Journal of Official Statistics
vol. 35, 3 (Sep 2019)

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 487–508, http://dx.doi.org/10.2478/JOS-2019-0021

Probing for Informal Work Activity

Katharine G. Abraham[1] *and Ashley Amaya*[2]

The Current Population Survey (CPS) is the source of official US labor force statistics. The wording of the CPS employment questions may not always cue respondents to include informal work in their responses, especially when providing proxy reports about other household members. In a survey experiment conducted using a sample of Amazon Mechanical Turk respondents, additional probing identified a substantial amount of informal work activity not captured by the CPS employment questions, both among those with no employment and among those categorized as employed based on answers to the CPS questions. Among respondents providing a proxy report for another household member, the share identifying additional work was systematically greater among those receiving a detailed probe that offered examples of types of informal work than among those receiving a simpler global probe. Similar differences between the effects of the detailed and the global probe were observed when respondents answered for themselves only among those who had already reported multiple jobs. The findings suggest that additional probing could improve estimates of employment and multiple job holding in the CPS and other household surveys, but that the nature of the probe is likely to be important.

1. Introduction

Information on employment and hours of work is critical to policy makers and other decision makers for assessing the state of the labor market and the economy more broadly. In the United States, much of this information comes from the Current Population Survey (CPS), a monthly survey of approximately 60,000 households carried out by the U.S. Census Bureau on behalf of the Bureau of Labor Statistics (BLS). In the CPS, an individual is considered to be employed if he or she "did any work at all for pay or profit during the survey reference week. This includes all part-time and temporary work, as well as regular full-time, year-round employment" (Bureau of Labor Statistics, undated).

One potential concern about the CPS data is that the wording of the survey's employment questions may not adequately cue respondents to report work activity outside of a conventional job or business. The CPS employment questions are asked for each household member age 16 and older. The initial employment question asks whether the household member did any work during the survey reference week for 'pay' (or, if

[1] University of Maryland, Joint Program in Survey Methodology and Department of Economics, College Park, Maryland, 20742, U.S.A. Email: kabraham@umd.edu.
[2] RTI International, Survey Research Division, 701 13 St NW, #750, Washington D.C., U.S.A. Email: aamaya@rti.org.
Acknowledgments: The authors are grateful to Frederick G. Conrad, Monica Dashen, Susan N. Houseman, Frauke Kreuter, and James R. Spletzer for helpful comments and suggestions on an earlier draft of the paper. Support for collection of the data analyzed in the paper was provided by the U.S. Census Bureau under Contract YA132312CN0037 with the University of Maryland, which provided support for the Joint Program in Survey Methodology (JPSM) generally and the 2016 JPSM Survey Practicum specifically.

applicable, for 'pay or profit'). Later questions in the sequence ask about having more than one 'job'(or, if applicable, more than one 'job or business'). The "pay or profit" and "job or business" formulations are used in cases in which the CPS respondent has indicated that someone in the household has a business. It is not clear, however, that respondents necessarily will think of money earned through informal work activity as either 'pay' or 'profit' or consider such activity to be a 'job' or 'business.' The consequence may be that the reporting of informal work activity is incomplete.

The use of proxy respondents is a second potential challenge to accurate reporting. Although CPS interviewers attempt to collect employment information from each household member age 16 years or older, time and availability constraints often lead to the use of a proxy reporter, a household member who answers the survey questions on behalf of other household members. Responses for roughly half of CPS sample persons are collected from proxy reporters (U.S. Census Bureau 2006). Even if the respondent understands that all work to earn money should be reported, irregular or casual work performed by other household members may be less salient to the proxy respondent than similar work performed by the respondent herself and thus less likely to be reported. In some cases, the respondent simply may not know about informal work performed during the survey reference period by other household members.

This article seeks to understand the nature of potential biases in the reporting of work activity in the CPS and similar surveys. Our first research question can be stated:

1) Is there informal work for pay or profit done during the survey reference week that is not captured by the standard Current Population Survey (CPS) employment questions?

To answer this research question, we examine whether asking questions focused specifically on informal work as a follow-up to the standard CPS employment questions identifies additional work activity. We also are interested in whether different ways of asking such added questions are more or less effective and in whether this varies according to whether a respondent is reporting for themselves versus another household member or, in the latter case, according to the relationship of the respondent to the other household member:

2) Does the way in which questions that probe for informal work are asked affect the number of additional jobs identified?
3) Does the relative effectiveness of different ways of probing for informal work vary by whether the survey respondent is answering for herself (self-report) or for another household member (proxy report)? If answering for another household member, does the relative effectiveness of different ways of probing vary by the closeness of the survey respondent to the other household member?

Finally, we are interested in the potential effects of under-reporting of informal work during the survey reference week on key labor force measures:

4) How does any under-reporting of informal work in answering the standard CPS questions affect estimates of the employment rate (the share of the sample that is categorized as employed) and the multiple job holding rate (the share of employed persons in the sample who hold more than one job)?

2. Background

To understand how question wording might affect reports of work activity, we must first identify how respondents formulate responses. The most common model of the response process suggests four steps: (1) understanding the question, (2) recall, (3) inference and estimation, and (4) mapping the answer onto the response format and editing the response (Sudman et al. 1996; Tourangeau et al. 2000). We limit our discussion to the first two of these steps because they are the most relevant to our research questions and experimental conditions.

Before a respondent can provide a response to a survey question, she must first understand what information is being requested. Even questions that appear to be clear can be interpreted in different ways. For example, in one study, respondents were asked: "Have you smoked at least 100 cigarettes in your entire life?" Respondents disagreed on whether to include puffs where they did not inhale, whether to count cigarettes they had only partially smoked, and what constituted a cigarette (Schober et al. 2018). This sort of disconnect is due, in part, to the difference between literal interpretation and pragmatic interpretation. Individuals want to be responsive to what they think the researcher wants to know (pragmatic interpretation), regardless of exactly what was asked (literal interpretation) (Schwarz 1999). A respondent answering the CPS employment questions might decide that informal or irregular work activity that occurred during the survey reference week is not part of what the interviewer is asking about. For example, someone who performed as a magician at weekend children's parties or maintained a blog that generated ad revenue might think of this activity as a 'hobby' and not as 'work', and fail to report it when asked the standard CPS questions. While any misconstruing of a question on the part of a respondent is problematic for achieving accurate estimates, there is no reason to think that the severity of this problem should differ between self and proxy reports. In either case, a probe asking specifically about informal work may change the respondent's understanding of what they should be reporting and thus uncover previously unreported work activity.

In the second step of the response process, the respondent must recall information relevant to formulating a response. They will use cues such as 'work', 'pay or profit', 'job or business,' and the reference week from the question wording and survey context to search their memory. Poor cues will increase the chance of retrieval failure (Tourangeau 2000). Because richer information is stored about the self than about others (Kuiper and Rogers 1979), strong cues may be especially important for proxy reporting. To the extent that individuals store more information about events that involve them directly, even a weak cue may spur retrieval of a given event, whereas stronger cues may be required to activate the retrieval of information about other individuals' activities. In the context of collecting information about informal work activity, we would expect a question that provides specific examples of the types of work that may have been performed (e.g., doing yard work or driving for a ridesharing service) to activate a respondent's memory more successfully than a question that asks in more general terms about informal work activity. We would expect this to be more the case for proxy reports than for self-reports and, among proxy reports, perhaps more so for individuals with whom the respondent has weaker ties (e.g., a roommate or other unrelated household member) as opposed to those with whom the respondent has stronger ties (e.g., a spouse).

The use of dispositional knowledge also may lead to failure in the recall process. Individuals may have two distinct types of knowledge about others: situational and dispositional (Schwarz and Wellens 1997). Situational knowledge includes details about specific events whereas dispositional knowledge is information that can be inferred about an individual based on her typical behavior. In a study of consumer expenditures, for example, respondents used a combination of situational and dispositional knowledge to report their own spending behavior but relied primarily on dispositional knowledge when reporting on behalf of their spouse (Dashen 2000). When individuals use dispositional knowledge to answer questions about employment, they may be less likely to report sporadic or casual work activity because it is not a 'usual' behavior (Sudman et al. 1996; Schwarz and Wellens 1997). This reasoning suggests that, to the extent that probing encourages the respondent to tap into situational knowledge, it may be differentially effective for uncovering added informal work among proxy reports compared to self-reports.

Individuals also may fail to retrieve the necessary information if it was not encoded in the first place. While this should be relatively rare for self-reports of employment, it could be more of an issue for proxy reporting. If another household member did work during the reference week but did not tell the respondent, the respondent would not know to report it. More generally, it may be difficult for a proxy respondent to estimate the extent of another household member's participation in an irregular behavior over a particular interval of time (Phillips et al. 2006). The closeness of proxy reporters to the subject of their reporting has been found to be correlated with the accuracy of the proxy report, perhaps because individuals who are closer to one another are more likely to share information about their activities (Bower and Gilligan 1979; Phillips et al. 2006). As an example, Kojetin and Miller (1993) found stronger agreement between spouses' reports about their partners' spending and the partners' own reports than between parents' reports about their children's spending and the children's own reports. In general, making spending decisions jointly with another household member, discussing spending with the other household member, or observing items that the other household member may have purchased all contributed to stronger agreement between reports made by the proxy respondent and those made by the person doing the spending. If lack of encoding is problematic in the context of reporting about work activity, we might expect the effects of probing to differ depending on the relationship between the respondent and the person about whom she is reporting. In this case, probing could be more effective when used to elicit information about individuals, such as a spouse who are closer to the respondent, since the information is more likely to have been encoded in the first place, and less effective when used to elicit information about other household members.

Three primary methods have been tested to improve accuracy of reports about behavior. First, definitions have been used to clarify questions, thus improving comprehension. In an experiment described by Fowler (1992), definitions intended to ensure that respondents' interpretations of a set of questions related to health behaviors were consistent with the researcher's intent were provided to half of the participants but not to the others. While no information was collected on respondent interpretation, the distribution of responses differed significantly between the two conditions. Inclusion of definitions or instructions can be more important in complex situations. For example, in one experiment, subjects asked a series of questions about employment status, housing, and household purchases

based on complex fictional scenarios answered accurately about 87% of the time when interviewers had the flexibility to clarify definitions, but just 28% of the time when no definitions were provided. Answers to the same questions based on simpler scenarios were accurate 97% or more of the time regardless of whether the interviewer had the opportunity to provide clarification (Schober and Conrad 1997).

Second, adding examples to questions offers additional cues that the respondent may be able to use to recall more complete information. The choice of examples provided may affect the responses that are given. In a study of food consumption, Tourangeau et al. (2014) varied the examples for different food categories by the frequency of consumption (e.g., bread versus barley for grains) and by whether the item would be considered a typical example (e.g., milk versus sour cream for dairy). Overall, individuals reported more consumption when any examples were provided. Further, when asked to list what they ate, they were more likely to mention consumption of the example items. This suggests that individuals retrieve enough information to make a judgement but do not try to recall everything.

Finally, and perhaps most relevant for this study, researchers have tested the use of decomposed questions to offer additional cues and enhance recall. Menon (1997) conducted a diary experiment in which individuals were asked either open-ended questions about the number of times they had done each of six behaviors or a set of questions that explicitly cued the respondent to think about the different circumstances under which each of the same things might have been done. The second, decomposed condition improved the accuracy of recall for the three irregular behaviors studied (making unplanned stops to talk to friends, snacking, and drinking from a water fountain), but not for the regular behaviors (washing hair, having dinner, and attending class).

Other research has identified circumstances under which decomposed questions may perform less well. In a survey experiment reported by Belli et al. (2000), respondents either were asked a simple question about the total number of local or long-distance phone calls they had made during a specified period or were asked decomposed questions about the same behavior that cued the respondent to think separately about calls at different times or to different destinations. Subjects who received the decomposed question had a greater tendency to over-report the number of phone calls they had made than subjects who received the simple question. Members of the study population in the Belli et al. (2000) study made a sufficiently large number of phone calls that they most likely used an estimation strategy to formulate their answers rather than enumerating each call individually (Blair and Burton 1987). We would expect respondents reporting on informal work activities during the prior week to enumerate rather than estimate, meaning that the findings reported by Menon (1997) are likely to be more applicable to our context than the findings reported by Belli et al. (2000).

Additional research has suggested that probing or decomposed questions also may result in overreporting due to forward telescoping, that is, the inclusion of activities that in fact had occurred prior to the specified reference period (Sudman and Bradburn 1973). Forward telescoping is more likely to occur when events are highly salient. Events that are less salient are more likely to suffer from backward telescoping, that is, the exclusion of events that occurred within the reference period because individuals think they occurred longer ago. To the extent that work identified through the additional questions is work that is less salient, the results will be more likely to suffer from backward than forward

telescoping, trending the additional work identified toward zero (Brown et al. 1985). Moreover, telescoping (of any kind) is more likely to occur the further back the period for which the respondent is asked to recall (Martin 2006). Given that our survey asks about events that occurred over the most recent calendar week, we would not expect telescoping to be a large problem in our context. On balance, the existing literature leads us to expect that adding questions to the CPS questionnaire to identify previously unreported work should improve the accuracy of the information collected.

The material importance of any potential underreporting of informal work activity for our understanding of the labor market will depend in part on the prevalence of such activity in the CPS target population. This is something that several recent surveys have attempted to measure. Robles and McGee (2016) analyzed data from the Enterprising and Informal Work Activities (EIWA) survey fielded by the Federal Reserve Board in October and November of 2015. In their sample, during the six months prior to the survey, 36% of the adult population participated in informal work that involved either selling or renting property or providing services. The estimate from the 2016 Survey of Household Economics and Decision making (SHED), which included similar questions, is that 28% of adults earned money from informal work outside of a main job during the month prior to the survey (Board of Governors of the Federal Reserve System 2017). The two waves of the Survey of Informal Work Participation (SIWP) carried out during 2015 asked whether respondents were "currently engaged" in informal paid activity or side jobs, exclusive of selling property, renting property or responding to surveys (Bracha and Burke 2019). This was the case for 21% of adults age 21 and older categorized as employed, 25% of those categorized as unemployed, and 12% of those categorized as out of the labor force based on the CPS employment questions. An important caveat is that all three of these estimates are based on online panel surveys. One might be concerned that engagement in other sorts of informal work is higher among those willing to participate in an online panel than among the general population. While perhaps the case, at least in the SHED, even after excluding all informal work done by anyone who reported any online work, the estimated prevalence of informal work activity remained substantial (Abraham and Houseman 2018).

Existing research provides some insights regarding the set of questions about the measurement of informal work that motivate our research. In a study based on data collected during the early 1990s, Martin and Polivka (1995) explored the effect of probing for informal work activity on measured employment rates. In one portion of their study, household survey respondents were asked questions very similar to the current CPS employment sequence. Then, in cases in which there was at least one adult member of the respondent's household with no reported employment, a question about informal work activity was asked regarding the first such person listed on the household roster. Additional work activity identified through this probing raised the estimated employment rate by 2.3 percentage points, with proportionally larger effects for household members under age 20 and age 65 and older. Martin and Polivka did not attempt to learn about underreporting of informal work as a secondary work activity (i.e., about multiple job holding) or about differences in the effects of probing for self versus proxy reporters, nor did they experiment with alternative wordings for their probe.

More recently, analyzing data from the two waves of the SIWP fielded in 2015, Bracha and Burke (2019) estimated that accounting for informal work activity identified through

probing would raise the overall employment rate by 4.5 percentage points above that estimated based on responses to the CPS employment questions and raise the multiple job holding rate by more than 11 percentage points. In contrast to Martin and Polivka (1995), Bracha and Burke asked first about informal work and then administered the CPS employment questions. This question ordering could have affected the responses to the CPS questions and thus their conclusions. The wording of their question about informal work—which asks whether a respondent is "currently engaged" in such work rather than about whether the respondent did any such work during the survey reference period – is also potentially problematic. Bracha and Burke (2019) do not provide evidence on possible differences in the reporting behavior of self-reporters versus proxy reporters, nor was their study designed to learn about the effectiveness of different ways of asking about respondents' participation in informal work.

Another relevant study is Katz and Krueger (2019a), which reported on a 2015 survey of respondents recruited via the Mechanical Turk website, Amazon's crowdsourcing platform, that was designed primarily to learn whether people answering the CPS employment questions under-report multiple job holding. They first asked the CPS employment questions and then asked "Did you work on any gigs, HITS or other small paid jobs last week that you did not include in your answer to the previous question?" Taking the additional small jobs mentioned by respondents into account raised the share of workers in the Katz and Krueger sample who were multiple job holders from 39% to 77%. Similar to the other studies we have discussed, the Katz and Krueger study was not designed to shed light on possible differences between the reporting behavior of self-reporters versus proxy reporters nor to assess the relative effectiveness of different ways of probing to learn about informal work activity.

Finally, in a novel analysis, Allard and Polivka (2018) used data from the American Time Use Survey (ATUS) to gauge the effects of accounting for informal work on the employment and multiple job holding rates. The ATUS, which uses the CPS as a sampling frame, includes CPS-style questions about individuals' labor force status and also collects information on each respondent's allocation of time during one 24-hour period. Allard and Polivka focused on time devoted to labor-intensive income-generating activities such as hobbies, crafts, food, performances or services that are not part of a job or business. They estimated that, in the ATUS over the 2012–2016 period, accounting for such activities would have raised the employment rate by between 0.4% and 3.0% and raised the multiple job-holding rate by between 3.0% and 20.7%. In both cases, the range reflects uncertainty about the extent to which average daily participation in such activities reflects the same people engaging in the activity on multiple days as opposed to different people engaging in the activity on different days. These estimates suggest that the standard CPS questions miss relatively little informal work activity, but depend both on the definition of income-generating activities adopted and, perhaps more importantly, on the ATUS doing a good job of capturing time devoted to those activities.

3. Methods

To answer our research questions, we use data from the 2016 Joint Program in Survey Methodology (JPSM) practicum project. For this project, a task visible only to US

residents was posted to the Mechanical Turk website, asking for individuals to complete a survey about employment referred to in the posting as the Current Employment Survey. Individuals who clicked on the task were told that they would receive USD 2.50 for completion of a survey about the employment status of themselves and other household members. A total of 4,991 people completed the survey on August 16 and 17, 2016, taking an average of 13.55 minutes to answer the questions asked. Given the non-probabilistic nature of the survey, response rates were not calculated. We excluded 52 cases due to item non response, and analysis was conducted on the remaining 4,939 completed interviews.

The first section of the survey collected information on the characteristics of all members of a respondent's household. It included questions concerning age, sex, education, race and ethnicity, marital status and relationship to the household respondent (opposite sex spouse, opposite sex unmarried partner, same-sex spouse, same-sex unmarried partner, child, grandchild, parent, brother/sister, other relative, foster child, housemate/roommate, roomer/boarder or other non-relative). The second section of the survey asked questions to identify each household member's employment status; for those who were employed, whether they held more than one job; and, as applicable, the hours worked on the main and other jobs. With the exception of some experimental questions concerning sexual orientation and gender identity, all of the questions about household members' characteristics and work activity were taken directly from the CPS questionnaire. The use of the CPS employment questions on the JPSM practicum survey means that the responses can be used to construct CPS-like measures of both employment and multiple job holding during the survey reference week ("last week," defined as the most recent completed week beginning on a Sunday and ending on a Saturday).

For the respondent (in single adult households) or for one randomly-selected member of the household (in multiple adult households), the CPS employment questions were followed by additional questions probing for activity to earn money outside of a regular job. This is the sample of people on which the analysis reported here is based. As can be seen in Table 1, the analysis subjects are younger and considerably more educated than the population as a whole.

The specific questions asked about informal work activity were varied experimentally. In one treatment condition, randomly assigned to half the cases, respondents were asked a global yes/no question about whether any such activity had occurred during the survey reference week (the global question). If no work activity had been reported for the subject household member in response to the standard CPS questions, the global question was:

> Sometimes people who don't have a job do other things to earn money. Did [you/[NAME]] do other things to earn money last week?

For those with work activity reported in answer to the CPS questions, the global question was:

> Sometimes, in addition to working at a job [or business] where there is a definite arrangement for regular work on a continuing basis, people do other things to earn money. Outside of a job [or business], did [you/[NAME]] do other things to earn money last week?

Table 1. Characteristics of Analysis Sample versus American Community Survey Estimates (Percent Distributions).

	Respondent	Other household members	ACS (2016)[3]
Age			
18–24/16–24[1]	11.7	18.7	12.8
25–34	45.8	31.7	17.7
35–44	23.9	17.5	16.6
45–54	11.1	14.1	17.7
55–64	5.7	11.4	16.4
65 and over	1.7	6.6	18.9
Female[2]	50.6	47.3	51.4
Education			
Less than high school	0.3	6.7	12.6
High school	8.7	21.3	27.7
Some college or associate degree	36.2	33.6	31.0
Bachelors degree or higher	54.7	38.3	28.7
Race/ethnicity			
Hispanic	7.3	10.7	16.0
Non-hispanic white	73.8	70.9	65.5
Non-hispanic african american	7.0	6.9	12.3
Non-hispanic other race	8.0	8.9	4.8
Non-hispanic multiracial	3.9	2.6	1.5
Sample size	2,704	2,235	–

[1]All survey respondents were age 18 or older, but respondents were asked to report for other household members age 16 and older. The survey sample includes N = 93 other household members age 16 or 17. The ACS numbers show the age distribution of the population age 18 and older.
[2]The survey sample includes N = 22 respondents and N = 19 other household members reported as transgender or not identifying as either male or female, or for whom no report on gender identity was provided. They are included in the denominator when calculating the percent female in our sample.
[3]All sample distributions are significantly different from the corresponding ACS distributions at $p < 0.001$.

