004; Angelini et al. 2008 for extensive details) consists of a series of SARIMA models. First, \hat{x}_t is estimated. Next, a SARIMA model is estimated with y_t^Q as dependent and x_t^Q as auxiliary variable. Then, by using \hat{x}_t^Q , an

estimate of \hat{y}_t^Q is obtained. Finally, a BM model with \hat{x}_t and \hat{y}_t^Q gives the estimate $\hat{y}_{t|T_m}^{M,BM,Bridge}$.

The MIDAS model approach (see Ghysels et al. 2004, 2007 for extensive details) provides (just like the Bridge model) a quarterly estimate $\hat{y}_{t|T_m}^{Q,MIDAS}$ and a monthly estimate $\hat{y}_{T_m|T_m}^{M,BM,MIDAS}$. The MIDAS model is a regression and filtering technique that incorporates different frequencies. The difference with the Bridge model is that the MIDAS model allows for the modeling of lags in both the quarterly and monthly series simultaneously. According to the literature (e.g., Asimakopulos et al. 2013) an advantage of the MIDAS model as compared to some alternatives such as state space and mixed frequency VAR models, is that the MIDAS model is more parsimonious and less sensitive to specification errors due to the use of non-linear lag polynomials. We estimated the MIDAS model in the R package “midasr” (see Ghysels et al. 2016) and the function `midas_r()` with some basic settings.

A STS model approach (see Durbin and Koopman 2012) is not focused on obtaining $\hat{y}_{T|T_m}^{Q,NC}$ alone, but instead decomposes a time series into a trend, a seasonal component and additional noise. The details of the STS model for this application are described in the supplemental file. In this paragraph we only highlight the novel aspect of modelling the seasonal component. Both y_t^Q and x_t with $t = 1, \dots, T_m$ are used as input of a multivariate STS model, that is both series are modelled jointly in a bivariate setting. Whereas for the other methods the quarterly value is repeated three times in the quarterly series, for the STM the value of this series is missing in the first and second month of each quarter. The quarterly value is used in the third month of each quarter. The STS model approach allows for missing values in the series. In this application, there are missing values in the last quarter, the model estimates for these periods are used as nowcasts. Similarly, as under the other approaches discussed, the auxiliary monthly series is used to improve the accuracy of the nowcasts. But under the STS model approach, this happens by explicitly modelling a correlation between trend disturbance terms of both series. We test two different trend models, the local and smooth trend model. For the local trend model, in the case of zero correlation between y_t^Q and x_t , the predictions are a flat line. For the smooth trend model, in the case of zero correlation between y_t^Q and x_t , the predictions are a linearly increasing or decreasing trend. In both models, in case of non-zero correlation, the predictions are adjusted by the auxiliary monthly series. The STS model can only disaggregate the trend component in monthly estimates. For the seasonal component, an additional BM step, just like with the Bridge and MIDAS model, is required.

We developed a new approach to model the seasonal component of the target series y_t^Q . This seasonal model, which takes missing values into account, is an extension of the known dummy seasonal model.

$$\beta_t^y = \begin{cases} S_t^y & \text{if } t \text{ third month of the quarter} \\ 0 & \text{if } t \text{ first or second month of the quarter,} \end{cases} \tag{4}$$

with

$$S_t^y = \begin{cases} S_{t-10}^y & \text{if } t \text{ first month of the quarter} \\ S_{t-1}^y & \text{if } t \text{ second month of the quarter} \\ -S_{t-3}^y - S_{t-6}^y - S_{t-9}^y + \omega_t^y & \text{if } t \text{ third month of the quarter,} \end{cases} \tag{5}$$

$$E(\omega_t^y) = 0, \tag{6}$$

$$\text{conv}(\omega_t^y, \omega_{t'}^y) = \begin{cases} \sigma_{\omega,y}^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} .$$

With (4)–(6) only the quarterly seasonal pattern of the target series y_t^Q can be estimated. The seasonal component is related with the observed series in the measurement equation of the state space model through ϕ_t^y defined in Equation (4). Since the observations are missing for the first two months of every quarter, ϕ_t^y is equal to zero for the first two months and set equal to the quarterly pattern defined with S_t^y in the third month of the quarter. Equation (5) defines a quarterly seasonal pattern for the months. It is assumed that the monthly seasonal pattern is constant within each quarter. The first two rows of (5) show that during the first two months of a quarter, the seasonal pattern is equal to the value of the quarter in the previous year. The third row of Equation (5) is like the standard dummy seasonal model for a quarterly series. In the third month of each quarter, except for the last quarter, the quarterly observation becomes available and the seasonal pattern for the last quarter (S_t^y) is updated using the values of the previous three quarters (S_{t-3}^y , S_{t-6}^y and S_{t-9}^y) and small change via ω_t^y . The seasonal pattern of the monthly auxiliary series x_t is modelled with a standard trigonometric seasonal component defined at a monthly frequency.

Under the assumption that the seasonal patterns of the monthly declarants and the quarterly declarants is similar, it is desirable that this monthly pattern is adopted by the quarterly series. Nevertheless, (smaller) differences between the seasonal patterns of the monthly and the quarterly series should be considered. This cannot be achieved with the structural time series model. Instead, DC or CL can be applied. It is expected that DC is suboptimal because this model cannot handle negative values easily.

The other components of the STS model are standard. Some adjustments are needed to take the relationship of the monthly and the quarterly figures of the y_t into account. Other adjustments are necessary to consider that the input series are partly based on the same enterprises.

The ideas of the STS model approach are like the Bridge approach. In both cases the auxiliary series is used to predict the target series. There are, however, some differences. First, in the STS model approach the auxiliary series is included as another dependent series and a correlation between trend disturbance terms of target series and auxiliary series is modelled. Both this correlation and the regression parameter in the Bridge approach are assumed to be constant over time. Second, in the STS approach all series are modelled and estimated simultaneously, including prediction of auxiliary series and target series and BM of the trend. Only the BM of the seasonal pattern is performed afterwards.

It might be worthwhile to express the various steps required to fit Bridge models in one state space model. The major advantage of such an approach is that it gives a more realistic approximation of the uncertainty of the nowcasts, since it avoids that estimates obtained in a particular step are treated as known in the next step. Casting a Bridge model in state space form requires that the BM of the target series in the final step is conducted with CL. Subsequently the target series and the auxiliary series are combined in a bivariate state

space model, where both series are modelled with a SARIMA model, see [Durbin and Koopman \(2012, chap. 3\)](#) for details. The SARIMA model for the target series must also include the auxiliary series as a regression component. At the same time the target series, observed at a quarterly frequency, must be modelled on a monthly frequency. Investigating the possibilities of this approach is left for further research.

2.4. Evaluation Method

Altogether we can distinguish eleven different nowcasting models (i.e., SE, CL extrapolation, BIR, ARIMA, SARIMA, Bridge, MIDAS, and the local and smooth trend STS models with and without correlation) that are combined with two BM model variants. To compare their quality, a standard method to evaluate models is to make out-of-sample predictions and check how close these predictions are to the actual outcome. This can be measured by calculating for example the mean absolute error (MAE), which in this case can be computed on the monthly and quarterly level. We choose to look at the MAE because it is also applied in the quality assessment of the short-term business statistics at Statistics Netherlands.

As a benchmark series we use $y_{t|T_{\max}}^{M,CL}$, which is the with CL BM series of y_t^Q (with x_t) with $t = 1, \dots, T_{\max}$, where T_{\max} is simply the last month for which both monthly and quarterly data are available. This allows us to compare each $\hat{y}_{T_m|T_m}^{M,BM,NC}$ with $\hat{y}_{T_m|T_{\max}}^{M,CL}$, where $\hat{y}_{T_m|T_{\max}}^{M,CL}$ is an estimate that is based on a maximum amount of information. The MAE of each series (denoted as $MAE^{M,BM,NC}$) can be written as:

$$MAE^{M,BM,NC} = \frac{\sum_{T_m=T_0}^{T_{\max}} \left| \hat{y}_{T_m|T_m}^{M,BM,NC} - \hat{y}_{T_m|T_{\max}}^{M,CL} \right|}{T_{\max} - T_0}, \quad (7a)$$

where T_0 is the first month of the evaluation period. T_0 should not be too early in the time series, because each model requires a period of calibration. The quarterly MAE (denoted as $MAE^{Q,BM,NC}$) can be obtained by:

$$MAE^{Q,BM,NC} = \frac{\sum_{T_m=T_0}^{T_{\max}} \left| \left(\sum_{T_m=Q(T_m)}^{Q(t)+2} \hat{y}_{T_m|T_m}^{M,BM,NC} \right) / 3 - y_{T_m}^Q \right|}{3(T_{\max} - T_0)}. \quad (7b)$$

The $MAE^{Q,BM,NC}$ has the advantage that the estimations are compared with the observed series y_t^Q instead of the estimated $\hat{y}_{t|T_{\max}}^{M,CL}$.