In these questions, as applicable, the text filled based on the person selected (e.g., if the respondent is answering about another household member, NAME refers to that person's name) and whether or not the respondent had reported work by the individual in a family business.

In the second treatment condition, survey respondents were asked essentially the same question, but with potential informal work activity decomposed into seven different categories (the detailed question). The seven categories of work activity outside of a regular job that a respondent might report were (1) provided services to other people, (2) provided services to a self-employed individual or business, (3) performed as an actor, musician or entertainer, (4) drove for a ridesharing service, (5) assisted with medical, marketing, or other research, (6) posted videos, blog posts, or other content online, or (7) did other informal work or side job. Examples were provided for all but the 'other' category.

For anyone categorized as CPS employed for whom informal work was reported, the respondent was asked to indicate whether the informal work mentioned in response to additional probing had been included in the CPS job count. Both among those who

received the global probe and among those who received the detailed probe, only about half of the informal work mentioned when we probed had been included when answering the CPS employment questions. Respondents also were asked to report the number of hours devoted to the informal work reported in response to the probing question. Appendix A (Supplental material) provides information on the age, sex, education, race, and ethnicity of self-reports and proxy reports by assignment to the global versus the detailed question treatment. The question treatment groups are well balanced with respect to these characteristics. The only statistically significant differences between the characteristics of the global and detailed question treatment groups are among other household members, with those assigned the global question somewhat less likely than those assigned the detailed question to have some college or an Associate degree (30.6% versus 36.7%) and somewhat more likely to have a Bachelors degree or higher (40.5% versus 36.1%).

To answer our first research question – whether there is informal work for pay or profit done during the survey reference week that the CPS employment questions do not capture—we look at the proportion of individuals for whom additional probing identified work that was not included in the answers to the CPS questions. We use a one-tailed one sample *t*-test to determine whether this proportion is significantly greater than zero.

To address our second research question on whether the method used to probe for informal work affects the answers obtained, we compare the share of people for whom additional work is identified by the global versus the detailed question. We use a two-tailed two-sample *t*-test to determine whether the two probes–the global question and the detailed question–elicit different amounts of additional work activity. To address our third research question, we carry out these same comparisons separately for respondents reporting for themselves (self-reports) versus respondents reporting for other household members (proxy reports) and then, within the latter group, separately for respondents reporting about a spouse or unmarried partner (which we will refer to simply as a spouse) versus respondents reporting about another household member.

We are most interested in the effects that probing for informal work activity has on the estimated employment rate (the percent of people in the sample who were employed) and the multiple job holding rate (the percent of employed persons with two or more jobs). Additional work activity identified among those initially classified as not employed could raise the employment rate; additional work activity identified among those with a single CPS job could raise the multiple job holding rate. In principle, the identification of multiple jobs for someone initially classified as not employed also could raise the multiple job holding rate. For the purpose of comparing the effects of the detailed and global questions on the multiple job holding rate, however, we do not want to allow for an outcome that is possible for those receiving the detailed question but not for those receiving the global question. In contrast to the detailed question, the global question allows us to determine only that an individual had done some work that was not initially reported, not whether they had more than one unreported job.

We look first at how asking one or the other of the probing questions (either the global question or the detailed question) affects the statistics of interest (the employment rate and the multiple job holding rate). We use one-tailed paired *t*-tests to determine whether these effects pass the threshold of statistical significance. The differences in the effects of interest then are compared across the two treatments–the detailed question treatment

versus the global question treatment–using a two-tailed two-sample *t*-test. These analyses related to our final research question are carried out first for the full sample and then separately by household member status (self-report or proxy report), with the latter also broken out according to whether the report is for a spouse or other household member.

All analyses are unweighted. The implications of the sample design and lack of weights are considered in the concluding discussion.

4. Results

Our first research question asks whether individuals engage in informal work during the reference week that is not captured by the standard CPS employment questions. We begin by looking at the patterns of employment for the sample as a whole. As shown in Table 2, based on their employment status as determined using the responses to the standard CPS questions, 16.6% of sample members are categorized as not employed, 63.6% as employed with one job, and 19.8% as employed with more than one job. By comparison, in CPS data for August 2016, 38.7% of individuals 18 and older were not employed, 58.3% were employed with one job, and 3.0% were employed with two or more jobs.

When respondents are prompted with follow-up questions about work activity outside of a regular job, additional work not reflected in the answers to the standard CPS employment questions is reported for 21.9% of the sample. Additional work is identified for members of all three employment-status groups–among those the CPS questions identified as not employed, as employed with a single job, and as employed with two or more jobs.

Because our sample was recruited through Amazon's Mechanical Turk, we know that all of our respondents have been involved in gig work at least to some extent. This means that the incidence of additional work we uncovered by probing likely is higher than in the general population. We do not have good information on the types of informal work done by those who received the global probe, but we do have that information for those who received the detailed probe. About a third of those receiving the detailed probe who did any added informal work reported work in the research category, which is where

Table 2. Additional Work Activity Identified by Probing, Full Sample.

	Sample size	Employment status based on CPS questions, percent of full sample	Additional work activity identified by probing, percent of full sample[1]	Additional work activity identified by probing, percent of row category
Total	4,939	100.0	21.9	21.9
CPS not employed	820	16.6	3.9	23.5
CPS employed, 1 job	3,142	63.6	14.8	23.3
CPS employed, 2 plus jobs	977	19.8	3.1	15.9

[1]All reported values for percent in full sample with additional work activity identified by probing significantly different from zero at $p < 0.001$.

Mechanical Turk activities should be listed. Looking across the remaining categories, among those receiving the detailed probe for whom we identified added work, 17% performed services for others, 12% performed services for a business or self-employed person, 7% earned money by posting content online, 3% drove for a ride-sharing service, 3% performed as an entertainer, and 31% did other types of informal work not captured in the more specific categories. These numbers add up to slightly more than 100 percent because there were some people who reported more than one type of added work.

As a sensitivity check, using the portion of our sample that received the detailed probe, we reran the tabulations reported in Table 2 excluding all additional informal research work. Without this exclusion, 25.8% of those receiving the detailed probe reported additional work activity; excluding research work, this share was smaller but remained substantial at 19.7%. As shown in Appendix Table B1 (Supplemental material), even with research work excluded, probing identified substantial added work activity in all three employment status groups as determined based on the answers to the CPS employment questions.

Another natural question to ask about the added work activity identified through probing is whether it involved more than a minimal amount of individuals' time. We collected information on hours for informal work identified through probing both for those receiving the global probe and for those receiving the detailed probe. Among those responding to the global probe, after asking the hours question, we then asked whether any reported informal work activity had been included when answering the CPS employment questions. Some subjects receiving the global probe could have done more than one type of informal work during the survey reference week; if the answer to the question was no, we assume that none of it had been reported. For those responding to the detailed probe, we asked separately about hours and their inclusion in answering the CPS employment questions for each type of reported informal activity. As a check on whether our conclusions would have been different had we excluded informal work activity identified through probing that involved only a minimal amount of time, we recomputed the numbers reported in Table 2 but counting added informal work only for those with at least four hours of such work identified through probing. As can be seen in Appendix Table C1 (Supplemental material), the share of respondents with added work is about 40 percent lower – 13.0% rather than 21.9% – but the general trends in the estimates are otherwise unaffected.

Among the full set of people reporting additional work during the survey reference week after probing, including those with very low hours, some 17.6% said that they spent an estimated 15 or more hours on that additional activity (15.2% for those receiving the global probe and 19.3% for those receiving the detailed probe). Added work activity during the reference week identified through probing occupied an average of 8.2 hours during the survey reference week (7.0 hours for those receiving the global probe and 9.1 hours for those receiving the detailed probe), roughly equivalent to a full normal work day. Those with no CPS employment for whom unreported work activity was identified by probing are somewhat more likely than those with one or more CPS jobs to have spent 15 or more hours on that activity during the reference week (24.7% versus 16.0% overall, 17.1% versus 14.9% for those the receiving the global probe, and 29.2% versus 16.9% for those receiving the detailed probe). Among those for whom added work activity was

identified, the group with no CPS employment also spent more hours than those with one or more CPS jobs (9.9 versus 7.9 hours overall, attributable entirely to the difference of 11.8 versus 8.5 hours for those receiving the detailed probe).

Our second research question asks whether the form of the follow-up question about informal work affects the number of people for whom additional work activity is identified. The first two rows of Table 3 report estimates of the distribution of the sample by CPS employment status and the distribution of additional employment identified by probing across the three employment status groups. Here, these estimates are shown separately for the cases receiving the global prompt and those receiving the detailed prompt. As anticipated given that the assignment to the global versus the detailed probe was random, the shares of the sample cases in each of the three CPS employment status groups do not differ significantly between the two treatments. The share of cases for which added employment was identified through probing, however, is significantly greater under the detailed question treatment than under the global question treatment (25.8% versus 18.0%, a statistically significant difference of 7.8 percentage points). This overall difference is spread across individuals with no CPS employment, one CPS job, and more than one CPS job; in each of the three groups, the detailed question identifies significantly more added employment than does the global question.

The third research question we posed was whether the effects of prompts to uncover work activity outside of a regular job differ depending on whether they apply to the individual herself (self-report) or to another household member (proxy report) and, in the latter case, whether the effects differ according to the relationship between the respondent and the other household member. The next two panels of Table 3 report estimates separately for the self-report and proxy report cases in our sample. The prevalence of work activity reported in response to the CPS questions is much higher for the people for whom we obtained self-reports than for the people for whom we obtained proxy reports. Those in the self-report group are much less likely to have no CPS employment, equally likely to have a single CPS job, and much more likely to have two or more CPS jobs. Consistent with the random assignment of respondents to treatments, within each of these two groups (self-reports and proxy reports), there are no significant differences in the prevalence of work activity elicited by the standard CPS questions between those receiving the global prompt and those receiving the detailed prompt.

The self-report cases in our sample differ from those for whom we have proxy reports not only in their level of work activity as captured by the CPS questions but potentially also with respect to the prevalence and nature of any work activity not captured by those questions. Differences in the amount of additional work activity identified by prompting for the self-report cases versus the proxy report cases could be due to differences in how people report about themselves as compared to how they report about others. They also could be due, however, to real differences in the labor force activity of the self-reports versus the proxy reports. Given that respondents were assigned randomly to be asked the detailed question versus the global question, however, we can attribute differences across question treatments within either the self-report or the proxy report group to the type of probe each treatment group received.

Asking the detailed question rather than the global question raises the share of proxy report cases for which additional work activity is identified by 10.2 percentage points,

Table 3. Additional Work Activity Identified by Probing, Global versus Detailed Probe.

	Sample size	Employment status based on CPS questions (percent of sample)				Additional work activity identified by probing (percent of sample)			
		Total	Not employed	1 Job	2 + Jobs	Total	CPS, Not employed	CPS, 1 Job	CPS, 2 + Jobs
Full sample									
Global prompt	2,492	100.0	16.6	63.5	19.9	18.0	2.9	13.3	1.8
Detailed prompt	2,447	100.0	16.6	63.8	19.7	25.8	4.9	16.4	4.5
Detailed minus global	–	–	0.0	0.3	−0.2	7.8	2.0	3.0	2.8
(*p*-value)			(0.980)	(0.845)	(0.828)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Self reports									
Global prompt	1,364	100.0	5.4	64.4	30.3	27.9	2.3	22.7	3.0
Detailed prompt	1,340	100.0	5.3	64.6	30.2	33.7	3.4	23.4	6.9
Detailed minus global	–	–	−0.1	0.2	−0.1	5.7	1.2	0.7	3.9
(*p*-value)			(0.862)	(0.921)	(0.965)	(0.001)	(0.070)	(0.664)	(<0.001)
Proxy reports									
Global prompt	1,128	100.0	30.2	62.4	7.4	6.0	3.7	2.0	0.3
Detailed prompt	1,107	100.0	30.3	62.8	7.0	16.3	6.7	7.9	1.7
Detailed minus global	–	–	0.0	0.4	−0.4	10.2	3.0	5.8	1.5
(*p*-value)			(0.987)	(0.856)	(0.386)	(<0.001)	(0.002)	(<0.001)	(0.001)
Spouse									
Global prompt	583	100.0	17.3	73.2	9.4	6.2	2.7	2.9	0.5
Detailed prompt	542	100.0	18.5	71.2	10.3	13.8	3.1	8.3	2.4
Detailed minus global	–	–	1.1	−2.0	0.9	7.7	0.4	5.4	1.9
(*p*-value)			(0.623)	(0.449)	(0.614)	(<0.001)	(0.697)	(<0.001)	(<0.001)
Other household member									
Global prompt	545	100.0	44.0	50.8	5.1	5.9	4.8	1.1	0.0
Detailed prompt	565	100.0	41.6	54.7	3.7	18.6	10.1	7.4	1.1
Detailed minus global	–	–	−2.4	3.9	−1.4	12.7	5.3	6.3	1.1
(*p*-value)			(0.411)	(0.198)	(0.250)	(<0.001)	(0.001)	(<0.001)	(0.014)

from 6.0% of cases with added work activity using the global prompt to 16.3% of cases using the detailed prompt. In contrast, the difference for the self-report cases is just 5.7 percentage points, with 27.9% reporting added work activity under the global prompt versus 33.7% under the detailed prompt. Putting these results somewhat differently, the number of proxy report cases with additional work identified by probing increases by 172% when the detailed question is asked instead of the global question, compared to an increase of just 21% for the self-report cases. Asking the detailed rather than the global question also has a larger effect on the number of hours devoted to additional work for proxy report cases for whom additional work is identified (11.1 hours versus 6.5 hours, a 4.6 hour difference) than for self-report cases with additional work (8.3 hours versus 7.1 hours, a 1.2 hour difference). Among the proxy reports, there are significant differences in the amount of additional work activity identified by the detailed prompt versus the global prompt for all three employment status groups – those without CPS employment, those with one CPS job, and those with more than one CPS job. Among the self-report cases, however, the only statistically significant difference arises for the subgroup who already had reported more than one job in response to the standard CPS questions.

The bottom two panels of Table 3 further break out how asking the global versus the detailed question affects the additional work activity reported when a proxy is answering for a spouse or unmarried partner (referred to for convenience as a spouse) versus some other household member. The rationale for making this comparison is that we expect a respondent generally to be closer to her spouse than to other household members and to communicate more with her spouse about daily activities. If this is correct, we might expect the amount of additional work activity identified by the global compared to the detailed questions be more similar when the proxy subject is a spouse than when the proxy subject is some other household member.

Among reports for spouses, the global and the detailed questions perform very similarly with respect to identifying previously unreported work activity for those with no CPS job, though a second or third job is more likely to be reported when the detailed question is asked. Among reports for other household members, the detailed question elicits significantly more reports of additional employment than the global question for all three CPS employment status groups (no CPS employment, one CPS job, or two or more CPS jobs). This is consistent with stronger cues being more important for activating respondents' memories or encouraging respondents to make use of situational knowledge when they are reporting for household members other than their spouse. Both for spouses and for other household members, among those with additional work identified, the detailed probe has a larger effect than the global probe on the number of hours reported (11.0 versus 6.1 hours for spouses and 11.2 versus 6.9 hours for other household members).

Because we have good information about the type of informal work performed only for respondents asked the detailed question, we cannot repeat this analysis with research work excluded. We have replicated the Table 3 tabulations excluding added work that involved less than four hours during the reference week. These results are shown in Appendix Table C2 (Supplemental material). As in our baseline results, the detailed probe elicits more unreported work activity than the global probe. This is especially true for proxy reports and, among the proxy reports, for other household members rather than a spouse.

Table 4 examines how taking into account the additional work activity identified by probing affects the estimated employment rate, defined as the share of the sample employed during the survey reference week, and the estimated multiple job holding rate, defined as the share of CPS employed persons holding more than one job during the reference week. The table reports estimated rates based on the responses to the CPS questions; augmented rates that add the additional work activity identified by probing to the numerator used to calculate the rate in question; and differences between each pair of estimated rates. In the full sample, as shown by the numbers in the first two rows of the table, probing to identify additional work activity consistently raises both the employment rate and the estimated multiple job holding rate. The increase in the employment rate is larger for those who received the detailed probe than for those who received the global probe. The difference in the effects of the detailed versus the global probe on the estimated employment rate in the full sample is a statistically significant 2.0 percentage points. Both the global and the detailed probe produce substantially larger effects on the multiple job holding rate. Again, in the full sample, the effect is larger with the detailed probe, which raised the multiple job holding rate by a statistically significant 3.6 percentage points more than the global probe.

Disaggregating by whether the respondent is reporting for herself or for another household member makes clear that the differences in the effects on the employment rate we observe in the full sample for the detailed question versus the global question arise primarily among the proxy report cases. For proxy reports, the effect on the employment rate of incorporating additional work activity identified by probing is a statistically significant 3.0 percentage points larger based on asking the detailed question as opposed to the global question. For the self-report cases, the corresponding difference in employment rate effects is smaller (1.2 percentage points) and not statistically significant.

The same general pattern holds for the multiple job holding rate. For proxy reports, incorporating additional work identified by probing raises the multiple job holding rate by a statistically significant 8.4 percentage points more when the detailed question is asked than when the global question is asked. For self-reports, in contrast, although both the detailed and the global question questions identify a sizable number of second jobs not reported in response to the CPS questions, the difference between the two effects is small and statistically indistinguishable from zero.

As with the results reported in Table 3, there is heterogeneity within the proxy report cases. Results are shown separately for spouses and other household members in the bottom two panels of Table 4. Recall that, among those reporting about themselves, the global and the detailed questions have statistically indistinguishable effects on both the employment rate and the multiple job holding rate. In the reports for spouses, the effect on the employment rate of asking the detailed question is statistically indistinguishable from the effect of asking the global question, but asking the detailed question has a notably larger effect on the multiple job holding rate. Adding work activity identified by probing raises the multiple job holding rate for a spouse by a statistically significant 6.7 percentage points more when the detailed question is asked than when the global question is asked. Finally, among reports for other household members, asking the detailed question rather than the global question has a larger effect on both the employment rate and the multiple job holding rate.

Table 4. Effect of Additional Work Activity Identified by Probing on Employment and Multiple Job Holding Rates, Global versus Detailed Probe.

	Employment rates					Multiple job holding rates				
	Sample size	CPS questions	Augmented by Probing	Difference	(*p*-value)	Sample size	CPS questions	Augmented by Probing	Difference	(*p*-value)
Full sample										
Global prompt	2,492	83.4	86.3	2.9	(<0.001)	2,078	23.9	39.9	16.0	(<0.001)
Detailed prompt	2,447	83.4	88.3	4.9	(<0.001)	2,041	23.6	43.2	19.6	(<0.001)
Detailed minus global	–	0.0	2.0	2.0	–	–	0.3	3.3	3.6	–
(*p*-value)	–	(0.984)	(0.035)	(<0.001)	–	–	(0.820)	(0.031)	(0.002)	–
Self reports										
Global prompt	1,364	94.7	96.9	2.3	(<0.001)	1,291	32.0	55.9	23.9	(<0.001)
Detailed prompt	1,340	94.7	98.1	3.4	(<0.001)	1,269	31.8	56.5	24.7	(<0.001)
Detailed minus global	–	0.1	1.2	1.2	–	–	−0.2	0.6	0.7	–
(*p*-value)	–	(0.951)	(0.042)	(0.070)	–	–	(0.933)	(0.769)	(0.667)	–
Proxy reports										
Global prompt	1,128	69.8	73.5	3.7	(<0.001)	787	10.6	13.5	2.9	(<0.001)
Detailed prompt	1,107	69.7	76.4	6.7	(<0.001)	772	10.0	21.2	11.3	(<0.001)
Detailed minus global	–	0.0	2.9	3.0	–	–	−0.6	7.8	8.4	–
(*p*-value)	–	(0.987)	(0.110)	(0.002)	–	–	(0.710)	(<0.001)	(<0.001)	–
Spouse										
Global prompt	583	82.7	85.4	2.7	(<0.001)	482	11.4	14.9	3.5	(<0.001)
Detailed prompt	542	81.6	84.7	3.1	(<0.001)	442	12.7	22.9	10.2	(<0.001)
Detailed minus global	–	−1.1	−0.7	0.4	–	–	1.3	7.9	6.7	–
(*p*-value)	–	(0.623)	(0.730)	(0.697)	–	–	(0.557)	(0.002)	(<0.001)	–
Other household member										
Global prompt	545	56.0	60.7	4.8	(<0.001)	305	9.2	11.1	2.0	(0.007)
Detailed prompt	565	58.4	68.5	10.1	(<0.001)	330	6.4	19.1	12.7	(<0.001)
Detailed minus global	–	2.4	7.8	5.3	–	–	−2.8	7.9	10.8	–
(*p*-value)	–	(0.411)	(0.007)	(0.001)	–	–	(0.184)	(0.005)	(<0.001)	–

To assess the sensitivity of these findings, we have replicated the Table 4 tabulations for respondents who received the detailed prompt but with added research work excluded (results reported in Appendix Table B2, Supplemental material). We also have replicated the full Table 4 analysis but with added work involving less than four hours during the reference week excluded (results reported in Appendix Table C3, Supplemental material). Even with these exclusions, incorporating the added work identified by probing produces a statistically significant increase in the estimated employment rate and has an even larger effect on the estimated multiple job holding rate. In the tabulations that exclude added work involving less than four hours, we can examine the effects of asking the detailed versus the global question about informal work. All of the qualitative findings from our Table 4 analysis are robust to the exclusion of very-low-hours added work.

5. Discussion and Conclusion

The results we have reported suggest that there may be a substantial number of people involved in informal work that is not captured by the standard CPS questions. In our sample, additional probing using either a global question or a decomposed question identified a sizeable number of reports of additional work activity. This was true whether a respondent was reporting for themselves or for another household member, and also whether the other household member was a spouse or someone else. Accounting for this additional work activity raised both the employment rate and the multiple job holding rate, defined in each case in the same way as in the monthly labor force statistics published by the BLS.

Further, our results suggest that different ways of probing for additional work activity may produce different results depending on the person about whom a respondent is reporting. For those in our sample reporting about themselves, the effects of a global probe are not very different from the effects of a more detailed probe that decomposes various possible types of work activity a person might have carried out and provides examples. Among these self-reports, the detailed probe elicits a significantly greater number of reports of additional work activity only for those who already had mentioned two or more jobs in response to the standard CPS questions. In contrast, for proxy reports, the detailed probe more consistently elicits a greater number of such reports. This is especially true when a respondent is reporting for a household member other than her spouse.

For a self-report, asking the detailed question rather than the global question has essentially the same effect as asking the global question on both the employment rate and the multiple job holding rate. For reports about a spouse, asking the detailed question produces a larger effect on the multiple job holding rate but not the employment rate. Finally, for reports about other household members, asking the detailed question has a larger effect on both the employment rate and the multiple job holding rate.

The added work activity identified through probing in our survey most likely is attributable either to respondents not having understood that this activity should have been reported in answering the CPS employment questions or to the cue offered by the probe activating their memories of the activity. The fact that, for self-reports, the detailed probe generally does not produce larger effects than the global probe may suggest that memories about own recent work activity tend to be relatively accessible. In contrast, for proxy

reports – and especially proxy reports pertaining to household members other than the spouse – the detailed probe more consistently produces more reports of added work activity, suggesting that strong cues are likely to be useful when seeking information from household survey respondents about work done by others in their households.

An important limitation of our study is that the sample for which we collected data is not representative of the population as a whole. All of our respondents are individuals who are active on Mechanical Turk and thus likely (though not certain) to have been involved at least in that form of informal work activity during the survey reference week. We would not expect the same necessarily to be true of other members of respondents' households, but even that group is younger and more educated than the population as a whole and may be atypical in other respects. For these reasons, even if we were to reweight the data we have collected to match the observable demographic characteristics of the broader population, the estimates derived from our survey responses could not be generalized to that universe. Another caution about drawing conclusions from our study about biases in the responses to the CPS employment questions is that our survey was conducted online, whereas the CPS responses are collected via telephone or face-to-face interviews. The survey findings nonetheless provide important evidence about the sensitivity of survey estimates to asking more probing questions and structuring the probes in different ways.

To the extent that irregular or informal work has become more common, under-reporting of work activity in response to the standard CPS questions could have become more prevalent over time. The fact that the share of people reporting self-employment income on their tax returns has been rising while the share reporting self-employment income in household survey data has been flat or declining is consistent with this possibility (Katz and Krueger 2019b; Abraham et al. 2018). On the other hand, surveys designed specifically to capture informal work activity do not show continued overall growth in the rate of participation in such activity in recent years, though participation in online platform work appears to have become more prevalent and cyclical effects could have masked a continuation of an underlying positive trend (Bracha and Burke 2018). It is important in any case to understand clearly what the CPS employment questions are and are not capturing, and to think about whether and how they could be improved or supplemented.

As the agency responsible for producing official U.S. labor force statistics, the BLS has a strong interest in producing the best possible information about individuals' work arrangements and how they are evolving. The Contingent Work Supplement (CWS) to the CPS, administered on five occasions between 1995 and 2005 and again in 2017, provides valuable information on this topic (see, e.g., Polivka 1996, Cohany 1996, and Bureau of Labor Statistics 2018). Because the CWS takes as its starting point the employment reported in response to the standard CPS questions, asking additional questions only about the main job reported for each person, it provides no information about any work not reported in answer to the standard CPS questions or work that is secondary to a main job. There is a need, we would argue, for efforts to design questions that can be used to obtain information about informal work more broadly. That said, if the types of informal work that people are doing change over time, the questions that are most appropriate to ask may change as well, something that could make it more difficult to produce estimates of informal work activity that are consistent over time.

In future research, it would be of value to examine whether our findings can be replicated in samples that have different characteristics and, ideally, are more representative of the general population. There also would be value in replicating our analysis using the survey modes that are employed in the CPS (telephone and face-to-face interviews) rather than collecting responses to an online instrument. In this study, we have compared the effects of asking a global question to the effects of asking a particular decomposed question for learning about informal work not reported in response to the standard CPS questions. The categories and examples included in our decomposed question focused on activities in which compensation is received mainly for a person's labor, as opposed to being provided in connection with selling a product (e.g., selling crafts on e-Bay) or providing temporary use of a capital asset (e.g., renting out a room in a house through Airbnb). It is not yet clear, however, which categories and examples of activities should be mentioned to obtain the most complete accounting of work done for pay or profit. Further research on how best to ask about such activity would be desirable. Additional testing also might incorporate follow-up questions about when any added activities were performed (to determine whether and to what extent activities that occurred prior to the survey reference period may have been reported), how much was earned from any missed activities (as a means of gauging their importance), and why the activities were not reported initially.

6. References

Abraham, K.G., J.C. Haltiwanger, K. Sandusky, and J.R. Spletzer. 2018. "Measuring the Gig Economy: Current Knowledge and Open Issues." National Bureau of Economic Research Working Paper No. 24950. August. Doi: http://dx.doi.org/10.3386/w24950.

Abraham, K.G. and S.N. Houseman. 2018. "Making Ends Meet: The Role of Informal Work in Supplementing Americans' Income." Upjohn Institute, unpublished working paper. December. Available at: https://www.aeaweb.org/conference/2019/preliminary/paper/QreAaS2h (accessed July 2019).

Allard, M.D. and A.E. Polivka. 2018. "Measuring Labor Market Activity Today: Are the Words Work and Job Too Limiting for Surveys?" *Monthly Labor Review*. November. Available at: https://www.bls.gov/opub/mlr/2018/article/pdf/measuring-labor-market-activity-today.htm (accessed April 2019).

Belli, R.F., N. Schwarz, E. Singer, and J. Talarico. 2000. "Decomposition Can Harm the Accuracy of Behavioural Frequency Reports." *Applied Cognitive Psychology* 14: 295–308. Doi: http://dx.doi.org/10.1002/1099-0720(200007/08)14:4<295::AID-ACP646>3.0.CO;2-1.

Blair, E. and S. Burton. 1987. "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions." *Journal of Consumer Research* 14(2): 280–288. Doi: https://doi.org/10.1086/209112.

Board of Governors of the Federal Reserve System. 2017. *Report on the Economic Well-Being of U.S. Households in 2016*. Washington, D.C: Board of Governors of the Federal Reserve System. Available at: https://www.federalreserve.gov/publications/files/2016-report-economic-well-being-us-households-201705.pdf (accessed May 2019).

Bower, G.H. and S.G. Gilligan. 1979. "Remembering Information Related to One's Self." *Journal of Research in Personality* 13: 420–432. Doi: https://doi.org/10.1016/0092-6566(79)90005-9.

Bracha, A. and M.A. Burke. 2018. "The Ups and Downs of the Gig Economy, 2015–2017." Federal Reserve Bank of Boston Working Paper 18–12. October. Available at: https://www.bostonfed.org/publications/research-department-working-paper/2018/the-ups-and-downs-of-the-gig-economy-2015-2017.aspx (accessed May 2019).

Bracha, A. and M. A. Burke. 2019. "How Big is the Gig?" Federal Reserve Bank of Boston, unpublished working paper. January.

Brown, N.R., L.J. Rips, and S.K. Shevell. 1985. "The Subjective Dates of Natural Events in Very-Long-Term Memory." *Cognitive Psychology* 17(2): 139–177. Doi: https://doi.org/10.1016/0010-0285(85)90006-4.

Bureau of Labor Statistics. Undated. "Labor Force Statistics from the Current Population Survey: Frequently Asked Questions." Available at: https://www.bls.gov/cps/faq.htm (accessed June 2018).

Bureau of Labor Statistics. 2018. "Contingent and Alternative Employment Arrangements, May 2017." Available at: https://www.bls.gov/news.release/pdf/conemp.pdf (accessed June 2018).

Cohany, S.R. 1996. "Workers in Alternative Employment Arrangements." *Monthly Labor Review*. October: 31–45. Available at: https://www.bls.gov/mlr/1996/10/art4full.pdf (accessed May 2019).

Dashen, M. 2000. "The Effects of Retention Intervals on Self- and Proxy Reports of Purchases." *Memory* 8(3): 129–143. Doi: https://doi.org/10.1080/096582100387560.

Fowler, F.J., Jr. 1992. "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56(2): 218–231. Doi: https://doi.org/10.1086/269312.

Katz, L.F. and A.B. Krueger. 2019a. "Understanding Trends in Alternative Work Arrangements in the United States." NBER Working Paper No. 25425. Cambridge, MA: National Bureau of Economic Research. Doi: https://doi.org/10.3386/w25425.

Katz, L.F. and A.B. Krueger. 2019b. "The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015." *ILR Review* 72(2): 382–416. Doi: https://doi.org/10.1177%2F0019793918820008.

Kojetin, B.A. and L.A. Miller. 1993. "The Intrahousehold Communications Study: Estimating the Accuracy of Proxy Responses at the Dyadic Level." Paper presented at the 48th Annual Conference of the American Association for Public Opinion Research, St. Charles, Illinois. May. Available at: http://www.asasrms.org/Proceedings/papers/1993_188.pdf (accessed May 2019).

Kuiper, N.A. and T.B. Rogers. 1979. "Encoding of Personal Information: Self-Other Differences." *Journal of Personality and Social Psychology* 37(4): 499–514. Doi: http://dx.doi.org/10.1037/0022-3514.37.4.499.

Martin, E. 2006. "Survey Questionnaire Construction." U.S. Census Bureau Research Report Series, Survey Methodology #2006-13. Available at: https://www.census.gov/srd/papers/pdf/rsm2006-13.pdf. (accessed May 2019).

Martin, E. and A.E. Polivka. 1995. "Diagnostics for Redesigning Survey Questionnaires: Measuring Work in the Current Population Survey." *Public Opinion Quarterly* 59(4): 547–567. Doi: https://doi.org/10.1086/269493.

Menon, G. 1997. "Are the Parts Better than the Whole? The Effects of Decompositional Questions on Judgments of Frequent Behaviors." *Journal of Marketing Research* 34(3): 335–346. Doi: https://doi.org/10.1177%2F002224379703400303.

Phillips, J.M., B.A. Bickart, and G. Menon. 2006. "Reporting About Others' Behavior: The Role of Judgment Strategy, Knowledge, and Regularity." September. Available at: https://ssrn.com/abstract=946247 (accessed April 2019).

Polivka, A.E. 1996. "A Profile of Contingent Workers." *Monthly Labor Review*. October: 10–21. Available at: https://www.bls.gov/opub/mlr/1996/article/profile-of-contingent-workers.htm (accessed May 2019).

Robles, B. and M. McGee. 2016. *Exploring Online and Offline Informal Work: Findings from the Enterprising and Informal Work Activities (EIWA) Survey*. Washington: Board of Governors of the Federal Reserve System. Doi: https://doi.org/10.17016/FEDS.2016.089.

Schober, M.F. and F.G. Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 61(4): 576–602. Doi: https://doi.org/10.1086/297818.

Schober, M.F., A.L. Suessbrick, and F.G. Conrad. 2018. "When Do Misunderstandings Matter? Evidence from Survey Interviews about Smoking." *Topics in Cognitive Science* 10(2): 452–484. Doi: https://doi.org/10.1111/tops.12330.

Schwarz, N. 1999. "Self-Reports: How the Questions Shape the Answers." *American Psychologist* 54(2): 93–105. Doi: http://dx.doi.org/10.1037/0003-066X.54.2.93.

Schwarz, N. and T. Wellens. 1997. "Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Actors and Observers." *Journal of Official Statistics* 13(2): 159–179. Doi: https://doi.org/10.1.1.39.5355.

Sudman, S. and N.M. Bradburn. 1973. "Effects of Time and Memory Factors on Response in Surveys." *Journal of the American Statistical Association* 68: 805–815. Doi: https://doi.org/10.2307/2284504.

Sudman, S., N.M. Bradburn, and N. Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Tourangeau, R. 2000. "Remembering What Happened: Memory Errors and Survey Report." In *The Science of Self-Report: Implications for Research and Practice*, edited by A.A. Stone, C.A. Bachrach, J.B. Jobe, H.S. Kurtzman, and V.S. Cain, 29–47. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Tourangeau, R., F.G. Conrad, M.P. Couper, and C. Ye. 2014. "The Effects of Providing Examples in Survey Questions." *Public Opinion Quarterly* 78(1): 100–125. Doi: https://doi.org/10.1093/poq/nft083.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

U.S. Census Bureau. 2006. *Design and Methodology: Current Population Survey*, Technical Paper No. 66. Washington DC. Available at: https://www.census.gov/prod/2006pubs/tp-66.pdf (accessed July 2019).

Received August 2018
Revised January 2019
Accepted May 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 509–529, http://dx.doi.org/10.2478/JOS-2019-0022

Correlates of Representation Errors in Internet Data Sources for Real Estate Market

Maciej Beręsewicz[1]

New data sources, namely big data and the Internet, have become an important issue in statistics and for official statistics in particular. However, before these sources can be used for statistics, it is necessary to conduct a thorough analysis of sources of nonrepresentativeness.

In the article, we focus on detecting correlates of the selection mechanism that underlies Internet data sources for the secondary real estate market in Poland and results in representation errors (frame and selection errors). In order to identify characteristics of properties offered online we link data collected from the two largest advertisements services in Poland and the Register of Real Estate Prices and Values, which covers all transactions made in Poland. Quarterly data for 2016 were linked at a domain level defined by local administrative units (LAU1), the urban/rural distinction and usable floor area (UFA), categorized into four groups. To identify correlates of representation error we used a generalized additive mixed model based on almost 5,500 domains including quarters.

Results indicate that properties not advertised online differ significantly from those shown in the Internet in terms of UFA and location. A non-linear relationship with the average price per m^2 can be observed, which diminishes after accounting for LAU1 units.

Key words: Big data; non-ignorable missing data; representation error; self-selection error; INLA.

1. Introduction

Big data and the Internet as a data source have become an important issue in statistics, in particular in official statistics. There are number of multinational initiatives (e.g., ESSnet on Big Data) that focus on the quality and suitability of estimates based on new data sources to complement or supplement existing statistical information. Before these data can be used for official statistics, it is crucial to explore potential sources of nonrepresentativeness. In this context, Daas et al. (2015), Buelens et al. (2014), Beręsewicz (2016, 2017), and Citro (2014) discussed coverage, nonresponse and measurement errors. Japec et al. (2015), Pfeffermann (2015), and Beręsewicz et al. (2018)

[1] Poznań University of Economics and Business, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, Poland. Email: maciej.beresewicz@ue.poznan.pl

Acknowledgments: This study was supported by the National Science Center, Preludium 7 grant no. 2014/13/N/HS4/02999. I would like to kindly thank (1) Grupa OLX sp. z o.o. owner of the Otodom service, (2) Polska Press sp. z o.o. owner of the Dom.Gratka.pl service for providing access to the data via API and historical data in an aggregated form and (3) The Trade and Service Department, Statistics Poland for preparing aggregated data from the Register of Real Estates Prices and Values in Poland. I thank Alicja Szabelska-Beręsewicz, Kamil Wilak, Tomasz Klimanek, Michal Szczudlak and Tomasz Hinc for helpful comments and discussions. Moreover, I sincerely thank three anonymous reviewers and the associate editor for a valuable discussion that led to significant improvements in the article.

elaborate on the coverage and nonresponse error, which can lead to significant bias in big data sources, in particular if missingess is nonignorable.

A number of empirical studies compare new data sources with official statistics. For instance, Daas et al. (2015) studied the consumer confidence index based on social media and survey data in the Netherlands; the Billion Price Project is aimed at calculating CPI based on web-scraped data and Cavallo (2013) provided an insight into discrepancies between official indicators in Argentina.

Research is also conducted on the use of Internet data sources for the real estate market. Notable examples include *the number and prices of houses for sale* indices published by Statistics Netherlands from 2013 to 2016. The indices were calculated based on properties offered for sale on the JAAL.nl website (Statistics Netherlands 2018). Hoekstra et al. (2012) discussed details regarding data collection which, according to the authors' knowledge, can be considered the first use of online real estate data to produce official statistics. Unfortunately, the indices have been discontinued as part of cost cutting measures, being non-compulsory statistics.

Other examples can be found in the literature on real estate. For instance, there are a number of studies focusing on asking and transaction prices and values (Ihlanfeldt and Martinez-Vazquez 1986; Kiel and Zabel 1999) but they were mainly based on household surveys and register data (cf. Fleishman and Gubman 2015). In this context, Lozano-Gracia and Anselin (2012) describe the use of advertising signs and newspaper ads to survey asking prices of properties and link with cadastral records between 2002 and 2007 in Bogota, Columbia; Anenberg and Laufer (2017) used online ads to create an up-to-date list price index as a proxy for the price index based on administrative sources, and Beręsewicz (2016) investigated sources of bias in estimates of the average asking price per m^2 for residential properties by comparing survey data and advertising services in Poland.

New sources contain both measurement and representation errors associated with variables and objects (Wallgren and Wallgren 2014; Zhang 2012; Reid et al. 2017). Zhang (2012) proposed a two-phase life cycle model for integrated statistical microdata, where the first phase is based on a single source and the second one – on integrated sources. Reid et al. (2017) extended this model by including a third phase devoted to the evaluation or estimation of the quality of the final outputs, taking into account all sources of error.

In the article we focus on the representation aspect, emphasizing that it is crucial to keep in mind differences in measurements when using these data sources for statistics. Representation or non-observation errors include frame, selection and missing/redundancy errors. All of these errors are discussed briefly in the case of a single source.

Frame errors are differences between the target population and the accessible set. In this context Reid et al. (2017) distinguish the following measures: lag in updating population changes, undercoverage, overcoverage and authenticity (incorrect or multiple identifiers). In the case of Internet data sources, reporting lags can be linked to differences between the moment when information is posted online and the actual event (e.g., a flat is offered for sale a couple of weeks before it is published online). Undercoverage refers to a situation where some units are not observed online (e.g., properties advertised in newspaper or between friends), while overcoverage error occurs when a given unit does not belong to

the target population. The second situation is common in Internet data sources because online services rarely have tools to verify if a given unit is correctly classified (e.g., a house or flat) or whether it exists (e.g., a future investment or property that is already being built). Finally, information published online may not contain any identifiers (e.g., no property parcel number).

Selection errors arise when objects in the accessible set do not appear in the accessed set. Reid et al. (2017) propose four indicators: adherence to the reporting period, dynamics of births and deaths, readability and inconsistent objects/units. The reporting period may refer to a situation when a statistical agency has an agreement with a data provider, which states that the data should be delivered on fixed date(s). However, in most cases special tools are developed in order to scrape or access the databases directly to decrease the reporting burden on the data holder (Hoekstra et al. 2012). Readability is certainly an issue in Internet data sources, mainly owing to restrictions on publicly available data (e.g., 1% sample of public tweets) or limited access rights to data (e.g., query results from the browser and API may vary).

Finally, missing/redundancy errors arise from the misalignment between the accessed set and the observed set, which could be measured by unit nonresponse rate, share of duplicated records or share of units that have to be adjusted to create statistical units (Reid et al. 2017). In Internet data sources we only observe units that either use the Internet or place information online (e.g., property ads). Given unrestricted possibilities of creating multiple accounts and content, the number of duplicated records is significant (e.g., several advertisements for the same property). Finally, objects in Internet data sources may refer to multiple statistical units (e.g., advertisement for a property and a garage with separate mortgages).

To assess these errors, it is necessary to rely on external data sources. However, in the case of Internet data sources (e.g., advertisements services), this may be problematic. For instance, there are rarely official statistics on this topic, there is no sampling frame for such units or persons/institutions that publish information online, and research on why such services are used is scarce. Some information about sources of errors can be found, for instance, in *the Information and Communication Technology* surveys coordinated by Eurostat. Certainly, differences between the target and observed set result from the underlying selection mechanism that prompts persons/companies to use certain services. Thus, it may be difficult to disentangle error and bias in new sources into frame and selection error and bias, which is why we will use representation errors as a general term for these errors.

In situations where sources of representation errors are unknown, they can be detected by comparing new data sources with auxiliary sources already used in statistics that is, surveys or registers (Beręsewicz 2017; Pfeffermann 2015; Lohr and Raghunathan 2017). Data can be linked at a domain level to provide information about the selection mechanism and characteristics of units that are not present online. It is crucial to discover the underlying selectivity because it may be linked to the effects and methods of dealing with these errors (Brick 2015).