3. Empirical Evaluation of the Nowcast Models

In this section we apply and compare the models that were introduced in Section 2. We first describe the data in Subsection 3.1. Next, in Subsection 3.2, we discuss how the models compare in terms of their MAE. In the last Subsection (3.3) we discuss how they perform before, during and after a simulated economic shock.

3.1. Time Series Data

We apply all models from Section 2 on index time series from twelve PPCs, that cover the period January 2010 until June 2020, and that represent twelve different economic

activities in the Netherlands based on four-digit NACE (Nomenclature statistique des activités économiques dans la Communauté européenne, Eurostat 2008). Six of them represent the hospitality sector (i.e., Hotels, Other accommodation, Restaurants, Fast food, Catering and Pubs) and six of them represent other activities in the service sector (i.e., Publishers, Legal activities, Accountants, Employment activities, Other Business Support, Repair of household goods). We refer to them as PPCs 1–12, in the same order as above. For all twelve series both y_t^Q and x_t are available. The auxiliary monthly index series x_t is based on the raw turnover data from monthly declarants that declared turnover in all consecutive months of the series and is only corrected for new and bankrupt companies. The published quarterly index series y_t^Q can be considered of higher quality, because it is based on turnover data from both monthly and quarterly declarants, it is manually corrected for errors, for new and bankrupt companies, and is complemented with primary data collection for a small group of exceptionally large companies. Both y_t^Q and x_t are rescaled such that they have a mean value of 100 in the year 2015.

The series y_t^Q and x_t for PPC 1, 2 and 9 are presented in Figure 1 below. These PPCs illustrate how similar or different both series can be.

In the graphs of PPC 1, 2 and 9 we see that y_t^Q and x_t can be correlated to different degrees. The graphs also show that the seasonal pattern generally presents itself in both the

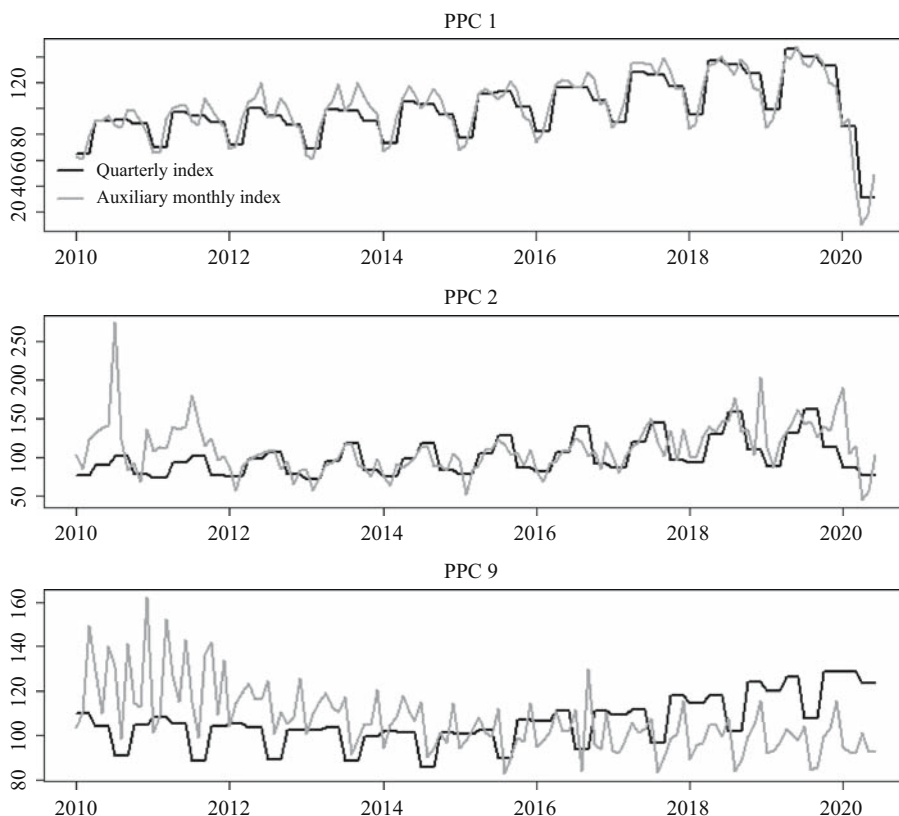


Fig. 1. Published quarterly series and monthly auxiliary index series of PPC 1, 2 and 9, over the period January 2010–June 2020.

monthly and the quarterly series. However, we also see that the relation between the series may or may not be stable over the entire period.

Furthermore, we see that the different PPCs are affected differently by the COVID-19 pandemic in the second quarter of 2020. For example, Hotels (PPC 1) show a big collapse in both the monthly and the quarterly series, while Accountants (PPC 9) seem to be hardly affected. The COVID-19 pandemic raises the question whether the models that perform best during a stable economic period also perform best during a crisis. This is investigated by evaluating the nowcast models separately over the period July 2016 until February 2020 (stable period) and over the period March 2020 until June 2020 (crisis period).

3.2. Nowcast Model Performance Before and During a Crisis

Table 1 shows the nowcasting results for all BM and NC model combinations and PPCs. The last two columns show the unweighted and weighted (by annual turnover in 2015) results over all PPCs.

In Table 1 the various models are evaluated on the monthly level during the stable economic period July 2016–February 2020. The grey cells in Table 1 are the cells in the top three of lowest MAEs in each PPC column. The rows that represent the Bridge and MIDAS models contain most of these grey cells, which implies they generally perform better than the other models in the table. The best model according to both the unweighted and weighted mean, is the Bridge model combined with the CL BM model. As was anticipated in Subsection 2.3.2, the STS models in combination with DC perform clearly worse.

Table 2 shows MAEs for each model/PPC combination, but now the MAE is calculated over the quarters in the period July 2016–December 2019 (The last quarter that was unaffected by the COVID-19 pandemic).

Table 2 yields the same conclusion as Table 1, as again the Bridge and MIDAS model perform quite well. To illustrate this graphically, Figure 2 shows $\hat{y}_{t|T}^{M,CL}$ and $\hat{y}_{T_m|T_m}^{M,CL,Bridge}$

Table 1. MAE^{M,BM,NC} over July 2016–February 2020 for all 12 PPCs, plus an unweighted and weighted mean.