The study described below focuses on residential properties in the secondary real estate market in Poland. The population of interest consists of residential properties offered for sale in the secondary market. This population may be of interest to official statistics,

particularly when it comes to estimating the asking-to-transaction-price ratio, price indices, measuring time-to-sale, or as an indicator of the situation in the real estate market. As this population is not dynamic, unlike, for instance, the population of mobile phone or social website users, the related data may not be considered as big data in terms of volume, but should be treated as such in terms of variety or complexity. One should keep in mind that, as Citro (2014) states, the Internet, (. . .), not only generates a great deal of today's "big data", but also provides ordinary-size data in a more accessible way – for example, access to public opinion polls or to local property records.

Given the nature of the population, as well as limited research on real estate market brokers and owners, we cannot separate representation errors into frame and selection errors. To investigate possible correlates of these errors, we used an auxiliary data source, namely the Register of Real Estates Prices and Values, which covers all sold properties that have an established ownership. Using this independent source on a different but related population, we can obtain information about types of properties that are not advertised online but are sold. Since we obtained these data a year after the last transaction took place, errors due to lags in registration can be regarded as negligible.

In the absence of access to a national unit-level register, quarterly data for 2016 were linked at a domain level defined as an interaction between the urban/rural distinction, the category of usable floor area and Local Administrative Unit (LAU 1, 380 districts) in Poland. To account for the time-lag between the moment of publishing advertisements and actual transactions, we linked transactions from q with advertisements from $q - 1$, where $q = \{2, 3, 4\}$ refers to the given quarter in 2016. This variable represents the time-to-sale, which is a lag between posting an advertisement online and the sale of the property. In total, 5,507 domains (including quarters) with a non-zero number of transactions for 376 districts were analysed. Bias was not examined because of the measurement error resulting from the difference in the definition of the target variables in the sources (asking vs transaction price).

The research questions that the article seeks to answer are as follows:

- what are correlates of non-observation errors?
- is the non-observation error non-ignorable?

The article has the following structure. Section 2 covers the data collection design and a thorough description of data sources used in the study. Section 3 defines the measure of selectivity, describes the modelling procedure and the model used to explore correlates of selectivity. Section 4 is devoted to exploratory data analysis and the presentation of modelling results. The article ends with conclusions and a discussion of the results.

2. Data

2.1. Internet Data Sources – Otodom.pl and Dom.Gratka.pl

The process of data acquisition was designed to minimize errors and interruptions. This is why web-scraping was not considered as a mode of data collection, as it is sensitive to changes in the structure of the webpage, the IP can be blocked and often not all data are present on the webpage.

Instead, we decided to approach owners of two leading Polish online real estate advertising services – Otodom.pl (https://www.otodom.pl, further on referred to as Otodom) and Dom.Gratka (http://dom.gratka.pl, later on referred to as Gratka) – to inquire about the possibility of accessing their databases via the Application Programming Interface (API). Acquiring data through the API is a more robust solution and results in structured and highly dimensional data.

We contacted the companies by sending a formal e-mail inquiry with a clear description of the purpose the data would be used for. Both companies were open to collaboration, interested in the results and shared their data. In the case of Otodom, access was free of charge, while data from Gratka were made available for a small fee. No special conditions were made by the companies, except for a request made by Gratka to prepare a short note for their blog about current trends in prices on their service. Finally, we were given special access tokens and passwords to connect to the databases through the Simple Object Access Protocol (SOAP) and Representational State Transfer (REST) APIs. In addition, we received monthly historical data in an aggregated form, which had been prepared according to our request.

The scope of access varied in each case. Gratka API only offered access to 26 variables, which described each advertised property, while the Otodom data set included 46 fields, which, in addition to property characteristics, contained anonymised information about the person/company that placed the ad, and ad characteristics (e.g., promoted, number of views). This corresponds to the distinction between 'accessible set' and 'accessed set' proposed by Zhang (2012) but with regard to the number of variables made available.

The data collection process started in Q4 2015 and still continues. Data were collected on a weekly basis. Each Saturday night a script was run to download all advertisements available for the whole of Poland. First, the raw data were stored in plain text files: the initial volume was about 100 GB for Gratka and 900 GB for Otodom. Then, the data were processed on a Linux server using `bash` and `jq` (data were stored in JavaScript Object Notation, JSON, format). Later on, the data were processed using R (R Core Team 2017) language with the help of the following packages: `data.table`, `tidyverse` and `stringi`. The initial number of advertisements from Otodom was over 20 million and from Gratka – 28 million. However, the data set contained duplicates as a result of the data collection process and the organization of the Polish property market.

The editing phase started by analyzing the data structure and metadata associated with the variables (e.g., definitions). Then, we removed objects that did not belong to the target population (e.g., not located in Poland, unfinished investments, primary market), contained erroneous or missing values in prices or usable floor area (e.g., properties with PLN 1 price or with UFA of over 30 000 m^2), did not belong to the reference period based on the date of the last modification or contained information regarding multiple properties. Then, some variables were harmonized (e.g., build year) in order to ensure consistency between partially standardized and non-standardized data (e.g., information provided directly by the owner/broker).

Properties in the Polish real estate market can be sold under closed or open agreements, which means that the same properties are advertised multiple times. Under an open agreement, multiple brokers can place an ad regarding the same property, often with different descriptions, and data holders cannot remove duplicates. This required additional

attention during the data processing stage. As the study involved quarterly data, deduplication was conducted within quarters based on the following *naive* procedure: (1) the most recent advertisements were selected based on the ad identification number, and then (2) using combinations of variables referring to province, price, usable floor area, number of rooms, year of construction and street name. Certainly, not all duplicates were located because of slight differences in values in floor area or street name. This problem could be resolved by probabilistic methods (cf. for recent advances Steorts et al. 2016; Chen et al. 2018). However, as most cases of duplicate ads were found in large cities (regional capitals), which are already covered by advertising portals, they did not pose a problem in the analysis at the domain level.

Finally, properties that did not have any information regarding location or could not be geocoded in order to be classified as either rural or urban were removed. All advertisements were geolocated based on the district and location using Google Geocode API. In view of the goal of the study no imputation was applied.

After the cleaning process, the final Gratka data set consisted of 816,100 ads with 526,720 and for Otodom – 699,958 ads with 394,953 unique objects.

2.2. *The Register of Real Estate Prices and Values*

The Register of Real Estate Prices and Values, later on referred to as the register, is a public register maintained by the district governor, which contains information about real estate prices included in notarial deeds and real estate values provided by real estate appraisers in appraisal reports, whose abridged versions are included in the register of land and buildings. It is worth noting that there is no single national register but each district (LAU1) in Poland maintains its own register (380 units). The data are reported to the Central Statistical Office in Poland five times a year – two months after each quarter and by April of the following year for the whole previous year. The register contains information on the population of sold properties: flats, buildings, built-up and land properties. The data are used to prepare an annual report entitled *Real Estate Sales*. Since 2015, these statistics have been broken down by market type – primary and secondary.

The Central Statistical Office divides residential properties into those sold in the primary and the secondary market. A primary market sale is defined as a transaction made in the free market, where the selling party is a legal person and the average price per 1 m^2 of usable floor area is at least 2,000 PLN. Transactions in the secondary market include other market transactions carried out in the free market and auction sales.

Aggregated quarterly data for 2015 and 2016 from the Register were obtained from the Trade and Service Department of the Central Statistical Office. The data contain information for the following variables: (1) district identifier and name, (2) market type (total, primary, secondary), (3) location (total, rural/urban, town with district status, town with district status with population of under 200,00, town with a district status with a population of over 200,000, unknown location), (4) categorized usable floor area in m^2 (total, under 40, 40–60, 60–80, and over 80 m^2), and variables (1) number of transactions, (2) value of transactions, (3) sum of floor area and (4) median price per m^2.

For the purpose of the study we selected domains defined as interactions between three quarters (2016 Q2-Q4), LAU 1, rural/urban distinction and four categories of usable floor

Table 1. Distribution of number of transactions within studied domains between Q2 and Q4 2016.

Total	No of domains	Min	1st Quantile	Median	Mean	3rd Quantile	Max
74 588	5 507	1	2	4	13.54	10	1 450

area. The analysis involved a total of 5,507 domains with a non-zero number of transactions in the secondary market for 2016. In that year transactions in the secondary market were reported in 376 out of 380 districts.

Table 1 presents the distribution of the number of transactions across the domains for 2016. The median number of transactions is 5 and the mean is 13.5, which indicates right skewness. In total, over 73,000 transactions were made in the secondary real estate market.

Table 3 and Figure 7 in the Appendix (Section 6) present a comparison between the register, Otodom and Gratka at domain levels. Note that there is a time-lag: online data refer to the Q1-Q3 2016 period and transactions to the Q2-Q4 2016 period. Pearson's correlation coefficient between the log-transformed number of transactions and the log-transformed number of ads in Otodom and Gratka is equal to 0.76 and 0.73 respectively, and between Otodom and Gratka – 0.88. Points over the black line indicate domains where the number of transactions is larger than that of ads.

2.3. Other Auxiliary Information

To account for the Internet coverage error, we used the broadband penetration ratio calculated as the number of buildings with access to broadband Internet (i.e., buildings for which Internet providers are able to provide broadband services) to all buildings in a given domain. This measure is calculated by the Office of Electronic Communications for all cities in Poland on a yearly basis. We calculated this indicator for urban/rural areas within LAUs using data for 31 December 2015.

3. Methods

3.1. The Approach

The following approach was adopted in the study. Otodom and Gratka were linked with register data at domain level. Because only domains with a non-zero number of transactions were selected, representation error was measured with reference to domains containing sold residential properties. Further, we defined the target variable, denoting non-observation error, by equation (1).

$$y_d^{(q)} = \begin{cases} 1, & \text{when } n_d^{(q)} > m_{d,otodom}^{(q-1)} \text{ and } n_d > m_{d,gratka}^{(q-1)}, \\ 0, & \text{else}, \end{cases} \tag{1}$$

Where $n_d^{(q)}$, $m_{d,otodom}^{(q-1)}$, $m_{d,gratka}^{(q-1)}$, denote the number of transactions, advertisements on Otodom, advertisements on Gratka in domain d and q-th quarter respectively. Domains were created by interaction between LAU1 units, urban/rural area and four floor area

categories for each quarter. The target variable refers to transactions that did not involve online advertisements. Moreover, as we are dealing with two sources, we are interested in domains that are not represented in any of the advertisement sources. In total, 1,533 out of 5,507 domains were not represented (27.8%).

To detect the correlates of representation errors, and to avoid measurement error, we constructed a generalized additive mixed model which only contains variables from the register

- usable floor area categorized into four groups (`floor_area`),
- urban or rural location (`urban_rural`),
- average price per m^2 (in 1,000 PLN) at domain level, (`average_pricem2`, centered at overall mean equal to 2,554 PLN), and
- broadband Internet coverage in urban and rural areas at LAU1 level (`net_coverage`, centered at overall mean equal to 78.5%).

In order to verify which variables are correlated with the target variable we built the following four models, each serving a different purpose:

- `Model 1` – a generalized linear model with `net_coverage`, `floor_area`, `urban_rural` and interaction `floor_area` and `urban_rural` – this model is used as a baseline model to verify the relationship with the dependent variable.
- `Model 2` – we extended `Model 1` by adding the `average_pricem2` variable, assuming a non-linear relationship, using smoothing spline and thus obtaining a generalized additive (mixed) model – this model is used to verify if representation error is non-ignorable.
- `Model 3` – we extended `Model 1` by adding the `LAU 1` (i.i.d.) random effect and thus obtaining generalized a mixed model – this model is used to account simultaneously for clustering and for the characteristics of the local market at `LAU 1` level.
- `Model 4` – we combined `Model 2` and `Model 3` to obtain a generalized additive mixed model – this final model is used to verify if the errors are non-ignorable by accounting for both the average price and characteristics of the local market.

3.2. *The Model*

We used Integrated Nested Laplace approximation (INLA), which is a new approach to Bayesian inference for latent Gaussian Markov random field (GMRF) models proposed by Rue et al. (2009). The basic idea behind INLA involves using a deterministic approach to approximate Bayesian inference for latent Gaussian models (i.e., GMRF), which, in most cases, makes INLA faster (i.e., a matter of seconds rather than minutes) and more accurate than MCMC alternatives to GMRF. It provides a number of likelihoods and latent models, including spatial random effects, but it is not as flexible as standard Bayesian approaches (see Chen et al. 2014 and Mercer et al. 2014 for an application of INLA for small area estimation with sampling weights and for an introductory book – Faraway et al. 2018). INLA is implemented in C++ but it can be applied by using `R-INLA` package (Lindgren and Rue 2015).

In the empirical study, we modelled the $\mathbf{y}$ variable defined in (1) (we drop q for simplicity), and therefore we assume that it has a binomial distribution given by

(2), where ρ_d denotes the propensity of representation error modelled by means of logistic regression:

$$\begin{aligned}(y_d|\rho_d) &\sim Binomial(n_d, \rho_d), \quad d = 1, \ldots, D.\\ \rho_d(\eta_d) &= \frac{\exp(\eta_d)}{1+\exp(\eta_d)}.\end{aligned} \tag{2}$$

We considered four models for $\rho_d(\eta_d)$:

- `Model 1`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1+\exp(\eta_d)} = \mathbf{x}_d^T\beta, \tag{3}$$

  where β are fixed effects parameters and $\mathbf{x}_d$ are independent variables at domain level (i.e., `net_coverage, floor_area, urban_rural and floor_area × urban_rural`),
- `Model 2`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1+\exp(\eta_d)} = \mathbf{x}_d^T\beta + \nu_j, \tag{4}$$

  where ν_j refers to smoothing spline for the price per m^2. In INLA it is modelled by random walk of order 2 (RW2) and implemented as a random effect. For the Gaussian vector $\nu = (\nu_1, \ldots, \nu_n)$ smoothing spline is constructed assuming independent second-order increments:

$$\Delta^2\nu_j = \nu_j - 2\nu_{j-1} + \nu_{j-2} \sim N\left(0, \tau_\nu^{-1}\right). \tag{5}$$

  The density π of ν is derived from its $n-2$ second-order increments as:

$$\pi(\nu|\tau_\nu) \propto \tau_\nu^{(n-2)/2}\exp\left\{-\frac{\tau_\nu}{2}\sum(\Delta^2\nu_j)^2\right\} = \tau_\nu^{(n-2)/2}\exp\left\{-\frac{1}{2}\nu^\tau\mathbf{Q}\nu\right\}, \tag{6}$$

  where $\mathbf{Q} = \tau\mathbf{R}$ and $\mathbf{R}$ is the structure matrix reflecting the neighborhood structure of the model given by:

$$\mathbf{R} = \tau_\nu \begin{pmatrix} 1 & -1 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}. \tag{7}$$

  We considered $n \in \{10, 15, 20, 25\}$ and, based on information criteria described in the next section, we selected $n = 20$.
- `Model 3`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1+\exp(\eta_d)} = \mathbf{x}_d^T\beta + \psi_i, \tag{8}$$

where $\psi_i \sim N\left(0, \tau_\psi^{-1}\right)$ refers to i.i.d. random effect for LAU 1 units, indexed by $i = 1, \ldots, 376$.

- `Model 4`

$$\rho_d(\eta_d) = \frac{\exp(\eta_d)}{1 + \exp(\eta_d)} = \mathbf{x}_d^T\beta + \nu_j + \psi_i, \tag{9}$$

where all parameters are defined as previously.

For both random effects (ν_j, ψ_i) we used the same penalized complexity (PC) prior suggested by Simpson et al. (2017). Under this new framework, a PC prior for the standard deviation $\sigma = 1/\sqrt{\tau}$ of a latent effect is set by defining parameters $(u,\ \alpha)$, so the interpretation is

$$P(\sigma > u) = \alpha,\ \ u > 0, 0 < \alpha < 1. \tag{10}$$

Hence, PC priors provide a different way to propose priors on the model hyperparameters. In this study, we believe that the probability of the standard deviation being higher than 1 is quite small, so we set $u = 1$ and $\alpha = 0.01$ in the following prior for τ

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp(-\lambda\tau^{-1/2}), \tau > 0, \tag{11}$$

for $\lambda > 0$ where

$$\lambda = -\frac{\mathrm{In}(\alpha)}{u}, \tag{12}$$

and $(u,\ \alpha)$ are the parameters of this prior.

3.3. *Model Selection*

Further, in order to select the most suitable model we used deviance information criterion (DIC; Spiegelhalter et al. 2002), Watanabe-Akaike information criterion (WAIC; Watanabe 2010) and the sum of the log of the conditional predictive ordinate (CPO; Held et al. 2010) values.

The DIC statistic is based on the deviance measure and the number of effective parameters. As in the case of AIC and BIC, models with smaller DIC are better supported by the data.

The WAIC statistic is a more fully Bayesian approach for estimating the out-of-sample expectation starting with the computed log point-wise posterior predictive density and then adding a correction for the effective number of parameters to adjust for overfitting (Gelman et al. 2014, Subsection 3.4).

However, as DIC may underpenalize complex models with many random effects, CPO statistic is often calculated. The CPO is based on the leave-one-out cross-validation procedure, which checks without re-running the model for each observation in turn. For more detail, see Held et al. (2010).

4. Results

4.1. *Exploratory Data Analysis*

Figure 1 presents the share of domains observed and not observed online for three categorical variables: four categories of usable floor area (`floor_area`), urban/rural area

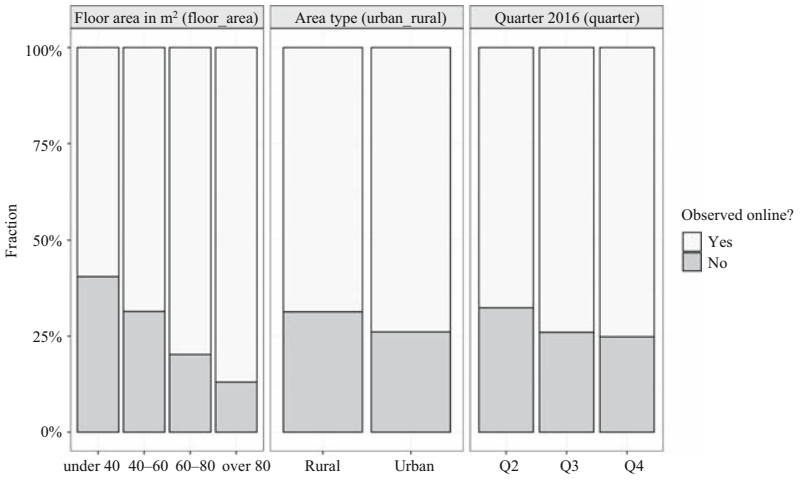

Fig. 1. Share of domains observed and not observed online for three categorical variables: usable floor area, location and quarters in 2016.

(`urban_rural`) and three quarters of 2016. The number of domains not observed online in the first quarter of 2016 is higher than in other quarters.

As expected, domains located in rural areas are less frequently observed online than those located in urban areas. This can be due to the lower broadband Internet coverage, as shown in Figure 2. Median Internet coverage for domains represented in the Internet sources was 83.5% and for those not represented online 75.6%. Another possible explanation is the difference in the use of the Internet by rural and urban dwellers and the fact that online advertising may not be equally necessary in small communities.

There is a linear relation between the categories of usable floor area (UFA) and the fact of being advertised online. Residential properties with UFA under 40 m^2 are more likely to be

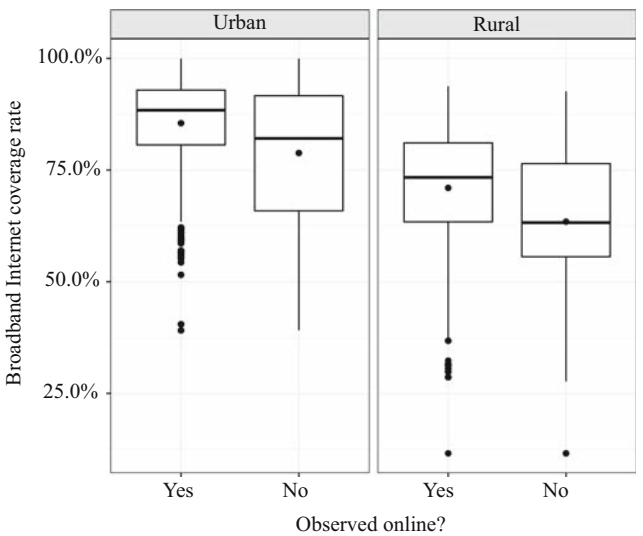

Fig. 2. Distribution of the broadband Internet coverage ratio in rural/urban areas and being observed online.