Series and NC type		BM model	Series/PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	6.5	15.2	4.7	6.1	8.6	4.9	7.1	7.1	7.0	3.5	4.7	2.8	6.5	5.5
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.4	3.3	0.5	0.7	0.6	1.0	1.2	0.3	1.2	3.0	2.7	2.1	1.4	1.7
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	2.3	26.0	1.8	8.6	4.3	4.2	11.8	4.8	9.1	4.2	5.9	6.6	7.5	5.9
			CL extrapolation	3.2	22.0	1.7	7.5	4.4	5.2	9.9	5.5	10.4	4.7	7.9	5.3	7.3	6.3
			ARIMA	2.9	20.0	1.6	7.1	4.6	3.9	8.5	4.8	8.2	4.2	5.4	4.9	6.4	5.3
			SARIMA	3.1	18.4	1.5	6.8	3.6	4.0	7.8	4.1	6.7	4.3	4.9	4.2	5.8	4.9
			BIR	3.1	13.1	1.5	5.5	4.3	3.1	4.5	2.4	5.6	6.9	5.6	7.5	5.3	5.2
		DC	Simple extrapolation	2.1	25.7	1.7	9.5	4.3	3.1	12.9	4.8	9.2	4.9	6.6	7.1	7.7	6.3
	ARIMA	1.9	20.5	1.4	7.2	3.7	3.3	7.9	5.8	6.9	4.7	3.9	7.3	6.2	5.2		
	SARIMA	1.9	10.9	1.2	5.9	4.1	2.5	4.6	2.4	4.8	4.9	4.5	5.6	4.4	4.2		
	BIR	2.1	22.0	1.5	8.7	4.1	2.9	10.2	3.7	6.0	4.9	5.4	6.4	6.5	5.3		
	NC before BM	CL	Bridge	3.1	6.1	1.3	3.3	3.2	2.2	3.6	1.8	3.8	4.0	3.2	2.9	3.2	3.3
			MIDAS	3.2	6.8	1.4	3.5	3.5	2.1	4.4	1.6	3.8	5.0	3.8	3.6	3.5	3.8
			DC	3.1	5.6	1.3	5.0	3.0	2.3	3.9	1.9	3.9	4.4	4.1	2.6	3.4	3.6
DC		Bridge	3.2	6.5	1.3	5.2	3.2	2.3	4.5	1.8	4.2	5.6	4.7	3.4	3.8	4.2	
		MIDAS	3.1	5.6	1.3	5.0	3.0	2.3	3.9	1.9	3.9	4.4	4.1	2.6	3.4	3.6	
		DC	3.2	6.5	1.3	5.2	3.2	2.3	4.5	1.8	4.2	5.6	4.7	3.4	3.8	4.2	
State-space	NC before BM	CL	STS, local trend model, no correlation	3.0	13.5	2.6	4.7	3.7	3.9	5.5	3.7	6.6	3.1	4.2	3.6	4.8	4.2
			STS, smooth trend model, no correlation	2.0	13.2	2.1	4.6	3.6	4.0	5.7	3.4	6.3	2.6	4.3	3.4	4.6	3.8
			STS, local trend model, correlation	2.6	13.6	1.8	4.7	3.6	3.9	5.6	3.7	6.6	2.6	4.1	3.4	4.7	3.9
		STS, smooth trend model, correlation	1.9	13.0	1.8	4.7	3.5	4.0	5.8	3.4	6.2	2.6	4.2	3.4	4.5	3.8	
		DC	STS, local trend model, no correlation	4.2	25.9	3.6	10.4	3.7	5.0	6.4	8.3	11.1	3.4	6.3	8.0	8.0	6.2
			STS, smooth trend model, no correlation	2.9	26.4	3.0	10.4	3.5	5.2	6.8	8.0	10.9	4.3	6.7	7.8	8.0	6.4
STS, local trend model, correlation	4.2		26.4	2.6	10.3	3.5	4.5	6.7	8.3	11.1	3.2	6.2	7.9	7.9	6.0		
STS, smooth trend model, correlation	3.4	25.9	2.8	10.5	3.3	5.4	7.1	8.0	10.6	4.3	6.1	7.8	7.9	6.3			

Table 2. $MAE^{O,BM,NC}$ over July 2016–December 2019 for all 12 PPCs, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	2.1	26.0	1.8	8.1	4.1	4.1	11.3	4.9	8.8	2.6	5.1	6.1	7.1	5.2
			CL extrapolation	3.0	21.3	1.7	6.9	4.0	5.3	9.2	5.5	10.0	4.7	7.5	5.2	7.0	6.1
			ARIMA	2.7	19.1	1.6	6.5	4.3	4.0	8.1	4.8	8.0	4.2	4.9	4.8	6.1	5.2
			SARIMA	2.8	17.6	1.5	6.4	3.2	4.1	7.2	4.2	6.5	4.3	4.4	4.1	5.5	4.8
			BIR	2.8	11.5	1.4	5.2	4.1	3.1	4.0	2.5	5.4	7.1	4.6	7.4	4.9	5.1
			DC	Simple extrapolation	1.8	26.2	1.7	9.0	4.0	2.9	12.2	4.9	9.0	3.3	5.7	7.0	7.3
	ARIMA	1.8	20.5	1.4	6.2	3.4	3.2	7.4	6.0	6.6	3.0	3.6	7.3	5.9	4.5		
	SARIMA	1.6	10.0	1.1	5.3	3.8	2.4	3.9	2.4	4.1	3.5	4.2	5.4	4.0	3.4		
	BIR	1.8	22.2	1.5	8.3	3.8	2.8	9.8	3.8	5.2	3.0	4.6	6.2	6.1	4.5		
	CL	Bridge	3.0	3.3	1.3	2.6	2.9	1.9	2.3	1.2	3.0	3.8	2.4	2.3	2.5	2.8	
	MIDAS	3.0	3.5	1.3	2.6	2.9	1.9	2.2	1.2	3.1	4.0	2.3	2.5	2.6	2.9		
	DC	Bridge	3.0	3.4	1.0	4.5	2.7	2.3	2.6	1.4	3.7	4.3	3.7	2.1	2.9	3.3	
MIDAS	3.0	3.2	1.0	4.5	2.7	2.2	2.5	1.5	3.9	4.7	3.7	2.3	2.9	3.4			
State-space	NC before BM	CL	STS, local trend model, no correlation	2.8	12.0	2.6	3.6	1.5	3.5	2.1	2.8	5.0	2.8	3.6	2.7	3.7	3.3
			STS, smooth trend model, no correlation	1.9	11.7	1.7	3.7	1.7	3.6	2.2	2.5	4.4	2.0	3.2	2.5	3.4	2.8
			STS, local trend model, correlation	2.4	12.0	1.7	3.7	1.4	3.4	2.2	2.8	5.0	2.2	3.5	2.5	3.6	3.0
			STS, smooth trend model, correlation	1.7	11.4	1.6	3.7	1.5	3.6	2.3	2.5	4.3	2.0	3.1	2.5	3.3	2.7
	DC	STS, local trend model, no correlation	3.2	17.5	3.4	7.1	1.9	4.4	5.0	6.0	5.8	3.0	5.4	4.3	5.6	4.5	
	STS, smooth trend model, no correlation	1.8	19.7	2.2	7.3	1.6	4.9	4.8	6.1	6.0	2.7	5.7	4.0	5.6	4.3		
	STS, local trend model, correlation	3.2	17.6	2.3	7.1	1.7	4.0	5.1	6.0	5.9	2.4	5.5	4.2	5.4	4.2		
	STS, smooth trend model, correlation	2.0	19.5	2.2	7.3	1.5	5.1	5.1	6.1	6.0	2.7	5.1	3.9	5.5	4.3		

for PPC 1, 2 and 9. It shows that $\hat{y}_{T_m|T_m}^{M,CL,Bridge}$ performs quite well for all three PPCs, except it underestimates the impact of the COVID-19 pandemic in the last few months. Therefore, to see if a different nowcast model should be preferred during a crisis, Table 3 shows the MAEs for the period March 2020–June 2020.

Table 3 shows that during a crisis, no longer the Bridge and MIDAS model that were fitted on data that is largely from periods prior to the crisis provide the most accurate nowcasts, but the more basic models perform somewhat better. This is not surprising, because they rely less on the past and more on recent data. The DC, (S)ARIMA method is the most accurate method among the direct ones. Surprisingly, the DC, (S)ARIMA model is clearly more accurate than the CL, (S)ARIMA model. The explanation lies in the AR(1) term that is part of both the CL and (S)ARIMA model. Therefore, the CL, (S)ARIMA model puts more weight on the past than the DC, (S)ARIMA model. This is a disadvantage during a crisis.

To further investigate how robust the above results are with respect to the MAE evaluation method, two other evaluation methods are applied. The first method counts in how many periods a specific method is more accurate than CL, Bridge (before the COVID-pandemic) or DC, SARIMA (during the first months of the COVID-pandemic). The second measure counts how often the relative prediction error, is very large. These alternative evaluation methods confirm the earlier results based on the MAE. Details about these alternative evaluation methods and the results are available from the authors on request.

A question that remains unanswered in this section, is which model should be preferred after a crisis? How long after a crisis will the CL, Bridge model start to outperform the DC, SARIMA model again? This question is the subject of the next section.

3.3. Nowcast Model Performance After a Crisis

To investigate the performance of the models during a longer crisis and after a crisis, we simulate three different types of economic shocks in January 2017. Each shock implies that

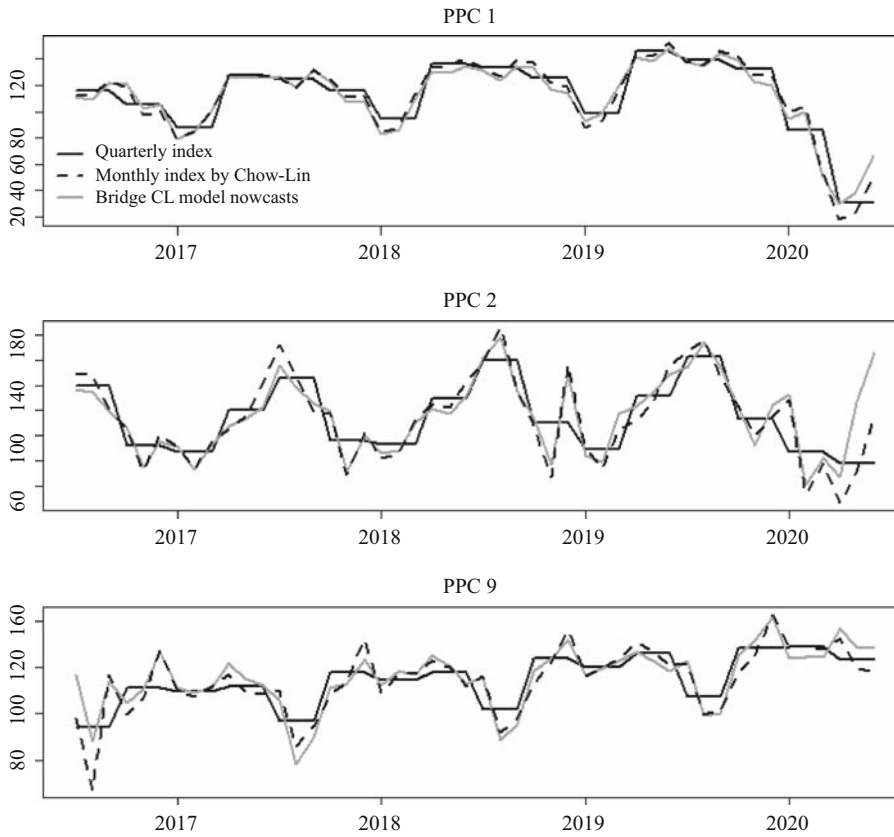


Fig. 2. Quarterly, monthly and Bridge CL NC index series of PPC 1, 2 and 9, over the period July 2016–June 2020.

Table 3. MAE^{BM,NC} over March 2020–June 2020 for all 12 PPCs, plus an unweighted and weighted mean.

Series and NC type	BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean	
Published	None	Quarterly series	18.2	20.6	23.3	5.3	12.1	22.6	3.5	3.2	4.6	3.2	5.3	4.8	10.6	8.0	
BM series	CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	DC	Monthly BM series	4.6	8.1	10.0	6.2	1.1	11.8	4.4	0.5	0.9	3.0	3.1	5.8	5.0	4.0	
ARIMA/ Regression	NC after BM	Simple extrapolation	7.3	28.3	10.0	30.0	7.8	13.6	13.4	7.9	12.1	2.0	15.3	7.5	12.9	9.2	
		CL extrapolation	6.9	27.6	6.9	28.2	6.2	5.6	22.5	7.8	7.0	4.9	7.8	6.6	11.5	8.9	
		ARIMA	6.5	18.0	6.4	29.2	6.6	7.0	18.7	6.2	2.9	5.6	6.4	4.7	9.9	7.9	
		SARIMA	6.7	23.9	6.4	29.0	6.2	7.8	15.5	5.7	2.7	5.5	6.1	4.7	10.0	7.7	
		BIR	11.1	28.6	9.2	26.5	6.5	15.5	9.9	2.2	2.9	1.7	9.5	8.5	11.0	6.9	
		DC	Simple extrapolation	5.9	26.6	12.0	30.4	7.2	10.9	11.6	7.5	13.0	5.4	12.4	6.2	12.4	10.0
	ARIMA	2.2	23.3	12.6	28.3	4.8	4.2	7.1	6.7	4.2	2.7	2.9	3.8	8.6	6.2		
	SARIMA	2.5	26.0	12.5	26.9	4.4	3.4	5.3	1.4	3.8	2.4	2.2	4.9	8.0	5.4		
	BIR	5.6	23.2	12.5	31.1	6.4	9.8	10.6	6.0	5.9	2.7	8.1	6.6	10.7	7.7		
	NC before BM	CL	Bridge	11.4	32.6	3.7	6.7	23.4	43.4	6.6	2.9	6.7	7.5	6.1	4.7	13.0	9.3
			MIDAS	14.4	34.6	7.6	10.7	22.3	58.7	7.3	2.6	6.0	8.2	6.2	8.4	15.6	10.9
			DC	71.7	31.4	16.1	17.3	27.0	15.4	4.4	2.9	7.7	14.0	9.6	8.6	18.9	16.0
DC		Bridge	76.6	32.4	11.3	19.1	25.8	26.7	6.9	2.6	7.2	14.6	8.8	10.9	20.2	16.7	
		MIDAS															
		DC															
State-space	NC before BM	STS, local trend model, no correlation	86.0	21.3	61.1	18.6	49.9	47.2	5.2	2.9	6.3	17.1	10.6	5.1	27.6	24.0	
		STS, smooth trend model, no correlation	93.2	24.9	73.3	24.8	49.9	47.3	4.9	2.5	5.8	17.8	13.5	7.4	30.4	26.3	
		STS, local trend model, correlation	37.7	18.8	2.0	12.9	34.1	36.1	3.5	2.7	6.3	12.8	9.8	4.8	15.1	12.6	
		STS, smooth trend model, correlation	52.9	27.7	21.8	6.6	37.8	40.9	4.0	2.5	5.7	15.9	12.7	7.3	19.7	16.8	
		DC	STS, local trend model, no correlation	90.3	31.8	68.7	22.1	52.2	41.5	6.0	4.9	7.0	17.6	9.5	5.9	29.8	25.6
		STS, smooth trend model, no correlation	112.8	34.3	127.3	25.0	53.8	42.5	4.8	3.8	6.8	18.3	12.2	8.7	37.5	33.0	
STS, local trend model, correlation	39.6	32.1	2.4	32.0	35.6	49.9	6.4	4.7	7.0	13.0	8.7	5.1	19.7	15.0			
STS, smooth trend model, correlation	70.9	27.6	31.5	15.6	40.6	51.8	5.1	3.9	8.3	16.3	11.9	8.6	24.3	20.4			

both $y_{\text{January } 2017}^Q$ and $x_{\text{January } 2017}$ are divided by 2. The first type (shock 1) does not recover, the second (shock 2) recovers during the next quarter and the third (shock 3) slowly recovers during a period of one year. Figure 3 shows these three shocks with PPC 1 as an example. In this simulation study the COVID-19 pandemic crisis period is excluded from the analysis.

To investigate the direct and long-term effect of a shock on the performance of different models, we calculate the MAE separately over the period January 2017–December 2017 for shock 1 and 3 (results in Table 4 and 6), over the period January 2017–March 2017 for shock 2 (results in Table 5) and January 2018–December 2019 for all shocks (results in Table 7–9). Just like during the COVID-19 pandemic crisis, the NC after BM models outperform the NC before BM models.

Table 4–6 shows that the BIR model performs among the best three models (according to both the weighted and unweighted mean) during all three shocks. This can be partly explained by the simulation setup because the BIR model is based on predicting the ratio $\frac{\hat{y}_t^{M,BM}}{x_t}$, which is by construction hardly disturbed by our artificial shocks, because both y_t^Q and x_t are multiplied with the same factor. In a real crisis, both series might be affected in different ways, which could make the other methods more competitive, as was seen in Table 3. Furthermore, Table 4–6 shows that some models have a serious problem in nowcasting the second type of