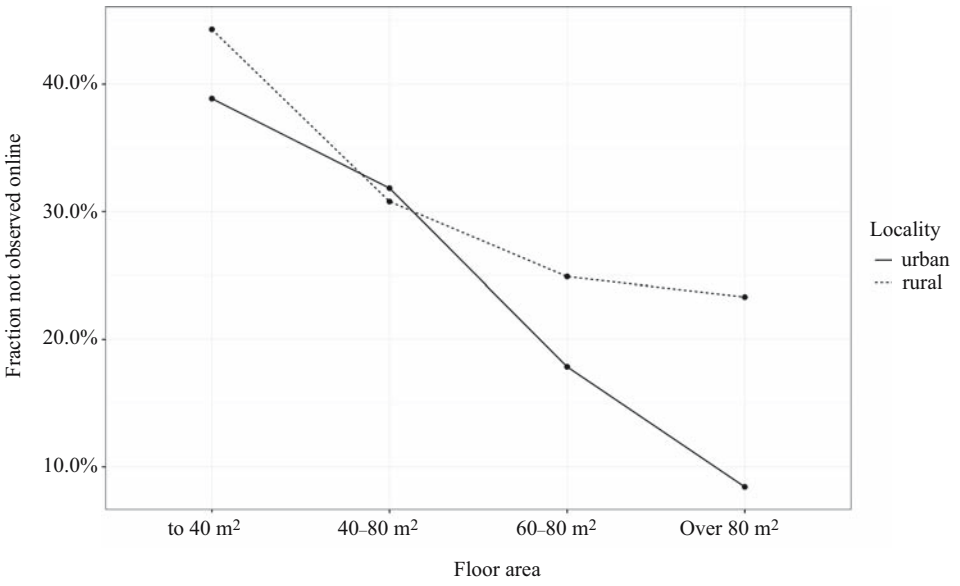

Fig. 3. Share of domains not observed online by locality.

sold without being advertised on two leading Polish portals. This indicates that the left tail of the UFA distribution is underrepresented on the Internet. Moreover, Figure 3 indicates the presence of an interaction between locality and UFA. Properties with floor area over 60 m^2 are more likely to be observed online in urban areas rather than in rural ones.

Figure 4 presents discrepancies in the distribution of `average_pricem2` between domains observed and not observed online. For clarity, `average_pricem2` is presented on a natural log scale and complemented by a rug plot under the density plots to visualize domain observations. The observed shift in the density plots suggests that Internet sources truncate the left tail of the price distribution.

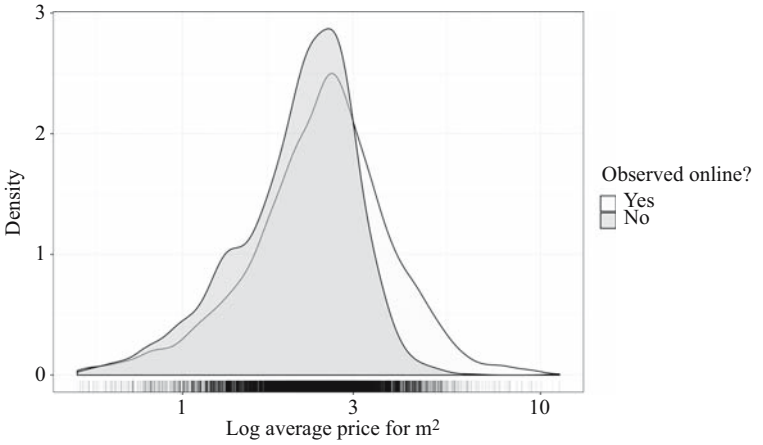

Fig. 4. Distribution of log-transformed average price m^2 depending on whether the domain was observed online or not (selectivity indicator) in 2016.

The results presented in the exploratory data analysis suggest the possibility of a non-ignorable selection mechanism at work in the secondary property market in Poland. In order to verify this hypothesis, we built a model that takes into account multiple covariates to detect the underlying data-creation mechanism.

4.2. *Modeling Results*

Table 2 consists of three parts and presents the summary of four estimated models described in Subsection 3.1. The top part presents estimates of odds ratios and standard errors for fixed effects. The middle part presents standard deviations and standard error for random effects. The bottom part contains three model selection measures.

If an estimate of a fixed effects parameter is larger than one, this means its odds of not being included in Internet data sources are high; if it is less than one, it means that domains with these characteristics are more likely to be observed online. More results regarding the model are presented in the Appendix (Section 6). Figure 8 in the Appendix presents posterior densities of the fixed effects estimated from Model 3 to facilitate the visual analysis of whether the model parameters differ from 0.

The results are in line with the analysis presented in Subsection 4.1. Bigger properties (over 40 m^2) are more often present online in comparison to those up to 40 m^2. The interaction between `urban_rural` and `floor_area` reveals differences between urban and rural areas. Bigger residential properties (over 60 m^2) in rural areas are more frequently absent from the Internet compared to urban areas. Only properties with UFA of 60–80 m^2 in urban areas seem to be equally represented in the online sources and the real estate register. As can be expected, the wider the Internet coverage, the smaller the non-observation propensity.

The parameters for `urban_rural` and `net_coverage` change slightly when random effect for LAU1 is introduced (Model 3 and 4). This is to be expected as these variables are characteristics of LAU1 units.

The random effects component accounts for the informativeness of selectivity measured is by adding `average_pricem2` to Model 1. WAIC and DIC statistics for Model 2 indicate that the average price per m^2 is a non-linear term because it improves Model 1. WAIC drops from 5905.1466 to 5676.7058 and DIC 5905.0123 to 5676.3313. This result suggests that selectivity might be non-ignorable given other characteristics of the real estate market.

However, if we introduce random effect for LAU1 unit rather than for `average_pricem2`, the drop in WAIC and DIC is significantly higher. WAIC drops by 1583.73 (in comparison to Model 1) and DIC decreases by 1559.917 (in comparison to Model 1). This suggests that Model 3 is better than Model 2. The variance component for LAU1 is almost twice as high as for `average_pricem2`.

Further, information criteria for Model 4 indicate that the model with both `average_pricem2` and LAU1 unit performs slightly less well than Model 3. This indicates that the LAU1 effect may account for prices within these units. This hypothesis is supported by Figures 5 and 6.

Figure 5 presents the relationship between the average price per m^2 and the non-observation propensity for Models 2 and 4. For both models we observe a non-linear relationship with the average price, and less expensive properties are more likely not to be observed online in comparison to more expensive properties. However, for Model 4 we

Table 2. Summary of fixed and random parameters and information criteria for the estimated models. For fixed effects only odds ratios are reported.

	Model 1		Model 2		Model 3		Model 4	
Parameters	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Fixed effects (odds ratios)								
Intercept	0.7572	0.0504	0.7311	0.0544	0.4908	0.0737	0.4842	0.0722
net_coverage	0.0168	0.0041	0.0227	0.0057	0.3504	0.1786	0.3321	0.1683
urban_rural: Rural	0.6741	0.0839	0.5607	0.0762	1.1612	0.1999	1.1134	0.2025
floor_area: 40-60	0.7112	0.0672	0.6782	0.0658	0.5526	0.0666	0.5428	0.0657
floor_area: 60-80	0.3267	0.0353	0.2983	0.0330	0.1700	0.0237	0.1660	0.0232
floor_area: over 80	0.1517	0.0242	0.1254	0.0203	0.0888	0.0178	0.0851	0.0172
urban_rural x floor_area interaction								
Rural and 40–60	0.7628	0.1265	0.7393	0.1250	0.6576	0.1348	0.6510	0.1343
Rural and 60–80	1.2024	0.2286	1.2666	0.2357	1.3543	0.3045	1.3560	0.3051
Rural and over 80	2.8316	0.6792	3.2472	0.7900	4.2850	1.2461	4.4505	1.3065
Random effects (σ)								
LAU1	–	–	–	–	2.2368	0.1287	2.1150	0.1254
average_pricem2	–	–	1.2704	0.3437	–	–	1.0382	0.3691
Model selection								
WAIC	5905.1464	–	5677.3259	–	4095.8715	–	4134.2167	–
DIC	5905.0121	–	5676.9841	–	4118.6910	–	4152.6810	–
$\sum \log(\text{CPO}_d)$	−2952.5730	–	−2838.6670	–	−2051.4940	–	−2071.2050	–

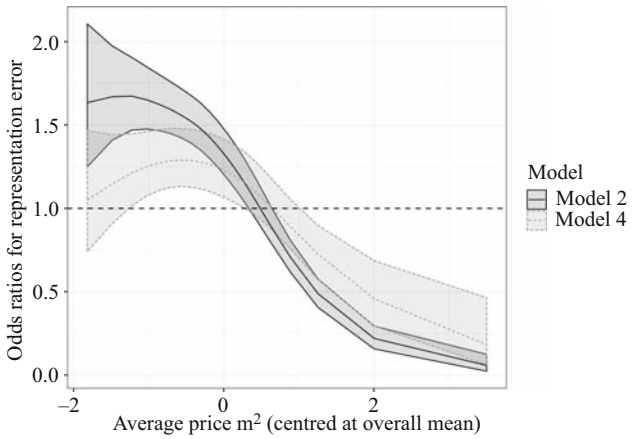

Fig. 5. Point estimates of odds ratios and 95% credible intervals estimated from random walk of order 2 for the average price per m^2 based on Model 2 and 4.

observe a diminishing effect for cheap properties, while for expensive ones, it remains more or less at the same level.

Figure 6 shows the distribution of the LAU 1 random effect for Model 4 in relation to the average price for m^2. This price was calculated as an average price for all domains within each LAU 1 unit. That is why the range of prices is different from that presented in Figure 5. There are several areas where properties, despite their average price, are always observed online (points below the dashed line) but the majority of LAU 1 units are above or close to the overall mean. 120 LAU 1 units (31%) have a credible interval over the dashed line denoting zero, which suggests that some domains within these LAU 1 units are not observed online. This, however, depends largely on the number of categories of floor area, and the time granularity considered.

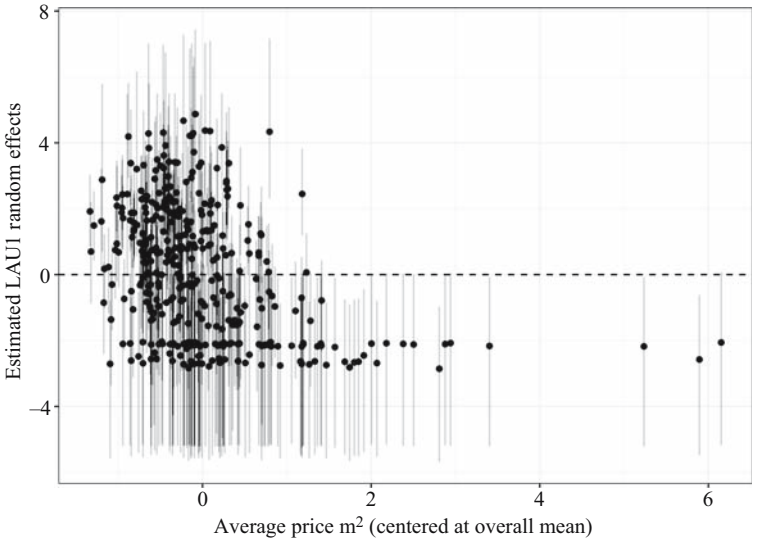

Fig. 6. Point estimates and 95% credible intervals of LAU 1 random effects and average price m^2.

5. Conclusion and Discussion

In the article we studied representation errors in Internet data sources for residential properties in the secondary market in Poland. We used the two biggest online advertising platforms that list real estate offers and one administrative source, which covers all transactions in this market. The auxiliary data source was used to detect correlates of representation errors and determine whether its missingness is non-ignorable.

The results suggest that representation errors are strongly correlated with usable floor area and Internet coverage. As expected, the selection mechanism is connected with the low level of aggregation (LAU 1 level), which is the dominant factor in random effects in the proposed models.

However, results of the estimated models are ambiguous. Based solely on information criteria, errors could be regarded as ignorable; however, when analyzing the relationship between the price and the fact of not being present online, a clear non-linearity is visible. This might also be connected with smaller properties (in terms of UFA) that are also characterized by lower prices.

A number of explanations can be proposed to explain such results. First, despite their size, these portals are mainly used by real estate brokers. Only 5% of offers listed on Otodom are placed by individual customers; data obtained from Gratka API do not contain such information and results in undercoverage. It is likely that brokers are the target group of premium customers because they place ads for more expensive properties. One potential way to overcome this problem is to use other services, which are targeted at different groups of people. In Poland, the OLX classified service can be a good example, as it also lists properties but, according to the OLX group, which owns Otodom and OLX, OLX users mainly include owners and people from rural areas. Second, properties in Poland do not have to be sold online, nor are they officially registered. Transactions involving properties not listed online can take place between family members or specific, small groups of customers.

Even though results are promising and support the research questions stated in the introduction, one should take into account that the study was conducted at domain level, which may have influenced the results. If units listed online could be linked with those included in the register, the analysis of correlates of self-selection error could be more accurate.

The problem of overcoverage regarding (1) duplicated entries, (2) outdated entries, (3) no longer for sale, and (4) false advertisements was not addressed in the data cleaning procedure. This issue cannot be easily tackled and requires additional attention. Therefore, to some (yet unknown) extent, results presented in the article may underestimate effects of the correlates of selectivity.

Keeping that in mind, the methods presented in the article can be used to select an appropriate method of correcting the selection bias. For instance, probabilities estimated on the basis of models described above could be used for propensity score weighting and then applied to online data. Another possible use involves the application of the model-based approach under the missing not at random (MNAR) mechanism to estimate asking prices for domains not covered by the online services. Other possible applications can be found in (Riddles et al. 2016; Sverchkov and Pfeffermann 2018; Sikov 2018; Heckman 1979; Marra et al. 2017).

Finally, the approach presented in the article could be applied to other sources given the availability of auxiliary variables (including proxies), both in these sources and in

independent data (e.g., administrative records, sample surveys). Without access to such covariates, it will not be possible to detect errors or reduce bias. In other words, researchers interested in big data for official statistics should focus on variables that are highly correlated with the target variable.

6. Appendix

Table 3. Distribution of number of transactions (Register) and advertisements (Gratka, Otodom).

Quarter	Source	Total	Min	Q1	Median	Mean	Q3	Max
2016 Q1	Gratka	224 719	0	0	3	119	19	26 079
	Otodom	165 374	0	1	5	87	22	15 344
	Register	–	–	–	–	–	–	–
2016 Q2	Gratka	205 443	0	0	3	112	18	25 937
	Otodom	180 724	0	2	7	98	31	15 247
	Register	27 436	1	2	4	14	11	1 276
2016 Q3	Gratka	193 786	0	0	3	109	18	23 456
	Otodom	174 418	0	2	8	98	31	14 314
	Register	25 428	1	2	4	14	10	1 450
2016 Q4	Gratka	–	–	–	–	–	–	–
	Otodom	–	–	–	–	–	–	–
	Register	21 724	1	2	4	12	9	1 140

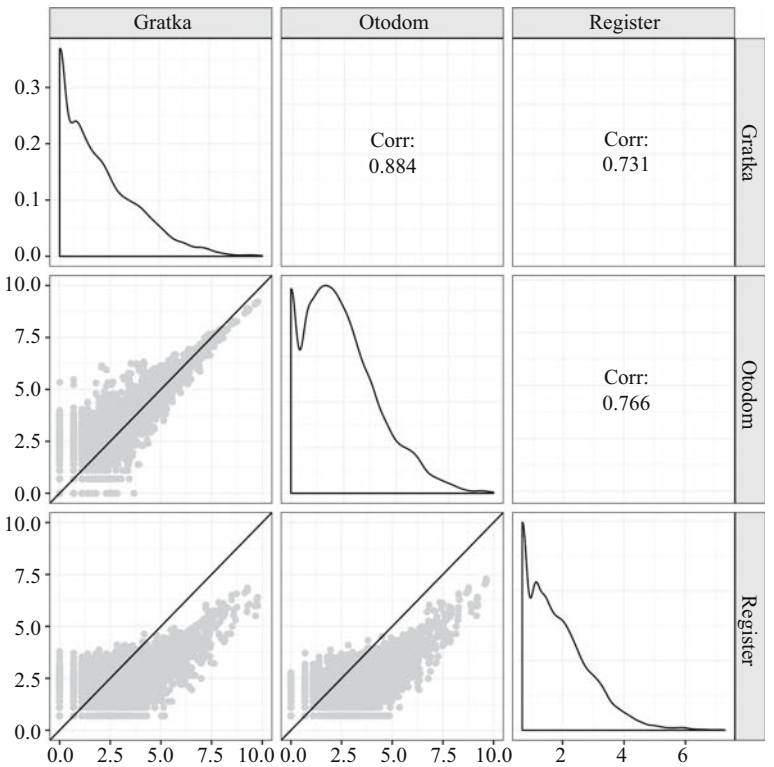

Fig. 7. Correlation of log number of transactions (Register) and advertisements (Gratka, Otodom) at domain level.

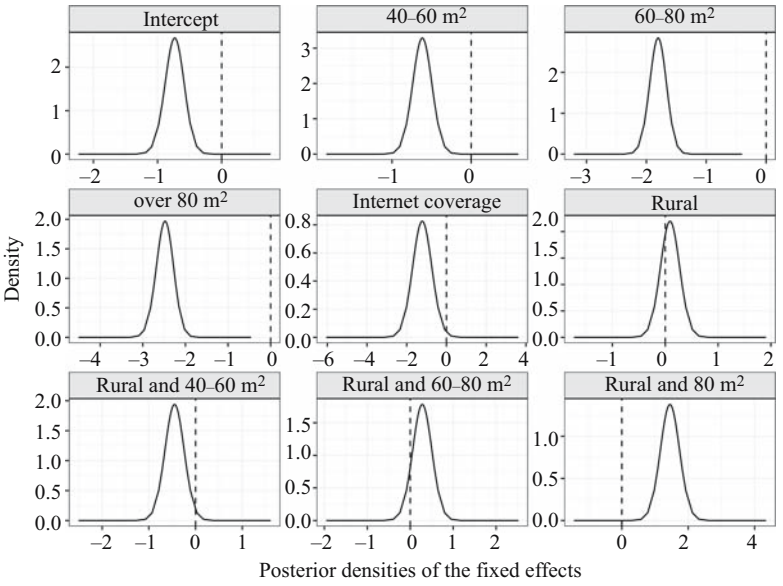

Fig. 8. Posterior densities of the fixed effects estimated under Model 3.

7. References

Anenberg, E. and S. Laufer. 2017. "A More Timely House Price Index." *Review of Economics and Statistics* 99(4): 722–734. Doi: https://doi.org/10.1162/REST_a_00634.

Beręsewicz, M. 2016. *Internet Data Sources for Real Estate Market Statistics*. PhD diss., Poznań University of Economics and Business. Available at: http://www.wbc.poznan.pl/dlibra/docmetadata?id=393454 (accessed February 2019).

Beręsewicz, M. 2017. "A Two-Step Procedure to Measure Representativeness of Internet Data Sources." *International Statistical Review* 85(3): 473–493. Doi: https://doi.org/10.1111/insr.12217.

Beręsewicz, 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. Doi: https://doi.org./10.2785/312232.

Brick, J.M. 2015. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29(3): 329–353. Doi: https://doi.org/10.2478/jos-2013-0026.

Buelens, B., P. Daas, J. Burger, M. Puts, and J. van den Brakel. 2014. *Selectivity of Big Data*. Discussion paper 201411. Statistics Netherlands, The Hague/Heerlen, The Netherlands. Available at: http://pietdaas.nl/beta/pubs/pubs/Selectivity_Buelens.pdf (accessed February 2019).

Cavallo, A. 2013. "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics* 60(2): 152–165. Doi: https://doi.org/10.1016/j.jmoneco.2012.10.002.

Chen, B., A. Shrivastava, and R.C. Steorts. 2018. "Unique entity estimation with application to the Syrian conflict." *The Annals of Applied Statistics* 12(2): 1039–1067. Doi: https://doi.org/10.1214/18-AOAS1163.

Chen, C., J. Wakefield, and T. Lumely. 2014. "The Use of Sampling Weights in Bayesian Hierarchical Models for Small Area Estimation." *Spatial and Spatio-Temporal Epidemiology* 11: 33–43. Doi: https://doi.org/10.1016/j.sste.2014.07.002.

Citro, C.F. 2014. "From Multiple Modes for Surveys to Multiple Data Sources for Estimates." *Survey Methodology* 40(2): 137–161.

Daas, P.J., M.J. Puts, B. Buelens, and P.A. 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.

ESSnet Big Data. 2018. "ESSnet Big Data." Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data (accessed February 2018).

Faraway, J.J., X. Wang, and Y.Y. Ryan. 2018. *Bayesian Regression Modeling with INLA*. Chapman/Hall/CRC.

Fleishman, L. and Y. Gubman. 2015. "Mass Appraisal at the Census Level: Israeli Case." *Statistical Journal of the IAOS* 31(4): 597–612. Doi: https://doi.org/10.3233/SJI-150939.

Gelman, A., J. Hwang, and A. Vehtari. 2014. "Understanding Predictive Information Criteria for Bayesian Models." *Statistics and Computing* 24(6): 997–1016. Doi: https://doi.org/10.1007/s11222-013-9416-2.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153–161. Doi: https://www.jstor.org/stable/1912352.

Held, L., B. Schrödle, and H. Rue. 2010. "Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA." In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, edited by T. Kneib and G. Tutz, 91–110. Heidelberg: Physica-Verlag HD. Doi: https://doi.org/10.1007/978-3-7908-2413-1_6.

Hoekstra, R., O. ten Bosch, and F. Harteveld. 2012. "Automated Data Collection from Web Sources for Official Statistics: First Experiences." *Statistical Journal of the IAOS* 28(3, 4): 99–111. Doi: https://doi.org/10.3233/SJI-2012-0750.