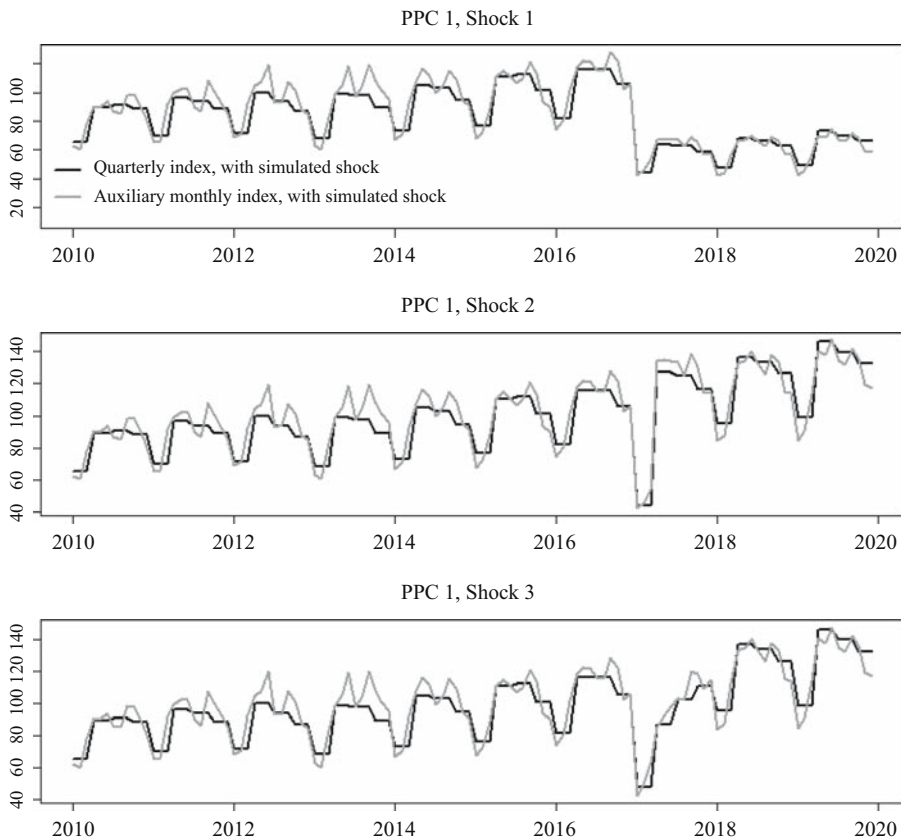


Fig. 3. Illustration of three types of shocks (as of January 2017) with PPC 1 as example.

Table 4. $MAE^{M,BM,NC}$ over January 2017–December 2017 for all 12 PPCs with shock 1, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	2.4	7.2	2.2	3.1	3.6	1.8	3.7	3.8	3.0	1.7	3.9	2.6	3.2	2.7
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.1	1.5	0.1	0.2	0.4	0.4	0.7	0.2	1.0	2.0	0.6	0.6	0.7	1.0
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	0.8	12.6	0.7	3.6	3.2	2.2	5.6	3.3	2.7	3.8	3.2	5.0	3.9	3.4
			CL extrapolation	0.7	11.0	1.1	3.2	4.0	2.7	8.6	4.2	4.1	12.3	4.1	11.3	5.6	6.9
			ARIMA	0.7	8.2	1.0	2.6	3.8	2.5	8.8	3.8	3.3	12.2	2.2	10.5	5.0	6.5
		SARIMA	1.0	7.6	0.9	3.0	3.1	2.0	8.2	4.1	3.3	11.5	2.4	10.6	4.8	6.2	
		BIR	0.7	7.3	0.6	2.4	3.3	1.6	2.0	2.7	2.9	5.5	3.2	4.8	3.1	3.4	
		DC	Simple extrapolation	0.9	11.8	0.7	3.8	2.9	1.6	5.6	3.3	3.3	2.6	3.7	4.5	3.7	3.0
	ARIMA		0.8	11.9	1.0	2.4	2.2	5.0	6.6	5.9	9.9	5.7	6.3	8.3	5.5	5.5	
	SARIMA		0.8	5.0	0.5	2.5	2.2	4.9	3.8	3.2	11.8	4.3	6.3	8.3	4.5	4.6	
	NC before BM	CL	Bridge	1.3	32.2	5.5	11.2	13.3	8.6	8.1	7.0	15.6	12.1	12.9	17.1	12.1	11.0
			MIDAS	1.3	28.7	5.0	12.2	12.7	7.3	8.4	6.6	14.4	10.4	12.0	15.6	11.2	10.0
			Bridge	5.0	33.0	8.0	7.9	10.4	8.4	11.3	7.1	12.0	13.5	11.9	19.3	12.3	11.4
		DC	MIDAS	5.2	29.8	7.7	8.8	10.3	6.2	11.7	6.2	10.9	11.9	10.9	17.6	11.4	10.5
STS, local trend model, no correlation			18.4	21.3	16.1	18.1	17.3	21.8	18.0	19.1	23.1	21.5	25.6	17.5	19.8	20.5	
STS, smooth trend model, no correlation			24.2	27.7	21.5	22.8	16.8	14.8	17.8	16.2	20.6	25.1	20.0	19.9	20.6	21.7	
State-space	NC before BM	CL	STS, local trend model, correlation	6.4	20.3	4.0	8.1	11.1	9.8	10.5	9.5	14.9	15.5	17.2	15.2	11.9	12.5
			STS, smooth trend model, correlation	9.1	26.5	9.1	11.4	14.8	17.5	16.7	11.9	16.8	15.5	19.5	15.1	15.3	14.9
			STS, local trend model, no correlation	25.2	38.6	14.9	30.5	18.2	29.8	17.1	19.3	27.6	24.6	24.2	16.6	23.9	23.6
	DC	STS, smooth trend model, no correlation	29.7	43.2	25.7	24.2	17.7	25.0	19.1	17.7	29.0	25.7	19.1	17.9	24.5	24.7	
		STS, local trend model, correlation	12.0	40.7	6.0	18.1	12.1	18.5	12.5	12.4	16.8	17.2	13.7	15.2	16.3	15.2	
		STS, smooth trend model, correlation	13.2	45.3	12.7	15.0	15.8	27.1	18.1	15.2	23.0	16.7	19.0	16.3	19.8	18.0	

Table 5. $MAE^{M,BM,NC}$ over January 2017–December 2017 for all 12 PPCs with shock 2, plus an unweighted and weighted mean.

Series and NC type		BM model	Series\PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	3.9	4.1	3.7	3.7	1.2	2.9	4.5	6.4	1.4	5.3	5.4	4.7	3.9	4.2
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.1	0.9	0.1	0.9	0.5	0.6	1.3	1.0	1.0	2.7	0.9	0.8	0.9	1.4
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	0.8	5.4	0.3	2.2	2.4	2.8	8.9	1.7	1.7	2.5	3.3	1.3	2.8	2.6
			CL extrapolation	0.8	6.9	1.8	2.6	2.2	1.8	22.6	4.5	2.5	38.3	3.8	29.8	9.8	16.6
			ARIMA	1.0	0.6	1.9	2.4	1.8	4.9	24.4	4.7	5.4	37.6	0.7	29.4	9.6	16.6
		SARIMA	0.2	0.7	1.5	3.5	2.3	2.0	22.7	6.3	2.1	38.1	0.6	29.8	9.2	16.2	
		BIR	0.6	2.8	0.2	1.5	0.7	1.9	3.4	1.7	5.0	5.1	3.9	2.9	2.5	3.3	
		DC	Simple extrapolation	0.8	5.8	0.4	2.2	2.7	2.2	7.7	1.8	1.7	2.2	2.4	1.6	2.6	2.4
	ARIMA		0.8	17.2	1.7	2.5	0.9	14.1	6.9	10.8	17.5	9.2	12.1	14.3	9.0	9.0	
	SARIMA		0.3	8.4	0.2	4.5	1.4	12.6	2.4	9.7	17.5	8.1	14.5	15.9	8.0	8.1	
	NC before BM	CL	Bridge	1.0	33.8	19.6	42.1	28.7	30.0	21.9	21.9	45.1	38.3	33.3	45.1	30.1	31.7
			MIDAS	0.7	29.6	17.8	46.3	26.4	25.0	20.1	19.9	38.6	32.6	31.0	39.7	27.3	28.1
			Bridge	16.4	36.9	28.0	26.9	24.7	31.0	23.1	23.3	35.2	43.7	36.6	40.8	30.5	33.7
		DC	MIDAS	17.1	32.4	26.6	29.4	23.0	21.8	22.1	20.1	29.1	38.5	33.7	35.7	27.5	30.1
STS, local trend model, no correlation			33.9	31.1	35.6	41.7	42.3	33.9	42.2	43.1	43.5	46.7	48.6	44.5	40.6	42.7	
STS, smooth trend model, no correlation			35.9	34.1	42.0	48.7	42.8	36.9	41.0	42.9	46.9	47.4	52.3	44.0	42.9	44.7	
State-space	NC before BM	CL	STS, local trend model, correlation	20.4	31.9	13.8	28.3	34.7	33.6	34.5	32.6	45.0	46.0	44.6	43.0	34.0	37.2
			STS, smooth trend model, correlation	27.3	33.9	31.7	38.0	39.0	35.8	35.6	36.4	43.7	48.5	52.0	41.4	38.6	41.6
			STS, local trend model, no correlation	59.0	118.6	31.5	75.2	44.9	69.0	44.8	44.5	34.7	46.0	24.5	38.1	51.2	46.2
	DC	STS, smooth trend model, no correlation	58.3	117.1	58.6	50.5	46.1	75.0	44.0	45.5	41.0	45.5	41.3	38.8	55.1	49.6	
		STS, local trend model, correlation	42.3	118.9	21.3	53.9	37.1	68.7	36.2	33.8	36.5	46.8	24.4	36.3	46.4	41.9	
		STS, smooth trend model, correlation	43.5	116.8	44.6	39.1	41.8	74.7	37.4	37.9	39.2	46.9	39.7	35.2	49.7	45.7	

shock, leading to very high MAEs. This concerns all the NC before BM models, the CL extrapolation and CL, (S)ARIMA models, which show large mean MAEs due to large MAEs for PPC 7, 10 and 12. The main reason for this last result is that the CL, (S)ARIMA model estimates a correlation between y_t^Q and x_t , which might be overestimated due to the artificial shock in both series. A final point of interest is that the NC after BM models have more problems with the gradual recovering shock 3 than shock 1 and 2.

Table 6. $MAE^{M,BM,NC}$ over January 2017–December 2017 for all 12 PPCs with shock 3, plus an unweighted and weighted mean.

Series and NC type		BM model	Series/PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	5.5	12.2	5.4	5.3	8.8	5.1	8.7	8.5	5.9	5.7	8.7	5.5	7.1	6.6
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.3	2.2	0.1	0.4	0.7	0.5	1.3	0.6	1.2	2.4	0.8	1.3	1.0	1.3
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	1.5	21.9	1.4	5.0	4.9	3.3	8.3	5.1	4.0	4.9	4.9	8.2	6.1	4.9
			CL extrapolation	1.3	17.9	1.6	4.8	6.5	4.8	10.4	5.7	7.2	15.3	6.8	13.1	7.9	9.2
			ARIMA	1.2	14.9	1.5	3.8	6.3	4.5	10.6	5.3	5.6	14.7	4.0	12.2	7.0	8.5
		SARIMA	1.5	13.5	1.4	4.1	5.5	4.2	9.4	5.3	3.4	15.1	4.5	12.2	6.7	8.2	
		BIR	1.3	12.4	1.1	3.1	5.9	2.7	2.6	3.6	4.0	7.6	4.9	7.8	4.8	4.9	
		DC	Simple extrapolation	1.8	20.9	1.3	5.4	4.1	2.5	8.9	5.3	4.7	3.2	5.5	7.8	6.0	4.5
	ARIMA		1.6	17.4	1.5	3.2	3.5	6.4	6.5	7.9	11.1	5.6	7.5	9.8	6.8	6.2	
	SARIMA		1.5	7.3	0.9	3.6	3.7	4.4	2.9	4.8	9.8	4.3	8.6	9.7	5.1	4.8	
	NC before BM	CL	Bridge	1.7	18.4	5.6	10.6	11.7	11.5	9.0	7.0	13.3	12.4	13.5	15.1	10.8	10.6
			MIDAS	1.8	18.4	5.4	12.3	11.3	11.1	8.7	6.9	12.8	11.5	13.5	14.6	10.7	10.3
			DC	Bridge	5.9	24.4	8.1	9.9	10.2	9.0	11.5	8.9	10.8	12.3	12.4	15.4	11.6
		DC	MIDAS	6.2	23.5	8.2	10.7	10.1	7.1	11.3	8.5	9.9	11.9	12.0	14.7	11.2	10.6
STS, local trend model, no correlation			18.3	19.1	17.5	18.1	19.8	14.2	17.7	17.2	20.1	22.5	21.2	18.3	18.7	19.9	
STS, local trend model, correlation			24.7	24.0	18.8	20.6	22.5	16.0	19.9	20.0	19.9	20.6	21.8	27.1	21.9	23.0	
State-space	CL	STS, local trend model, correlation	6.5	19.3	4.1	8.7	10.9	12.6	11.9	10.9	16.1	14.3	18.4	12.9	12.2	12.6	
		STS, smooth trend model, correlation	10.4	25.5	6.4	11.5	13.5	17.4	18.1	13.0	18.4	18.1	23.6	17.0	16.1	16.2	
		DC	STS, local trend model, no correlation	24.2	49.5	16.6	28.9	20.6	23.4	18.0	18.4	19.9	24.4	16.9	18.0	23.2	22.1
	DC	STS, smooth trend model, no correlation	29.2	55.1	23.2	23.0	22.2	25.7	23.9	22.1	21.4	30.2	22.1	22.4	26.7	26.4	
		STS, local trend model, correlation	12.0	50.2	6.1	18.9	11.3	20.8	13.3	14.3	17.3	15.9	14.3	15.5	17.5	15.5	
		STS, smooth trend model, correlation	14.0	57.6	6.6	15.3	14.4	26.9	23.6	17.3	22.1	18.7	24.3	16.8	21.5	19.2	

Next, in Table 7 we look at the MAEs of the different models in the second and third year after shock 3. Just as in the analysis of the real data in Subsection 3.2, the CL, Bridge model is again among the better performing nowcast models during this more stable period. We see that the NC models are less accurate than without the (simulated) crisis (also when we recalculate Table 1 over the same period January 2018–December 2019, not shown). Of course, this is no surprise.

Table 7 shows that in year two and three after the shock, the CL, Bridge model is back among the best performing models, overtaking the models that did well in the first year

Table 7. $MAE^{M,BM,NC}$ over January 2018–December 2019 for all 12 PPCs after shock 3, plus an unweighted and weighted mean.

Series and NC type		BM model	Series/PPC	PPC1	PPC2	PPC3	PPC4	PPC5	PPC6	PPC7	PPC8	PPC9	PPC10	PPC11	PPC12	Unweighted mean	Weighted mean
Published		None	Quarterly series	5.9	15.3	4.2	5.1	9.1	4.7	6.9	6.5	6.4	4.5	6.4	4.5	6.6	5.7
BM series		CL	Monthly BM series	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		DC		0.3	3.1	0.2	0.4	0.5	0.7	1.2	0.3	1.1	1.1	0.9	0.5	0.8	0.8
ARIMA/ Regression	NC after BM	CL	Simple extrapolation	2.5	26.7	2.0	10.3	3.9	3.7	10.4	5.5	6.6	4.1	4.4	7.9	7.3	5.5
			CL extrapolation	3.7	21.7	2.0	8.9	4.2	5.6	9.8	5.6	9.3	5.1	8.1	8.0	7.7	6.4
			ARIMA	3.4	19.0	1.9	8.2	4.8	3.9	8.5	4.9	6.7	4.6	5.4	6.8	6.5	5.4
		SARIMA	3.5	18.1	1.8	8.1	3.7	4.5	7.9	4.1	4.8	4.3	4.7	6.9	6.0	4.9	
		BIR	3.4	11.6	1.6	5.9	5.0	3.0	4.1	2.6	3.4	4.0	4.7	7.1	4.7	3.9	
		DC	Simple extrapolation	2.1	27.3	1.9	10.9	4.1	2.6	11.6	5.6	6.8	3.9	5.4	7.6	7.5	5.7
	ARIMA		2.0	20.0	1.5	7.6	4.5	4.3	7.8	6.4	4.6	3.8	3.8	7.5	6.1	4.7	
	SARIMA		2.0	15.0	1.4	6.4	3.9	3.6	4.5	2.5	5.7	3.5	4.5	7.2	5.0	4.0	
	NC before BM	CL	Bridge	2.2	22.9	1.7	9.8	3.9	2.6	8.8	4.1	3.6	3.9	4.4	6.9	6.2	4.7
			MIDAS	3.2	8.4	1.5	4.0	3.9	3.4	4.0	2.1	5.0	3.9	4.3	4.9	4.0	3.7
			DC	Bridge	3.2	9.9	1.5	4.1	3.8	3.3	5.0	2.1	4.7	4.0	4.2	4.9	4.2
		DC	MIDAS	3.6	9.3	1.8	5.3	4.6	3.2	8.4	3.4	3.5	3.3	5.2	7.7	4.9	4.0
STS, local trend model, no correlation			7.6	14.5	7.0	8.5	6.5	6.2	6.4	7.3	8.9	8.3	8.9	9.1	8.3	8.0	
STS, smooth trend model, no correlation			6.8	14.6	5.4	7.1	6.0	5.5	7.1	7.6	9.4	8.7	6.6	11.2	8.0	7.9	
State-space	CL	STS, local trend model, correlation	3.3	13.9	1.5	4.8	4.4	6.8	6.6	3.8	6.6	3.8	6.9	6.8	5.9	5.0	
		STS, smooth trend model, correlation	6.1	14.2	5.2	6.9	4.8	5.6	7.3	6.4	10.6	6.5	7.3	8.8	7.5	7.1	
		DC	STS, local trend model, no correlation	7.8	30.7	7.6	11.1	7.4	6.9	9.9	12.5	12.3	8.0	11.6	12.6	11.5	9.9
	DC	STS, smooth trend model, no correlation	7.3	29.7	5.7	9.3	7.2	6.1	10.8	11.1	12.6	8.4	9.1	14.6	11.0	9.6	
		STS, local trend model, correlation	4.0	30.2	1.8	8.7	4.6	7.3	11.1	6.8	11.3	4.6	8.8	10.4	9.1	7.0	
		STS, smooth trend model, correlation	6.4	30.7	5.6	9.0	5.5	6.6	10.6	9.0	14.0	7.2	9.8	11.3	10.5	9.0	