Ihlanfeldt, K.R. and J. Martinez-Vazquez. 1986. "Alternative Value Estimates of Owner-occupied Housing: Evidence on Sample Selection Bias and Systematic Errors." *Journal of Urban Economics* 20(3): 356–369. Doi: https://doi.org/10.1016/0094-1190(86)90025-2.

Japec, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher. 2015. "Big Data in Survey ResearchAAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839–880. Doi: https://dx.doi.org/10.1093/poq/nfv039.

Kiel, K.A. and J.E. Zabel. 1999. "The Accuracy of Owner-provided House Values: The 1978–1991 American Housing Survey." *Real Estate Economics* 27(2): 263–298. Doi: https://doi.org/10.1111/1540-6229.00774.

Lindgren, F. and H. Rue. 2015. "Bayesian Spatial Modelling with R-INLA." *Journal of Statistical Software* 63(19): 1–25. Doi: https://doi.org/10.18637/jss.v063.i19.

Lohr, S.L. and T.E. Raghunathan. 2017. "Combining Survey Data with Other Data Sources." *Statist. Sci.* 32(2) (May): 293–312. Doi: https://doi.org/10.1214/16-STS584.

Lozano-Gracia, N. and L. Anselin. 2012. "Is the Price Right?: Assessing Estimates of Cadastral Values for Bogotá, Colombia." *Regional Science Policy & Practice* 4(4): 495–508. Doi: https://doi.org/10.1111/j.1757-7802.2012.01062.x.

Marra, G., R. Radice, T. Bärnighausen, S.N. Wood, and M.E. McGovern. 2017. "A Simultaneous Equation Approach to Estimating Hiv Prevalence with Nonignorable Missing Responses." *Journal of the American Statistical Association* 112(518): 484–496. Doi: https://doi.org/10.1080/01621459.2016.1224713.

Mercer, L., J. Wakefield, C. Chen, and T. Lumley. 2014. "A Comparison of Spatial Smoothing Methods for Small Area Estimation with Sampling Weights." *Spatial Statistics* 8: 69–85. Doi: https://10.1016/j.spasta.2013.12.001.

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://dx.doi.org/10.1093/jssam/smv035.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: www.R-project.org/ (accessed February 2019).

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics* 33(2): 477–511. Doi: https://doi.org/10.1515/jos-2017-0023.

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.

Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)." *Journal of the Royal Statistical Society B* 71: 319–392. Doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Sikov, A. 2018. "A Brief Review of Approaches to Non-ignorable Non-response." *International Statistical Review* 86(3): 415–441. Doi: https://doi.org/10.1111/insr.12264.

Simpson, D., H. Rue, A. Riebler, T.G. Martins, S.H. Sørbye, et al. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32(1): 1–28. Doi: https://doi.org/10.1214/16-STS576.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. Van Der Linde. 2002. "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4): 583–639. Doi: https://doi.org/10.1111/1467-9868.00353.

Statistics Netherlands. 2018. *Indicatoren bestaande woningen in verkoop*. Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksb eschrijvingen/indicatoren-bestaande-woningen-in-verkoop (accessed November 2018).

Steorts, R.C., R. Hall, and S.E. Fienberg. 2016. "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association* 111(516): 1660–1672. Doi: https://doi.org/10.1080/01621459.2015.1105807.

Sverchkov, M. and D. Pfeffermann. 2018. "Small Area Estimation Under Informative Sampling and Not Missing At Random Non-response." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(4): 981–1008. Doi: https://doi.org/10.1111/rssa.12362.

Wallgren, A. and B. Wallgren. 2014. *Register-based Statistics: Statistical Methods for Administrative Data*. New York: Wiley.

Watanabe, S. 2010. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research* 11(Dec): 3571–3594. Available at: http://www.jmlr.org/papers/v11/watanabe10a.html (accessed February 2019).

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-based Statistics and Data Integration." *Statistica Neerlandica* 66(1): 41–63. Doi: https://doi.org/10.1111/j.1467-9574.2011.00508.x.

Received March 2018
Revised March 2019
Accepted April 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 531–576, http://dx.doi.org/10.2478/JOS-2019-0023

An Integrated Database to Measure Living Standards

Elena Dalla Chiara[1], *Martina Menon*[1], *and Federico Perali*[1]

This study generates an integrated database to measure living standards in Italy using propensity score matching. We follow the recommendations of the Commission on the Measurement of Economic Performance and Social Progress proposing that income, consumption of market goods and nonmarket activities, and wealth, rather than production, should be evaluated jointly in order to appropriately measure material welfare. Our integrated database is similar in design to the one built for the United States by the Levy Economics Institute to measure the multiple dimensions of well-being. In the United States, as is the case for Italy and most European countries, the state does not maintain a unified database to measure household economic well-being, and data sources about income and employment surveys and other surveys on wealth and the use of time have to be statistically matched. The measure of well-being is therefore the result of a multidimensional evaluation process no longer associated with a single indicator, as is usually the case when measuring gross domestic product. The estimation of individual and social welfare, multidimensional poverty and inequality does require an integrated living standard database where information about consumption, income, time use and subjective well-being are jointly available. With this objective in mind, we combine information available in four different surveys: the European Union Statistics on Income and Living Conditions Survey, the Household Budget Survey, the Time Use Survey, and the Household Conditions and Social Capital Survey. We perform three different statistical matching procedures to link the relevant dimensions of living standards contained in each survey and report both the statistical and economic tests carried out to evaluate the quality of the procedure at a high level of detail.

Key words: Propensity score; statistical matching; well-being; fused data; multidimensional poverty.

1. Introduction

In times of recession it is especially important to understand the multidimensional linkages among income, wealth and consumption and how costs and opportunities are distributed across social classes and territories. In France, the Fitoussi Commission (Stiglitz et al. 2010) set up by the French government to identify new tools to measure economic performance and social progress believes that it is now time to shift the attention from the measurement of economic production to the measurement of the well-being of people. To evaluate material welfare, the Commission proposes that income, consumption of both goods and time, and wealth, rather than production, should be evaluated jointly with the aim of broadening the measures traditionally used for family support, including the evaluation of non-market activities. Income or consumption alone cannot comprehensively

[1] University of Verona, Department of Economics, via Cantarane, 24, 37129, Verona, Italy. Emails: elena.dallachiara@univr.it, martina.menon@univr.it, and federico.perali@univr.it

describe a household's standard of living, although consumption inequality often mirrors income inequality (Attanasio et al. 2015). Consumption, defined by total household expenditure, including possibly an imputed income from housing, differs from income because a household can borrow or save, and it should better reflect long-term standard of living and lifetime resources (Slesnick 1993; Blundell and Preston 1995; Meyer and Sullivan 2011; Brewer and O'Dea 2012).

The measure of well-being is therefore the result of a multidimensional evaluation process no longer associated with a single indicator, as is usually the case when measuring gross domestic product. A person's standard of living depends on multidimensional circumstances such as health status, equal access to education, the ability to develop personal relationships, to enjoy a clean environment and to invest in activities creating social capital. The estimation of individual and social welfare, multidimensional poverty and inequality, which is especially important in light of the evaluation of the impact of Horizon 2020, requires an integrated living standards database where information about consumption, income, time use and subjective well-being are jointly available. Similarly, integrated information is necessary to properly model household production, male and female labor supply, the full cost of children, and fertility decisions accounting for the cost of time invested in child care (Caiumi and Perali 2015).

This integrated architecture is also appropriate for identifying the short and long-run actions guaranteeing the well-being of present and future generations as pursued, for example, by the ISTAT (Italian National Institute of Statistics) which, for the years 2013 and 2014, has produced a policy-relevant report on the Equitable and Sustainable Well-Being of Italians (ISTAT 2013, 2014). Integrated databases about living standards are also useful in epidemiological studies because they can serve as controls for case studies designed to capture all relevant quality-of-life dimensions in order to understand the causes of public health problems such as juvenile crime or public-health related aspects. The ecological framework, which is often used to explain why some groups in society are at a higher risk of exposure to public health problems while others are protected, views public "disease" as the outcome of interactions between many factors at four levels – the individual, the relationship, the community, and the societal (Krug et al. 2002).

An integrated database with a design similar to the one described in the present study has been built for the US by the Levy Economics Institute to measure the multiple dimensions of well-being. In the United States, as in Italy and most European countries, the state does not collect a unified database to measure household economic well-being. Hence data sources about income and employment surveys and other surveys on wealth and time use have to be statistically matched to form the Levy Institute Measure of Economic Well-being (LIMEW) database (Wolff and Zacharias 2003; Kum and Masterson 2010; Sharpe et al. 2011; Wolff et al. 2012).

The Living Standard Measurement Studies (LSMS) conducted by the World Bank in most developing countries, on the other hand, have been designed to capture all the dimensions affecting well-being and quality of life and, in most cases, do not need such a composite matching design. In a developing country context, it is more cost and time efficient to carry out an integrated survey rather than a survey specific to each relevant dimension, as is done in most developed countries where a higher level of statistical precision is required.

Our aim is to create an integrated data set to measure living standards that combines information available from different data sources using Italian data as an empirical example. Our main contribution to the literature is to evaluate both the statistical and economic robustness of the fused data. To this end, we show how to perform robust economic tests based on the fundamental Engel relationship verifying the viability of the fused database for economic analysis. We also illustrate the policy potential of the Italian integrated data set by presenting an excerpt of the results of a research measuring multidimensional poverty and of a causal investigation of juvenile crime in Italy. The matched data set contains information collected in four different surveys: the European Union Statistics on Income and Living Conditions survey (henceforth EUSILC), the Household Budget Survey carried out by the Italian National Statistical Institute (henceforth HBS), the Time Use Survey by the Italian National Statistical Institute (henceforth TUS), the Household Conditions and Social Capital survey of the International Center of Family Studies (henceforth CISF). We implement the statistical matching by using a propensity score approach (Rosenbaum and Rubin 1983; Caliendo and Kopeinig 2008). We also investigate uncertainty by calculating the Fréchet inequality for the contingency table associating income and expenditure classes, which is a special concern of the present analysis.

Our findings relate to Italian data. However, both the implementation method, which is rarely applied to the fusion of four data sets, and the evaluation method, adopting both statistical and economic tests of the quality of the matching, are of general interest. The matching performance is comparable with the matching results adopted by the Levy Institute (Kum and Masterson 2010; Masterson 2010, 2014; Wolff et al. 2012; Rios-Avila 2014, 2015, 2016; Albayrak and Masterson 2017) using mainly US and Canadian data, and with Eurostat (Leulescu and Agafitei 2013; Webber and Tonkin 2013). This evidence suggests that if the same method is applied to other EU countries, the performance is likely to be as statistically and economically robust.

This assertion does not imply that this work is exempted from limitations. In absence of auxiliary information, the present application is developed under the conditional independence assumption. We studied the inferential consequences of this assumption by analyzing the uncertainty associated with the lack of joint information about the variables of interest. Another important limitation relates to the matching of complex sample surveys. This aspect is particularly exacerbated when the final integrated database is obtained after more than two linkages. Because of the potential accumulation of sources of imprecision as more surveys are fused mixing data from different clusters and strata, the reliability of the results may be affected. This is a relevant issue that, in our view, deserves greater research attention.

The rest of the article is organized as follows Section 2 describes the methodology to implement statistical matching using the propensity score approach. The single data sets are delineated in Section 3. Section 4 provides a detailed description of the three statistical match procedures and analyzes both the statistical and economic robustness of the outcomes. Section 5 illustrates an empirical application about the measurement of multidimensional poverty in Italy that exploits the fused living standard database. Section 6 summarizes the main findings and draws conclusions that could be useful for future. The supplemental material consists of Tables A1–A18 and Figures A1–A8.

2. The Statistical Matching Method

Statistical matching techniques enable the integration of two or more data sources that refer to the same target population and share a common set of variables. Matching combines information observed in a donor data set, which can also be considered the control group, with units of a recipient data set, which can be considered as the treatment group, with missing values for those variables. The donor data set is the database that contains the extra information and normally includes the largest number of observations. In practice, statistical matching can be seen as a method of variable imputation from a donor to a recipient survey (Rubin and Schenker 1986; D'Orazio et al. 2006a; Kum and Masterson 2010; Tedeschi and Pisano 2013; Donatiello et al. 2014).

Let A and B be two independent samples of size n_A and n_B respectively, drawn from the same population. Variables Y are observed only in A, while variables Z are observed only in B. A set of variables X are collected in both samples and are correlated with both Y and Z. The main goal of statistical matching is to estimate the joint distribution of (Y, Z, X) or at least on the pairs of target variables that are not observed jointly (Y, Z). The relation between these common variables and the specific variables observed only in one of the data sets is used to impute from a donor data set A information on Y in the recipient data set B for similar units and a synthetic dataset is generated with complete information on X, Y and Z representative of the population of interest.

Statistical matching methods can be classified into three broad categories: non-parametric methods such as the constrained or unconstrained hot deck method; regression-based parametric methods; and mixed methods. Hot deck imputation involves replacing missing values with values from a donor unit similar in terms of common characteristics. A hot deck application is random when the donor is selected randomly from a donor pool. The constrained hot deck method ensures that each record in the donor file is used only once to impute the non-observed variables in the recipient file using values really existing in the donor file. Mixed methods involve a combination of parametric and non-parametric techniques in a two-stage process such as the predictive mean matching imputation method or the propensity score matching.

This study adopts the latter approach. Statistical matching is a delicate exercise because of the dimensionality problem related to the high number of shared covariates, the number of possible values of categorical variables, and the presence of continuous variables that can reflect many different values. The propensity score is one possible balancing score that deals with the high dimensionality of the procedure reducing the problem to one-dimension. There are other attractive ways to deal with the dimensionality problem, such as the predictive mean matching (PMM) also when integrated in hot deck matching schemes (Kum and Masterson 2010; Leulescu and Agafitei 2013). The hot deck matching tends to break down when the sample size is small or the set of selecting variables is large, because the pool of potential donors is limited and robust matches are rare (Mittag 2013). Andridge and Little (2010) contend that very little is known about the theoretical properties of hot deck procedures. On the other hand, because the hot deck is a nonparametric technique, it is less exposed to model misspecification.

Rosenbaum and Rubin (1983) proposed the use of *balancing scores* applied to the most relevant observed common variables. The balancing score $b(X)$ is a function of the

observed covariates X such that the conditional distribution of X given $b(X)$ is independent ($\perp$) of assignment in the treatment (D) $D \perp X|b(X)$. Originally, this technique was introduced to estimate causal effects between treated and control groups in non-randomized experiments.

The propensity score is estimated using a logistic or probit regression specified on the selected set of covariates that are common to all questionnaires, and its estimated score can be considered a synthetic indicator of the shared variables used in this function. The propensity score is the conditional probability of assignment to a particular treatment conditional on a set of observed covariates $p(X) = prob(D = 1|X)$, where D is an indicator equal to *1* if an observation refers to the treated group and *0* otherwise.

For a statistically robust application of the propensity score, the assumptions normally made when implementing a statistical matching procedure can be stated in a randomized trial context (Rosenbaum and Rubin 1983):

Conditional independence: given a set of common covariates that are not affected by treatment, the potential outcomes are independent of treatment assignment

$$D \perp Y_0, Y_1|X => D \perp Y_0, Y_1|p(X).$$

Common support: observations with the same covariate values have a positive probability of being both in treated and untreated

$$Y_0, Y_1 \perp D|X.$$

The conditional independence assumption asserts that the outcome in the control group is independent of the treatment D conditional on the selected set of covariates. In the early statistical matching implementations, it was frequent to assume the independence of the never jointly observed variables Y and Z given the set of common variables X, $f(x,y,z) = f_{Y|X}(y|x)f_{Z|X}(z|x)f_X(x)$ where $f_{Y|X}$ is the conditional density function of Y given X, $f_{Z|X}$ is the conditional density function of Z given X and f_X is the marginal density function of X (D'Orazio et al. 2006a). Conditional independence rarely holds in practice. In a statistical matching context where only A and B are available it is not possible to test the conditional independence assumption. Modern applications exploit, when possible, relevant information from an auxiliary data source to overcome the conditional independence assumption (Donatiello et al. 2014) and evaluate the uncertainty associated with the lack of joint information about the variables of interest (Conti et al. 2017).

The common support requirement states that the distribution of observed covariates is as similar as possible in both groups. This assumption ensures that there is an overlap in the characteristics of treated and untreated observations sufficient to have potential matches in the untreated group.

Note that when using the terms treated and control in the context of statistical matching rather than a randomized trial context, we refer to the treated group as the recipient data set and to the control group as the donor data set. This analogy says that the treated group is the recipient of the treatment, that is, the additional information coming from the control (donor) data set that donates information (treats) the recipient. In a multiple matching

exercise, as it is in our application, there are multiple donor data sets contributing information to the single recipient data set.

Another relevant assumption underlying the implementation of a statistical matching procedure is that the processes generating the missing data is missing completely at random (MCAR). There is no systematic relationship between the propensity of missing values and any data, either observed or missing, because missingness is induced by the sampling design (D'Orazio et al. 2006a). In general, ignorability assumes that missing data can be considered as occurring effectively at random, so that the effects of the unobserved, possibly confounding, factors and missing data can be ignored. Strong ignorability (Rosenbaum and Rubin 1983) presumes that the conditional independence assumption holds and that there is common support, or overlap, between the data sets. In most cases, it is difficult to validate the ignorability assumption because statistical matching suffers from the identification problem concerning the association of the variables never jointly observed, given that the variables common to both data sets cannot be estimated from the observed data. This is a general problem that affects all statistical matching procedures, not just the propensity score. The validity of a matching technique concerning the preservation of the true association of the variables never jointly observed depends on the explanatory power of the common variables (Rässler 2002, 2004; Kiesl and Rässler 2009). Given these common variables, the variables not jointly observed can be more or less independent after statistical matching.

For every variable specific to each data set to be fused, the marginal joint cumulative distribution function is bounded by the Fréchet inequality (D'Orazio et al. 2006a,b, 2009, 2017; Kiesl and Rässler 2009; Conti et al. 2012; Conti et al. 2017). The range of these bounds may be used to evaluate the data fusion procedure, although the bounds may not represent a sufficiently stringent interval to be useful in all practical situations. In general, the higher the explanatory power of the common variables and the narrower the bounds of the association, the more reliable are the matching results at all interesting levels of validity. In any event, it is important to recognize that, from the observed data, we are not able to uniquely recover the underlying joint distribution that could have generated the data because of the range indeterminacy.

In Subsection 4.1, we investigate uncertainty stemming from the identification problem associated with the lack of joint information on the variables of interest by calculating the Fréchet inequality for the contingency table associating income and expenditure classes. This is an especially important economic relation not only for the estimation of short-term savings but also for the related measures of well-being, poverty and inequality (Donatiello et al. 2014; Conti et al. 2016; Conti et al. 2017). The distance between bounds is affected by the number of classes and by the elements included in the set of matching variables. Shorter intervals decrease uncertainty and as a consequence increase trust in the conditional independence assumption. It is in this sense that the analysis of uncertainty can be viewed as a measure of the relevance of the conditional independence assumption and the overall quality of the procedure, and as a specification tool for selecting the most appropriate set of matching variables.

The assumption of conditional independence is especially untenable in the case of consumption and income, although conditional independence seems to be an innocent assumption when the matching variables include a reliable proxy for income as auxiliary

information (Singh et al. 1990; Donatiello et al. 2014; Conti et al. 2016; Conti et al. 2017). In the HBS survey, information about aggregate household incomes is recorded in large intervals as it is stated by respondents, while in the EUSILC database it is constructed with a much higher level of detail on all different types of income earned by all household members. Though affected by large measurement errors, it maintains a high correlation with income. Thus, it may serve as reliable auxiliary information (Singh et al. 1993; Coli et al. 2005; Donatiello et al. 2014). Because the income section of the HBS is not available to users that do not belong to ISTAT, we imputed income at the individual level using information from EUSILC and then summed individual incomes to determine household income. As predictors included in the multiple imputation procedure using the predictive mean matching method, we used the variables region, family type, age, gender, education level, occupational status, job, part-time or full-time worker, and the distinction between dependent or self-employed worker. Predicted income was then used as a matching variable and included in the specification of the logistic model estimating the propensity score, where it performed with high explanatory power.

2.1. Implementation of the Statistical Matching Method

We now describe in sequence the steps adopted to implement our statistical matching procedure.

1. **Harmonization of the data sets**. The first step of the matching procedure harmonizes the common variables across data sets by comparing and adjusting the definitions and classifications to make them homogeneous. We also need to choose the best set of "matching variables" observed in both data sets that have a significant relationship with the variables of interest. A correct selection of variables controls for differences within groups because the selected variables need to be independent of the group assignment, thus affecting the outcome but not the exposure. The model specification involves a trade-off between the common support condition and the plausibility of the conditional independence assumption. A parsimonious specification may not affect common support, but may affect the plausibility of conditional independence, while a full specification may give rise to a support problem by affecting the common support condition (Black and Smith 2004; Caliendo and Kopeinig 2008). The main purpose of the propensity score estimate is to balance all covariates, not to define the best selection into groups (Augurzky and Schmidt 2001).
2. **Compare the distribution of *X***. To inspect whether the common variables are independent of sample selection, we compare the marginal and joint distribution in the recipient and donor group by testing the similarity in distribution and calculating the between groups distance using both the absolute difference and Cramer's V test (Sisto 2006; Masterson 2010; Leulescu and Agafitei 2013). Distributions can be also compared using the Hellinger distance. In our context, this measure is always coherent and consistent with Cramer's V test, which is our selected test. Both the Hellinger distance and Cramer's V assume values between 0 and 1. A value close to 0 means that the relationship between the two distributions is weak. For Cramer's V test, the acceptance threshold of weak relationship is 0.15. Before matching, the

common set of variables may have statistically different distributions, but after the implementation of the propensity score matching procedure, the common set of variables should be balanced within the strata.