after the crisis. The tables of shock 1 and 2 (not shown) give the same result. It is interesting to look more closely at the absolute error over time. Therefore, in Figure 4 below we show the average absolute error over all twelve PPCs, for month 1 to 36 after the shock occurred.

Figure 4 shows that during the first three months the CL, Bridge model has much higher average absolute errors. This is a general pattern that holds for all NC before BM models (not shown in the figure). Another pattern in all three shocks is that during the first three months after a shock, the SE and BIR model perform best.

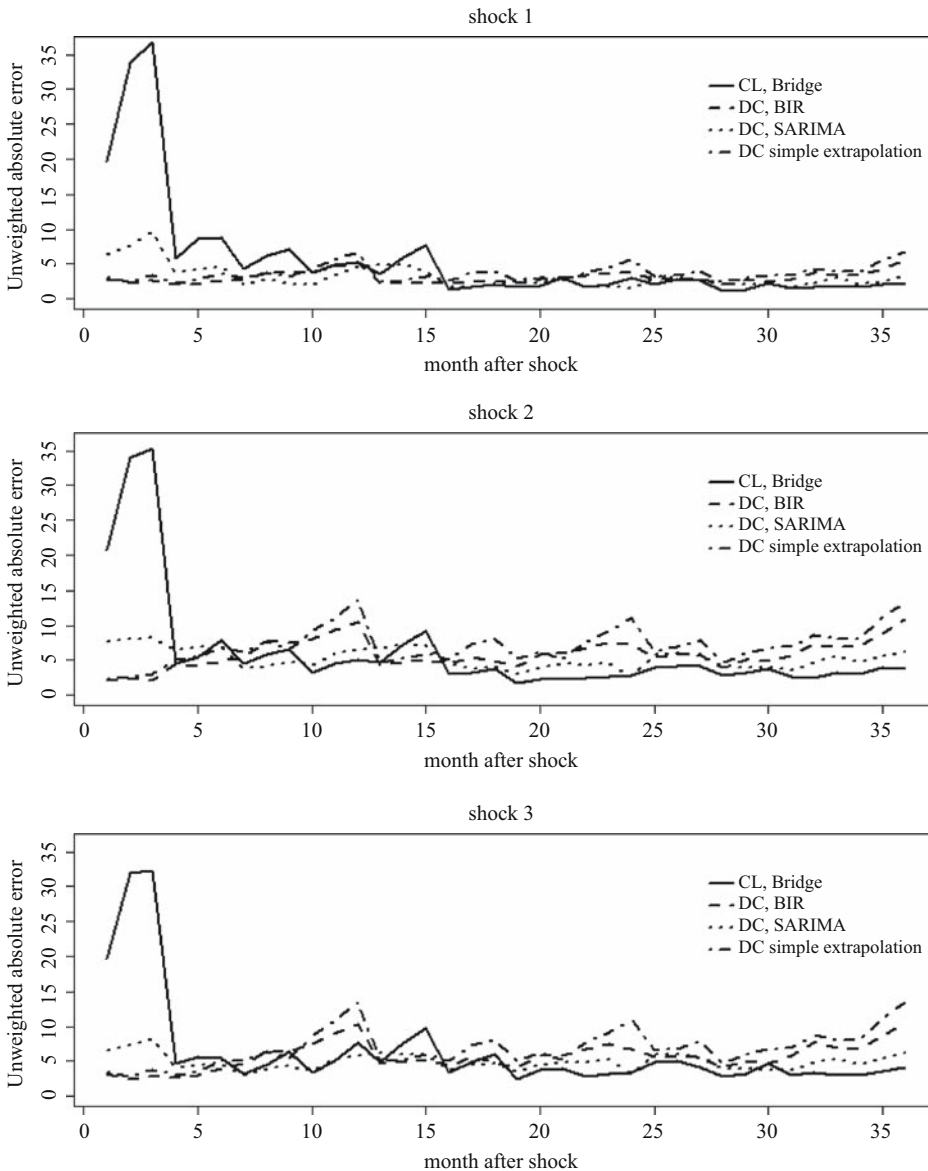


Fig. 4. Average absolute error over all PPCs, 1–36 months after occurrence of shock 1, 2 and 3, for selected models.

4. Conclusion

In this article the estimation of short-term monthly estimates based on a slow but accurate quarterly series and a potentially selective monthly auxiliary series is discussed. There are two problems involved. First, the quarterly series must be temporarily disaggregated (TD), using the monthly auxiliary series. This is done with well-known BM models. These models, Chow-Lin (CL) and Denton-Cholette (DC), transfer the monthly pattern of the monthly auxiliary series onto the quarterly series. Unfortunately, the plausibility of these transfers cannot be evaluated in this application, since the monthly patterns of the businesses that declare VAT on a quarterly frequency remains unknown. The fact, however, that subject matter specialists consider the results as plausible, gives us trust in the results.

A major part of the article concerns the second problem: monthly estimates must be computed before the quarterly figure is available, which means that a nowcast method must be applied. In the article different nowcast methods are compared. In the evaluation of the methods, we distinguish between a stable economic period, where the development of the series is quite stable and predictable, and a period of crisis in which a sudden shock occurs. The financial crisis of 2008 and the COVID-19 pandemic are two examples of such crises, of which the latter is considered in this article. The methods are applied to twelve series that are published by Statistics Netherlands.

It is found that during a stable period most of the methods we consider perform quite well. The so-called Bridge, CL model performs slightly better than the other methods. This method first predicts both the quarterly and the monthly series (using a SARIMA- model) of the current quarter. Then, a CL BM model is applied.

In a period of crisis, the Bridge model is no longer the most accurate model. However, during a shock most models perform worse than during a stable period. Right after a shock, NC after BM models perform better than the NC before BM models. The best method in a period of crisis seems to be the DC, SARIMA model with the monthly auxiliary series as regressor and use this model to nowcast the current month. The reason why DC is preferred in times of crisis, is because an (S)ARIMA model of a series obtained by CL leads to larger dependencies on the history of the series. A simulation study shows that one year after a sudden shock, the CL, Bridge method is again one of the best performing methods.

In this article, for two reasons, only one auxiliary monthly series is used for both BM and nowcasting. The first reason is that each auxiliary monthly series is based on turnover of companies with similar economic activity (i.e., primary publication cells, PPCs). This implies that an auxiliary series that measures the same phenomenon is potentially sufficient, while additional auxiliary monthly series might introduce error. This may hold especially in a period of crisis, where the relation between the additional auxiliary series and the target series might be disturbed leading to model misspecification and biased nowcasts. The second reason is that in the production of timely monthly official figures there is very limited amount of time for model checking and evaluation. From that point of view, relatively simple models that are easy to interpret are preferred above complex models.

In this article 12 PPC time series are considered as test cases, and nowcasts are computed for 44 months in normal times and four months in crisis times. The performance of the different models is quite consistent over these test cases and within periods. This

might indicate that our results can be generalized to other applications but more empirical results to support our findings are of course desired.

An issue with the simulation in this study is to find a benchmark to evaluate the accuracy of the proposed methods. In this study the monthly index series as obtained by BM are used as the benchmark. It is not clear whether this choice favours some methods above others. A simulation that does not favour or handicap particular methods requires a setup where artificial populations are created. This indeed gives more insight in the properties of the different procedures under different conditions. This is left as further research.

The index series which are used in this article are based on turnover sums, where the monthly auxiliary series are based on a selective subpopulation. It is also possible to (partly) correct for this selectivity by weighting, using the available background information about the involved enterprises. In this application, only limited information is available. When there is some information about the self-selection process available, this could be used in the correction process as well.

In this article, we investigated the index series that are published by Statistics Netherlands as short-term statistics. Each index series is based on two turnover series. In a preliminary analysis it is investigated whether the accuracy of the nowcasts could be improved when these underlying turnover series are modelled instead of the index series. It is found that modelling the index series is more promising. See Zult et al. (2020) for further details.

5. References

- Angelini, E., M. Banbura, and G. Rünstler. 2008. *Estimating and Forecasting the Euro Area Monthly National Accounts from a Dynamic Factor Model*. ECB Working Paper No. 953. Available at: <https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp953.pdf> (accessed March 2021).
- Antolín-Díaz, J., T. Drechsel and I. Petrella. 2021. *Advances in Nowcasting Economic Activity: Secular Trends, Large Shocks and New Data*. CEPR Discussion Paper No. DP15926. Available at: <https://ssrn.com/abstract=3805349>.
- Asimakopoulou, S., J. Paredes, and T. Warmedinger. 2013. *Forecasting fiscal time series using mixed frequency data*. ECB Working Paper No. 1550, Available at: <https://www.econstor.eu/bitstream/10419/153983/1/ecbwp1550.pdf> (accessed April 2021).
- Baffigi, A., R. Golinelli, and G. Parigi. 2004. "Bridge models to forecast the euro area GDP." *International Journal of Forecasting* 20: 447–460. DOI: [https://doi.org/10.1016/S0169-2070\(03\)00067-0](https://doi.org/10.1016/S0169-2070(03)00067-0)
- Baker, R., S.J. Blumberg, J.M. Brick, M.P. Couper, M. Courtright, J.M. Dennis, D. Dillman, M.R. Frankel, P. Garland, and R.M. Groves. 2010. "Research synthesis AAPOR report on online panels." *Public Opinion Quarterly* 74: 711–781. DOI: <https://doi.org/10.1093/poq/nfq048>
- Barcellan, R., and D. Buono. 2002. "Temporal Disaggregation Techniques – ECOTRIM Interface (Version 1.01)." User Manual, Eurostat. Available at: <https://ec.europa.eu/eurostat/documents/3859598/9441376/KS-06-18-355-EN.pdf/fce32fc9-966f-4c13-9d20-8ce6ccf079b6> (accessed September 2022).
- Bloem, A.M., R. Dippelsman, and N.O. Maehle. 2001. *Quarterly national accounts manual: Concepts, data sources, and compilation*. Washington, D.C. International

- Monetary Fund. Available at: <https://www.imf.org/external/pubs/ft/qna/2000/textbook/index.htm> (accessed June 2021).
- Box, G., and G. Jenkins. 1989. *Time series analysis: Forecasting and control*. Holden-Day, San Francisco.
- Chow, G., and A. Lin. 1971. "Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series." *The Review of Economics and Statistics* 53(4): 372–375. DOI: <https://doi.org/10.2307/1928739>.
- Dagum, E. B., and P.A. Cholette. 2006. *Benchmarking, temporal distribution, and reconciliation methods for time series* (Volume 186). Part of the book series: Lecture Notes in Statistics. Springer Science & Business Media.
- Daalmans, J.A. 2018. "Special Issue Article: Benchmarking, Temporal Disaggregation, and Reconciliation of Systems of Time Series." *Statistica Neerlandica* 72 (4): 406–420. DOI: <https://doi.org/10.1111/stan.12163>.
- Denton, F.T. 1971. "Adjustment of monthly or quarterly series to annual totals: An Approach based on quadratic minimization." *Journal of the American Statistical Association* 66 (333): 99–102. DOI: <https://doi.org/10.2307/2284856>.
- Doz, C., D. Giannone and L. Reichlin. 2012. "A quasi–maximum likelihood approach for large, approximate dynamic factor models." *Review of economics and statistics*, 94(4): 1014–1024. Available at: <http://www.jstor.org/stable/23355337> (accessed July 2021).
- Durbin, J., and S.J. Koopman. 2012. *Time series analysis by state space methods*, second edition. Oxford: Oxford University Press. Available at: <https://www.researchgate.net/publication/227468262> (accessed September 2022).
- Eurostat. 2008. NACE Rev. 2, *Statistical classification of economic activities in the European Community*. Available at: <https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF> (accessed April 2021).
- Eurostat. 2017. *Handbook on rapid estimates*. Available at: <https://ec.europa.eu/eurostat/documents/3859598/8555708/KS-GQ-17-008-EN-N.pdf/7f40c70d-0a44-4459-b5b3-72894e13ca6d?t=1513758176000> (accessed March 2021).
- Eurostat. 2018. *ESS Guidelines on temporal disaggregation, benchmarking, and reconciliation*. Available at: <https://ec.europa.eu/eurostat/documents/3859598/9441376/KS-06-18-355-EN.pdf/fce32fc9-966f-4c13-9d20-8ce6ccf079b6> (accessed March 2021).
- Fernández, R. 1981. A Methodological Note on the Estimation of Time Series. *The Review of Economics and Statistics* 63(3): 471–476. DOI: <https://doi.org/10.2307/1924371>.
- Ghysels, E., V. Kvedaras, and V. Zemlys. 2016. Mixed Frequency Data Sampling Regression Models: The R Package midasr. *Journal of Statistical Software*, 72(4): 1–35. Available at: <https://doi.org/10.18637/jss.v072.i04>
- Ghysels, E., P. Santa-Clara, and R. Valkanov. 2004. "The MIDAS Touch: Mixed Data Sampling Regression Models." UCLA: Finance. Available at: <https://escholarship.org/uc/item/9mf223rs> (accessed July 2021).
- Ghysels, E., A. Sinko, and R. Valkanov. 2007. MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, Taylor & Francis Journals 26(1): 53–90. DOI: <https://doi.org/10.1080/07474930600972467>.
- Giannone, D., L. Reichlin, and D. Small. 2008. "Nowcasting: The real-time informational content of macroeconomic data." *Journal of Monetary Economics*, 55(4): 665–676. DOI: <https://doi.org/10.1016/j.jmoneco.2008.05.010>.

- Hyndman R.J., and Y. Khandakar. 2008. “Automatic time series forecasting: the forecast package for R.” *Journal of Statistical Software* 27(3): 1–22. DOI: <https://doi.org/10.18637/jss.v027.i03>.
- Litterman, R.B. 1983. “A Random walk, Markov model for the Distribution of Time Series.” *Journal of Business and Economic Statistics* 1(2): 169–173. DOI: <https://doi.org/10.2307/1391858>.
- R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/>. (accessed March 2021).
- Sax, C., and P. Steiner. 2013. “tempdisagg: Methods for Temporal Disaggregation and Interpolation of Time Series.” Available at: <http://CRAN.R-project.org/package=tempdisagg>. (accessed April 2021).
- Sims, C.A. 1980. “Macroeconomics and reality.” *Econometrica* 48(1): 1–48. DOI: <https://doi.org/10.2307/1912017>.
- Steiner, C., and P. Sax. 2013. *Temporal disaggregation of Time Series*. Available at: <https://journal.r-project.org/archive/2013-2/sax-steiner.pdf> (accessed March 2021).
- Stock, J.H., and M.W. Watson. 2001. “Vector autoregressions.” *The journal of economic perspectives* 15(4): 101–115. DOI: <https://doi.org/10.1257/jep.15.4.101>.
- Zult, D., S. Krieg, B. Schouten., P. Ouweland, and J. van den Brakel 2020. *From quarterly to monthly turnover figures using nowcasting*. Statistics Netherlands, Discussion paper. Available at: https://www.cbs.nl/-/media/_pdf/2020/14/nowcasting-fribs-24-maart.pdf.

Received March 2020

Revised February 2021

Accepted September 2022