3. **Estimate the selected statistical matching method** (Propensity Score Matching). The matching variables are then used to estimate the propensity score value. The set of matching variables is specific to each pair-wise matching that we describe in the next sections.
4. **Validate the propensity score procedure** by a) computing balancing tests, and b) checking the overlap and region of common support between the two groups. As summarized by Lee (2013), to validate the result of the selected propensity score specification, four balancing tests are recommended: i) standardized differences proposed by Rosenbaum and Rubin (1985) for evaluating the bias reduction due to the success of the matching procedure, and consequently analysis of the distance in marginal distributions of the common variables; ii) t-tests to evaluate the equality of each covariate mean between the recipient and donor groups (Rosenbaum and Rubin 1985); iii) stratification test for testing the mean differences within strata of the propensity score (Dehejia and Wahba 1999, 2002); iv) Hotelling test or F-test to verify the joint equality of covariate means between the reciepient and donor groups (Smith and Todd 2005).

   The standardized difference was computed as the percentage of the ratio between the difference of sample means in the recipient and donor subsamples and the square root of the average of sample variance in both groups. Following Rosenbaum and Rubin (1985) a standardized difference is "large" if it is greater than 20. We also computed a t-test to verify if the mean of each common variable between the recipient and the donor database is not statistically different before and after the matching.

   The stratification test was developed in two steps. In the first phase the observations were divided into strata. To determine the number of strata, the estimated propensity score was split into ranges provided that its mean within each stratum was not statistically different in the recipient and donor group. In the second step, for each stratum a t-test was performed to test whether the common covariates presented the same distribution in both groups (Dehejia and Wahba 2002; Caliendo and Kopeinig 2008; Garrido et al. 2014). If the t-test is rejected in even only one stratum, then the propensity score model is not well specified and the specification should be corrected until there are no significant differences between the two groups and the conditional independence assumption is more likely to hold (Caliendo and Kopeinig 2008; Lee 2013).

   The Hotelling test is used to jointly test the equality of the means in all covariates used in propensity score specification, between the recipient and donor data set. If the null hypothesis is rejected, there is no balance in covariates between the two data sets. This test is adopted in multivariate tests of hypotheses and it is the generalization on the t-test used in univariate problems.

   To assess whether the characteristics observed in the recipient group are also observed in the donor group, it is important to verify the overlap of the region of common support of the propensity score value between these two groups (Lechner 2008). This investigation is crucial because the lack of common support may lead

to biased results since the donor group may not be sufficiently similar to the recipient one. A graphical analysis of the density distributions of the propensity score in the recipient and donor group permits a visual inspection of the range and shape of the propensity score distributions (Caliendo and Kopeing 2008). The estimated propensity score is then used to match each individual in the recipient group to an individual in the donor group.

5. **Choose the matching algorithm**. Rodgers (1984) distinguishes between the constrained and unconstrained algorithm types. An unconstrained method imposes no restrictions on the number of times a donor unit may be imputed because it takes simple random samples with replacement. It has the advantage of permitting the closest possible match to each record at the cost of increasing the sample variance of the estimators (Rodgers 1984; Rässler 2002; Kum and Masterson 2010). The distributions of the imputed variables are therefore more likely to represent the empirical marginal or conditional distributions of the selected sample, rather than the ones observed in the original donor file. Despite this disadvantage, unconstrained matching is still the method most frequently used (Rodgers 1984; Kum and Masterson 2010). On the other hand, a disadvantage of the constrained method is that the average distance between the recipient and donor values of the matching variables is plausibly larger, and sometimes unacceptably larger, than in the unconstrained case because matching is implemented without replacement. It is important to remark that the use of sampling weights make sure that donor records can be matched to more than one recipient and vice versa. From a practical point of view, constrained matching is computationally more demanding than unconstrained matching.

   The main matching algorithms are nearest neighbor, caliper and radius, stratification and interval, kernel and local linear, and weighting (Chen and Shao 2000; Caliendo and Kopeinig 2008; Kum and Masterson 2010). The choice in regard to performing a matching with or without replacement and the number of comparison units involves a trade-off between bias and variance. The two aspects are inter-related because, for example, a matching with replacement and a smaller number of comparison units reduces both the bias and the precision (Dehejia and Wahba 2002). All methods yield similar results with large samples, while the trade-off between bias and variance is mainly relevant for small samples. As a result, there is not a better matching algorithm, but its choice should be evaluated case-by-case on the data structure (Caliendo and Kopeinig 2008). We compared different matching algorithms. Our preferred choice was the nearest neighbor algorithm with replacement and one comparison unit because it was the most effective algorithm in preserving the distribution of the donor data set as it is described in Subsection 4.1. For each individual of the recipient database we selected the individual in the donor database with the closest distance in terms of propensity score. The matching algorithm imputed the missing values of the recipient sample using the information from the donor sample.
6. **Assess the statistical matching quality** by a) inspecting distributions, b) analyzing the trend of the imputed variables by the set of X covariates comparing the ratio of mean and median in the two groups, and c) performing uncertainty analysis by

computing Fréchet Bounds between the variables of interest. Rässler (2002) describes four levels of validity to evaluate a matching procedure: preserving individual values, preserving joint distributions, preserving correlation structures, and preserving marginal distributions. In most cases, only the last level, which establishes a minimum validity requirement, can be verified, although recent literature shows that both the preservation of the joint distribution and the preservation of the correlation structure can be evaluated (Conti et al. 2016; Conti et al. 2017). Statistical matching can be considered successful if the marginal and the joint distribution of the covariates and the imputed information show similar trends in the original and the synthetic databases. We assessed the matching procedure by both inspecting the distributions of the extra information in the two databases, and comparing the distribution of the imputed covariates by the set of common variables used in the propensity function, computing the ratio of mean and the ratio of median (Kum and Masterson 2010; Webber and Tonkin 2013). The ratio of mean (median) is calculated as the ratio between the mean (median) of the recipient data set and the mean (median) of the donor data set. To demonstrate whether the two groups are different in the means or medians, we consider the distance of the ratio from 100, being the value that represents the perfect similarity in the means or medians of the two groups. There is no defined threshold to establish if the imputed information in the two samples can be considered comparable, but the closer the ratio is to 100, the greater the similarity of the extra information.

Further, as part of the statistical evaluation, it is important to deal with the source of indeterminacy stemming from the conditional independence assumption and improve the overall quality of the procedure by exploring the degree of uncertainty associated with the matching results, as we did in our empirical application, and possibly exploiting auxiliary information when available, or introducing meaningful logical constraints (D'Orazio et al. 2006b; Conti et al. 2016; Conti et al. 2017).

7. **Assess the economic matching quality** using Engel curves, poverty and inequality analysis.

3. Data Sets Description

In the following section, we briefly describe the four surveys used in this work. Subsequently, we analyze the characteristics and properties of each statistical matching performed.

The implementation of the Stiglitz et al. (2010) proposal to measure well-being in a comprehensive manner based on an extended notion of income that accounts for the value of private and public consumption, working and nonworking time, financial and social assets, requires the integration of several sources of information about households. We now describe the data sets related to the consumption, income and wealth, time use and social dimensions that we combined to construct a multidimensional measure of economic well-being representative of the Italian population. This objective requires adopting a matching procedure that is careful to preserve at least the marginal distribution of the main

economic and social variables of interest, paying especial attention to the varying sampling designs of each data set.

3.1. European Union Statistics on the Income and Living Conditions Survey (EUSILC): The Recipient Survey

EUSILC is an annual statistical survey that gathers comparable cross-sectional and longitudinal data for the EU Member States. In Italy, the National Statistical Institute (ISTAT) conducts the survey. The EUSILC sample is drawn with a two-stage sampling design where primary units are municipalities and secondary units are households. A sample of 760 municipalities is selected, according to a conditional Poisson design with inclusion probabilities proportional to demographic sizes within strata. From each selected municipality, households are drawn by simple random sampling. We use the 2010 sample of 19,147 households corresponding to 47,551 individuals. The sampled households are selected with a rotational design where a fraction of the sample of the previous survey is dropped and replaced with a new sample of equal size maintaining the same representativeness of the whole population. The survey collects information on incomes, wealth and living conditions at both the household and individual levels. EUSILC also gives detailed information on socio-demographic characteristics, housing conditions, health and education, employment status, economic activity and other firm-specific attributes.

3.2. Household Budget Survey (HBS)

The ISTAT consumption survey collects detailed information on household expenditure on goods and consumer services in diverse categories, such as foodstuffs, clothing, housing, transport, education, health and holidays. Expenditures in the HBS are classified using the United Nations' five-digit Classification of Individual Consumption According to Purpose (COICOP) classifications. The main aim of this survey is to analyze and evaluate the trend in household expenditure in relation to the socio-demographic characteristics of family members. We used the data collected in 2009. The HBS sample is drawn with a two-stage sampling design. The primary sampling units are municipalities. They amount to around 470 selected among two groups according to a conditional Poisson design with inclusion probabilities proportional to demographic sizes within strata. From each selected municipality, households are drawn by simple random sampling. The sample is composed of 23,005 households.

3.3. Time Use Survey (TUS)

The TUS records the time employed in daily activities by each household member. The respondent keeps a diary reporting the main activity undertaken, any other activity taking place at the same time, and the places in which the activities are carried out. Each family, selected according to a random procedure, compiles a diary for either one day of the week, Saturday or Sunday according to the day of the visit. To implement the matching procedure, we first imputed the time spent on each activity for those days that the household member did not have to fill in the diary. The TUS also reports on socio-demographic characteristics, education, economic activity, housing, and health conditions.

The sampling design is implemented in two stages. The first stage units are municipalities (508) and the second stage units are households. The interviewees are each family member aged three or over. The 2008–2009 cross-sectional wave interviewed 18,250 households and 44,606 individuals.

3.4. Household Conditions and Social Capital Survey (CISF)

The survey on household conditions and social capital was designed by the International Center of Family Studies (CISF) in 2009 with the aim of describing the well-being of Italian families and their stock of social capital. The survey was carried out through telephone interviews by COESIS. It collects household level data about socio-demographic characteristics, income and overall economic condition, and a detailed set of questions on social capital and relational well-being. The sampling design is stratified by geographic areas and family types. The sample includes 4,017 households and has both national and macro-regional representativeness. Unlike the others, the CISF survey is not scheduled with regular frequency. It is the only survey not implemented by the Italian National Statistical Institute (ISTAT) included in our integrated data set.

In general, multi-stage cluster and stratified sampling are two distinctive features of complex surveys such as those used in this statistical matching exercise. As a consequence, observations cannot be assumed to be independent and do not have equal probability of being selected, as is the case of simple surveys. Observations that are from the same cluster or strata are likely to be more similar to each other. Ignoring the sampling design may introduce serious bias in both the imputation method and the outcome models. Several authors (D'Orazio et al. 2006a; Ridgeway et al. 2015; Conti et al. 2016; Austin et al. 2018) have analysed how to account properly for complex designs and the different survey weights when implementing a statistical matching procedure, placing especial emphasis on Renssen's two-step procedure (Renssen 1998) based on calibration of the weights and Rubin's (Rubin 1986) file concatenation.

In our analysis, we minimized the adding complexity of different survey designs by selecting the three main surveys to be matched (EUSILC, HBS, and TUS) from the same statistical institute. Instead of integrating the income information from EUSILC, we could have selected the Survey on Household Income and Wealth (SHIW) that is conducted by Banca d'Italia every two years. The SHIW survey, which is part of the Household Finance and Consumption Survey of the European Central Bank, is also conducted in two stages. Municipalities with more than 40,000 inhabitants are all included in the sample, while smaller primary units are selected using a probability sampling scheme proportional to size. Secondary sampling units are then selected by simple random sampling. On average, the sample comprises about 8,000 households (20,000 individuals) distributed across around 300 Italian municipalities. The SHIW size of both primary and secondary units is about 1/3 of the HBS size. Conti et al. (2016, Table 2) show that the estimated proportions of households, conditional on two main design variables such as macro-region and household size are not significantly different between EUSILC and HBS. In the context of the present application, this is also the case for all the ISTAT data bases EUSILC, HBS and TUS. There are no significant differences also for the CISF database, which is not produced by ISTAT. Therefore, with

the intent of not adding complexity, we preferred EUSILC to SHIW even though we recognize that SHIW is interesting for the higher value of the information on the value of assets, debts and regular savings with respect to EUSILC. Our choice was also due to the consideration that ISTAT is actively committed to improving the *ex-ante* harmonization of the EUSILC, HBS, and TUS social surveys and in complementing the wealth dimension as part of the revision process under development within the new European Framework Regulation on Social Statistics. Moreover, in recent years, the Italian version of EUSILC has consistently made use of registered data that cross-verify the income data collected through surveys using available social security and tax records.

As part of our specification strategy of the propensity score regression, in the set of matching variables we included some relevant variables of the sampling design such as regions and household characteristics. According to Kum and Masterson (2010), the propensity score matching method's dimensionality reduction is effective in minimizing the potential bias that may stem from the complex designs of the fused data sets.

Figure 1 illustrates how consumption, time use, and social capital donor data sets have been linked to the income and wealth survey. The donor data sets include the extra information missing in the recipient database. The recipient data set contains the most detailed and accurate information about common variables gathered in all surveys. Combining these relevant dimensions of well-being yields a "new" database, to which we refer as the Italian Integrated Living Standard survey (IILS).

To respect the temporal correspondence between income and related variables, we used the 2010 cross-sectional wave for the EUSILC survey because the information on income refers to the previous reference period. We used the 2009 cross-sectional wave for the HBS and the 2008–2009 wave for the TUS.

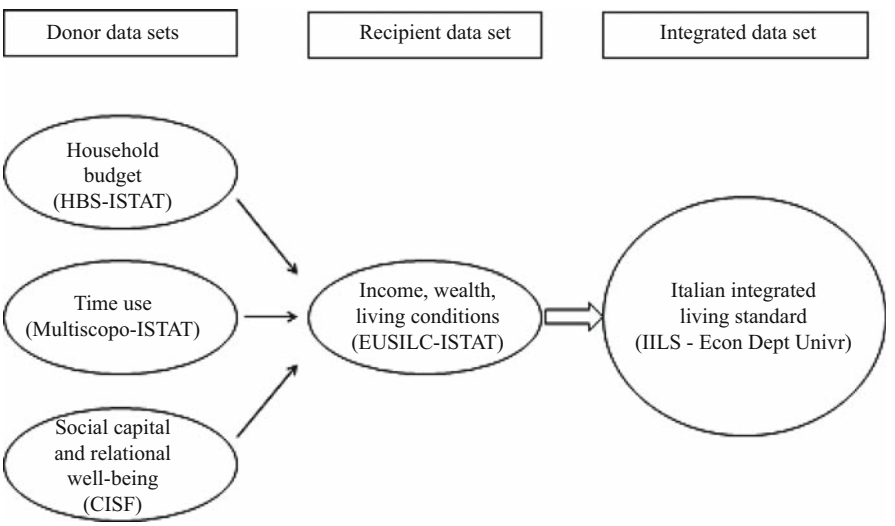

Fig. 1. The data sets used to create the integrated database.

In the next section, we describe the features of each one-to-one matching implemented following the sequential representation of Figure 1 and evaluate the statistical and economic quality of the linking procedure.

4. Results of the Statistical Matching Procedures

We implemented three different statistical matching procedures using the EUSILC data set as the recipient sample because this survey includes the most detailed information regarding socio-demographic characteristics, household conditions, occupational status, income, wealth, health and education.

In sequence, the first linking procedure performed the data fusion between the EUSILC and HBS data set to impute the information related to household consumption. The second statistical matching associated the information about household time use with the EUSILC data set. The third matching filled in the missing values of the EUSILC data for social capital, family relationships and family well-being, using the CISF survey.

The three data fusions were implemented using the method outlined in Section 2. For illustrative convenience, we report the EUSILC-HBS match only. For this matching, we describe a) the alignment of common variables, b) their frequency distributions, c) the standardized differences and t balancing tests, d) the distribution of propensity score, e) the distribution of the extra information imputed with the propensity score procedure, in the original and matched data sets, f) their ratio of mean and median by covariates, and g) the investigation of uncertainty constructing the Fréchet bound, and implement an economic evaluation of the statistical procedure. The results of the EUSILC-TUS and EUSILC-CISF matching procedure are reported in the supplemental material.

4.1. Data Fusion Between the EUSILC and HBS

The EUSILC database does not record data about family consumption that is typically collected in household budget surveys. As shown in Figure 1, we add household consumption to the former survey. We aggregate detailed household expenditures into nine categories: cereals; meat, fish and dairy products; fruit and vegetables; other food products; clothing; housing; transport and communication; recreation and education; health and hygiene.

The two basic conditions for implementation of the statistical matching are satisfied. Both samples refer to the same target population and share a set of covariates related to socio-demographic characteristics, household characteristics and working status conditions. The common variables are defined in the same way in both surveys. Table 1 documents how we harmonized and aggregated the variables of major interest to achieve the same alignment omitting trivial reclassifications.

The adopted propensity score specification that satisfied the balance property includes: region of residence (five dummies coded as North-West, North-East, Center, South, Islands), a dummy variable to indicate the presence in the family of children between 0–5 years old, and between 6–14 years old, a dummy variable to denote the presence in the family of at least one self-employed worker, a single-parent dummy, home-ownership (dummy variable equal to 1 if the household head is a home-owner), average family

Table 1. Alignment of common variables in EUSILC – HBS match.

Variable	EUSILC	HBS	Harmonized variable
Education	*istr_c* 1 = unqualified, illiterate 2 = unqualified, can read and write 3 = primary school 4 = first grade secondary school 5 = second grade secondary school (2–3 years) 6 = second grade secondary school (4–5 years) 7 = certificate post-A levels 8 = bachelor's degree or master's degree 9 = superior graduate school 10 = Ph.D.	*titstu* 1 = Ph.D or superior graduate school 2 = master's degree 3 = bachelor's degree 4 = second grade secondary school (4–5 years) 5 = second grade secondary school (2–3 years) 6 = first grade secondary school 7 = primary school 8 = unqualified	1 = unqualified 2 = primary school 3 = first grade secondary school 4 = second grade secondary school (2–3 years) 5 = second grade secondary school (4–5 years) or certificate post-A levels 6 = bachelor's degree or master's degree 7 = superior graduate school or Ph.D.
Status in employment	*pl040* 1 = self-employed with employees 2 = self-employed without employees 3 = employee 4 = family worker	*posprof* 1 = executive 2 = manager 3 = clerk 4 = intermediate categories 5 = foreman 6 = other employee 7 = trainee 8 = homeworker 9 = military force (armed force) 10 = entrepreneur 11 = self-employed 12 = independent contractor	1 = self-employed 0 = other

Table 1. Continued.

Variable	EUSILC	HBS	Harmonized variable
		13 = partner of cooperatives 14 = assistant 15 = project worker 16 = occasional contractor (from code 1 to code 9 employee, from code 10 to code 16 self-employed)	
Tenure status of the house	*hh020* 1 = owner 2 = tenant or subtenant paying rent at prevailing or market rate 3 = accommodation is rented at a reduced rate (lower price that the market price) 4 = accommodation is provided free	*tipoccup* 1 = tenant or subtenant paying rate 2 = owner 3 = accommodation is in usufruct 4 = accommodation is provided free by relatives or friends	1 = owner 0 = other

Note: The name of the variables in each survey are indicated in italics.

education (five dummies coded as Primary, Middle, Middle-High, High, University) and total disposable household income.

Table 2 shows the frequency distribution of the variables used in the propensity score specification. Geographical area shows the largest absolute differences. The value of Cramer's V test supports the hypothesis that the common variables are independent of the group assignment. Therefore, considering a threshold of 0.15 associated with a weak

Table 2. Comparison between frequency distributions for some common variables.

	EUSILC	HBS	Absolute difference	Cramer's V*
Geographical area				0.094
North-West	23.03	23.58	0.55	
North-East	24.04	21.15	2.89	
Center	22.97	17.62	5.35	
South	21.36	26.61	5.25	
Islands	8.60	11.04	2.44	
Children 0–5 years old				0.020
No	88.53	89.75	1.22	
Yes	11.47	10.25	1.22	
Children 6–14 years old				0.011
No	84.01	83.18	0.83	
Yes	15.99	16.82	0.83	
Self-employed				0.005
No	80.51	80.15	0.36	
Yes	19.49	19.85	0.36	
Single-parent				0.025
No	91.41	92.78	1.37	
Yes	8.59	7.22	1.37	
Homeownership				0.009
No	25.50	24.72	0.78	
Yes	74.50	75.28	0.78	
Average family education				0.036
Primary	26.95	26.83	0.12	
Middle	24.28	27.24	2.96	
Middle-High	19.16	18.15	1.01	
High	23.16	21.48	1.68	
University	6.44	6.29	0.15	
Household income				0.025
1st quintile	19.51	20.41	0.90	
2nd quintile	19.48	20.44	0.96	
3rd quintile	20.17	19.85	0.32	
4th quintile	19.85	20.13	0.28	
5th quintile	20.99	19.17	1.82	

∗The acceptance threshold of a weak relationship is 0.15.

Table 3. Test for standardized differences and t-test on the equality of means.

Variable	Test for Standardized differences		T-test	
	Standardized difference before matching	Standardized difference after matching	P-value before matching	P-value after matching
Geographical area				
North-West	−1.30	0.00	0.1820	0.9750
North-East	6.90	−1.70	0.0000	0.1650
Center	13.30	2.00	0.0000	0.0880
South	−12.30	−0.20	0.0000	0.8980
Islands	−8.20	−0.20	0.0000	0.8510
Children 0–5 years old	3.90	−2.80	0.0000	0.0180
Children 6–14 years old	−2.30	−1.50	0.0220	0.1990
Self-employed	−0.90	−1.00	0.3560	0.3900
Single-parent	5.10	−0.80	0.0000	0.4930
Homeownership	−1.80	0.40	0.0640	0.7180
Average family education				
Primary	0.30	4.10	0.7790	0.0010
Middle	−6.80	−1.40	0.0000	0.2500
Middle-High	2.60	−1.80	0.0080	0.1240
High	4.00	−1.50	0.0000	0.2030
University	0.60	0.60	0.5280	0.6030
Household income	7.60	−1.60	0.0000	0.1850

relationship, we can conclude that all these variables are independent of the groups. This conclusion is generally supported by the evidence presented in Table 3. Before matching, all standardized differences between recipient and donor groups were less than 20%, indicating that the two data sets are similar. The magnitude of these differences decreased after matching, becoming very close to zero. We use the test of standardized differences to illustrate the reduction in bias that can be attributed to matching on common variables (Rosenbaum and Rubin 1985; Lee 2013). Table 3 also shows the p-values of the t-test to compare the means of the common variables. As pointed out by Rosenbaum and Rubin (1985) and Caliendo and Koepinig (2008), it is reasonable to expect differences before the matching execution. After matching, the covariates should be balanced in both groups and hence no significant differences should be found, as is the case in Table 3. In general, the balance in covariates is less likely to be achieved by covariates that do not significantly impact the outcome (Garrido et al. 2014). Before matching, there are many covariates that do not have the same proportion, but after matching the proportions in the recipient and donor groups become equal. The sole exception is represented by the "Primary" category of education, which is balanced before matching but after matching does not show the same mean in the two samples. The Hotelling test also confirms that the covariates are balanced between the two groups. The null hypothesis of joint equality of the means is not rejected (Table 4).

These statements are supported by the evidence presented in Table 5. Conditioning on the propensity score, all variables are balanced within the two samples. The upper part of the table shows t-test values verifying whether the density distributions of the propensity score are equal in the two selected samples within each stratum. The lower part shows the

Table 4. Hotelling test after matching.

Variable	Mean of HBS	Mean of EUSILC
Geographical area		
North-West	0.230	0.230
North-East	0.248	0.240
Center	0.221	0.230
South	0.214	0.214
Islands	0.087	0.086
Children 0–5 years old	0.124	0.115
Children 6–14 years old	0.166	0.160
Self-employed	0.199	0.195
Single-parent	0.088	0.086
Homeownership	0.743	0.745
Average family education		
Primary	0.251	0.269
Middle	0.249	0.243
Middle-High	0.199	0.192
High	0.238	0.232
University	0.063	0.064
Household income	3.197	3.158
Hotelling p-value	0.069	

Table 5. Test statistic of t-test in each stratum.

	Stratum										
	2	3	4	5	6	7	8	9	10	11	12
Balance of propensity score distribution across recipient and donor groups	−0.077	1.368	1.332	−0.076	−1.218	−2.238	−0.654	−0.425	−1.980	−0.593	−1.381
Balance of covariates across recipient and donor groups											
Geographical area – ref. cat. "North-West"											
North-East	−1.287	−1.200	0.747	−1.346	0.336	0.681	2.501	0.279	1.127	−0.586	−0.775
Center	–	–	−0.984	0.576	−0.330	−0.104	−2.117	−0.254	−0.719	−0.840	−0.775
South	−0.987	2.089	−2.197	0.194	−1.116	−0.561	−2.189	–	−0.873	–	–
Islands	1.471	−1.586	−0.961	−0.169	−0.731	−0.589	0.033	–	−0.873	−0.622	–
Children between 0–5 years old	0.591	−0.539	−0.717	−0.429	0.036	0.884	0.483	−1.016	1.029	1.066	−0.775
Children between 6–14 years old	0.842	0.284	0.746	0.081	−0.563	0.813	−1.892	−1.766	0.331	−1.568	−0.775
Self-employed	1.210	0.907	0.343	1.842	−1.040	−0.768	−0.226	−2.346	−0.877	−1.031	–
Single-parent	−0.279	2.350	0.287	0.627	−1.750	−0.062	−0.050	0.357	−0.100	−0.048	–
Homeownership	1.121	−1.716	2.048	0.118	−1.077	2.482	−1.434	−0.905	−1.465	−2.269	–
Average family education -ref. cat. "Primary"											
Middle	−0.253	0.823	1.416	0.226	−0.529	1.156	0.550	−0.329	0.480	−1.090	–
Middle-High	−2.494	1.061	−0.595	1.186	0.733	−0.961	1.547	−0.014	−0.117	−0.992	−0.775
High	0.685	1.555	−1.714	1.661	−1.370	−0.612	0.042	0.601	−1.080	0.079	−1.549
University	−0.007	−2.082	−0.495	0.456	0.388	0.774	−0.636	0.276	1.533	0.798	–
Household income	2.326	1.124	−0.092	2.256	−0.981	0.050	−0.217	−1.203	−1.518	−0.508	−0.793

t-test values carried out to determine whether the common covariates have the same distributions in the two data sets. The first stratum is not shown because the propensity score takes values higher than the first quintile into which the sample was initially divided. Considering a 0.01 significance level, the propensity score and the common covariates have the same distribution in the two samples.

We also performed a preliminary test to investigate the region of common support of the propensity score value. As shown in Figure 2, the estimated propensity score takes values in a similar range and displays comparable density distributions. Therefore, the observations have the same probability of belonging to the recipient or the donor group.

In addition, we implemented a comparative analysis of different matching algorithms such as radius, caliper, Epanechnikov and Gaussian kernels, nearest neighbor with and without replacement and multiple comparison units. The results of all algorithms are consistent in mean but they differ in distribution. Extra information imputed using radius, caliper and both kernel matching algorithms produce mean and median values that are similar to the same statistics of the original distribution, but standard deviations are significantly smaller compared to the original variables. On the other hand, the distribution of imputed values generated using the nearest neighbor algorithm is the most similar to the donor's distribution with and without replacement, and with different comparison units. Table 6 reports these results for the three main consumption categories: cereals, protein foods such as meat, fish and dairy products, and clothing. When adopting one comparison unit, there are no significant differences between distributions with and without replacement. As the number of comparison units increase, differences become more marked, especially in terms of standard deviations. In light of these results, for our matching exercise we adopt the nearest neighbor algorithm with replacement and one comparison unit.

To verify the matching quality, we analyzed the distribution of the extra information transferred from the donor to the recipient. We tested whether the extra information in the matched data set preserves the same distribution as the original data set. We also compared the distributions of the covariates used in the propensity score specification by computing

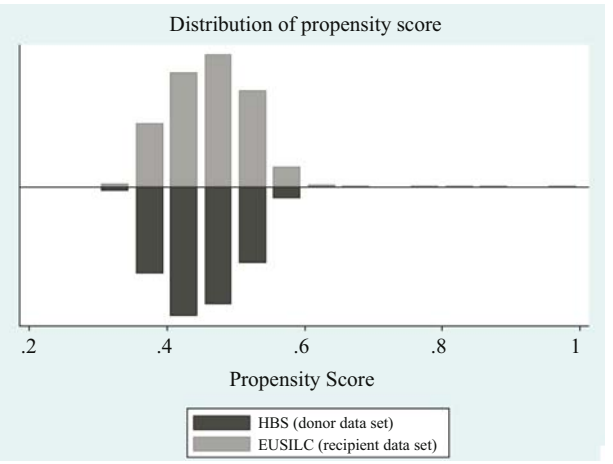

Fig. 2. Distribution of propensity score across recipient and donor data sets.

Table 6. Comparison of statistic distributions using different nearest neighbor matching algorithms (in euros).

		Nearest neighbor					
			With replacement and (*1*,....,*5*) comparison units				
	Donor database	Without replacement and 1 comparison unit	1	2	3	4	5
Cereals							
Mean	69.2	69.72	70.6	70.37	70.25	70.03	70.18
Median	59.28	59.69	60.33	64.91	66.44	66.93	67.95
Std. Dev.	46.1	46.43	46.96	33.26	27.19	23.45	20.92
Min	1.37	1.37	1.7	4.58	12.25	11.29	15.55
Max	469	469	469	291.97	222.27	203.22	194.14
Meat, fish and dairy products							
Mean	213.41	213.33	216.38	215.57	215.46	214.95	215.22
Median	181.53	181.46	184.52	196.71	203.8	205.1	207.5
Std. Dev.	142.5	147.08	147.93	106.23	86.77	75.36	68.18
Min	2.39	2.39	2.64	8.28	33.42	33.62	45.43
Max	1522.19	1522.19	1522.19	956.02	736.78	611.66	576.99
Clothing							
Mean	149.78	151.5	154.59	153.96	153.32	153.1	153.1
Median	142.53	144.88	147.98	151.36	152.02	151.52	151.82
Std. Dev.	76.03	76.04	78.96	56.47	47.25	41.87	38.12
Min	31.78	31.78	33.1	40.41	41.46	44.42	46.8
Max	2563.31	2563.31	2563.31	1349.77	977.84	747.98	636.32

the ratio of mean and the ratio of median. The ratio of median is not reported for the other two data fusions because the meaning of their imputed variables does not fit well since most of their values are concentrated in a single point of mass. Median is an indicator more robust for skewed distributions, but in this context mean is the more appropriate tool with which to evaluate the quality of the matching. A less accurate imputation can preserve the same central tendency between the two databases, but when imputed values are very different from those recorded in the donor data set, it is more difficult to preserve the average value since the mean is largely influenced by outliers.

The distributions of all categories of expenditure are very close to each other, showing that the matching procedure reproduced the same distribution as the original data set. For illustrative purposes, in Figures 3 and 4, we report only the distributions of the four main categories of expenditure and in Figure 5 we report the total household expenditure without disaggregations. This evidence is not sufficient to characterize the quality of the matching outcome completely. It is also necessary to inspect the marginal distribution of imputed variables by variables used to estimate the propensity score value and to compute the matching algorithm. Tables 7–8 and Tables A1–A3 in Supplemental material report the means and the medians of the extra information in the integrated and donor data sets and their ratio by the covariates used to estimate these values. These results show that the synthetic database well preserves the marginal empirical distribution of the common variables in the donor data set. Consequently, the original and matched groups are statistically similar. The lowest income category records the highest difference in mean and median between the two samples. Other discrepancies arise in the presence of children 0–5

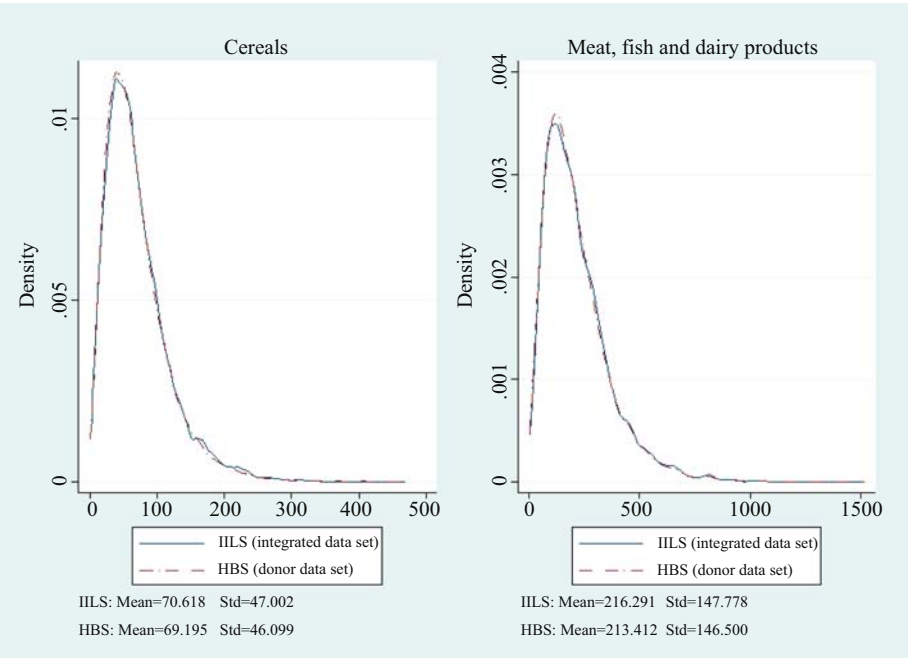

Fig. 3. Distribution of expenditure for "Cereals" and "Meat, fish and dairy products" in integrated and donor data sets.

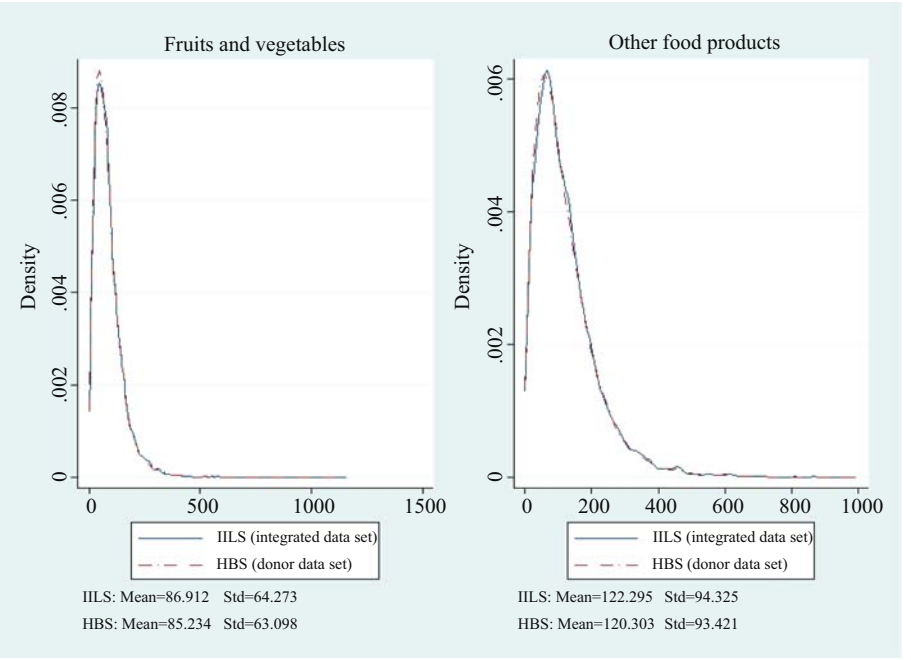

Fig. 4. Distribution of expenditure for "Fruits and vegetables" and "Other food products" in integrated and donor data sets.

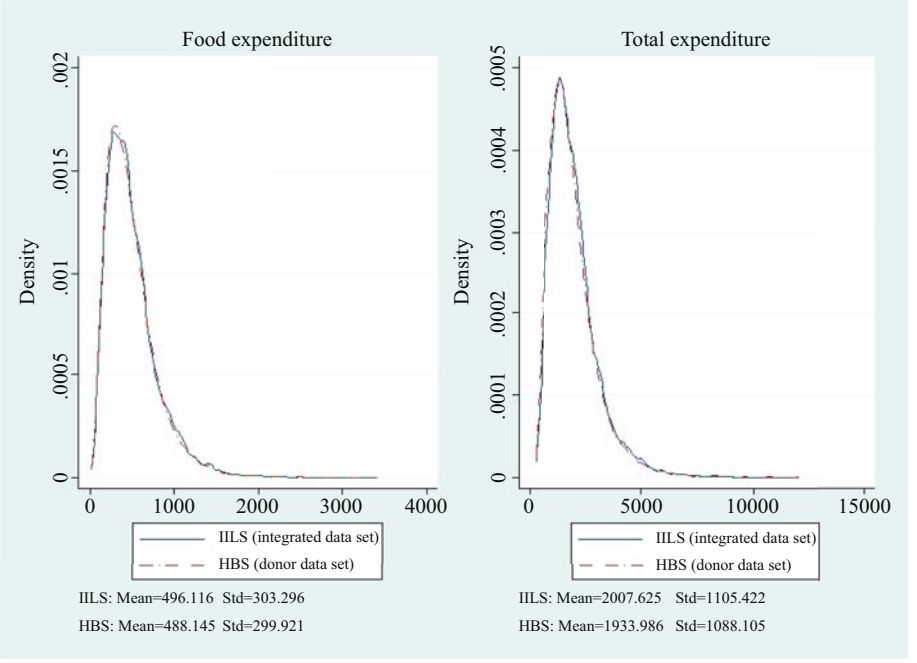

Fig. 5. Distribution of "Food expenditure" and "Total expenditure" in integrated and donor data sets.

Table 7. Cereals expenditure (in euros): Ratio of mean and median by covariates.

	Average			Median		
	HBS	IILS	Ratio	HBS	IILS	Ratio
Geographical area						
North-Western	71.23	71.05	99.74	60.84	60.79	99.92
North-Eastern	70.04	70.73	100.97	58.07	59.72	102.84
Center	69.30	72.67	104.87	59.95	62.35	104.00
Southern	67.80	68.38	100.84	58.97	58.83	99.76
Islands	66.40	69.24	104.27	56.96	59.36	104.21
Children 0–5 years old						
No	67.63	70.35	104.03	57.70	60.18	104.30
Yes	82.88	72.65	87.66	72.65	60.99	83.96
Children 6–14 years old						
No	65.10	70.47	108.24	55.32	60.11	108.66
Yes	89.44	71.40	79.84	78.93	60.79	77.02
Self-employed						
No	65.89	70.46	106.94	56.09	60.18	107.29
Yes	82.54	71.25	86.32	72.57	60.65	83.57
Single-parent						
No	69.41	70.63	101.75	59.36	60.14	101.31
Yes	66.44	70.51	106.13	57.97	62.59	107.97
Homeownership						
No	64.02	70.48	110.09	54.16	60.65	111.98
Yes	70.89	70.66	99.68	60.80	60.22	99.05
Average family education						
Primary	59.70	69.63	116.64	50.01	59.20	118.38
Middle	72.84	71.19	97.73	62.97	61.40	97.51
Middle-High	78.90	71.47	90.59	69.72	60.65	86.99
High	70.58	70.95	100.52	60.39	60.22	99.72
University	61.17	68.86	112.56	50.72	59.69	117.71
Household income						
1st quintile	51.05	58.84	115.26	42.89	48.60	113.30
2nd quintile	62.65	64.51	102.98	53.28	54.95	103.13
3rd quintile	68.90	69.78	101.28	60.49	61.85	102.25
4th quintile	77.12	76.29	98.93	68.29	68.18	99.83
5th quintile	87.48	81.60	93.28	76.71	69.64	90.78

or 6–14 years old and where a member of the family is self-employed. In this case, the divergence may be due to the number of children in each age group, rather than simply their presence.

We also investigated uncertainty generated by the lack of identifiability given the available data by calculating the Fréchet inequality for the contingency table associating income and expenditure classes. The Fréchet inequalities bound the probabilities of two joint events given the probabilities of the individual events conditioning on a set of common variables. In the present context where we use categorical variables, if we only know the conditional distributions $F(y|x)$ and $G(z|x)$ it is not possible to learn something about the association between y and z given x, but we can identify the bounds $max(0, F(y|x) + G(z|x) - 1) \leq H(y,z|x) \leq min(F(y|x), G(z|x))$ describing how uncertain the

Table 8. Total household expenditure (in euros): Ratio of mean and median by covariates.

	Average			Median		
	HBS	IILS	Ratio	HBS	IILS	Ratio
Geographical area						
North-Western	2154.79	2021.39	93.81	1891.81	1783.93	94.30
North-Eastern	2147.08	2119.94	98.74	1914.22	1882.80	98.36
Center	1967.24	2197.14	111.69	1765.92	1989.06	112.64
Southern	1705.08	1766.91	103.63	1539.43	1562.48	101.50
Islands	1552.99	1748.30	112.58	1404.63	1552.45	110.52
Children 0–5 years old						
No	1884.86	1994.32	105.81	1655.32	1772.07	107.05
Yes	2364.10	2110.34	89.27	2133.66	1879.10	88.07
Children 6–14 years old						
No	1836.22	2015.54	109.77	1608.69	1797.63	111.74
Yes	2417.41	1966.00	81.33	2164.89	1729.95	79.91
Self-employed						
No	1806.41	2004.22	110.95	1591.03	1782.42	112.03
Yes	2449.27	2021.69	82.54	2174.32	1800.23	82.79
Single-parent						
No	1939.83	1993.21	102.75	1718.85	1769.21	102.93
Yes	1858.89	2160.95	116.25	1639.10	1954.15	119.22
Homeownership						
No	1817.00	2032.74	111.87	1638.38	1829.96	111.69
Yes	1972.39	1999.03	101.35	1738.17	1769.07	101.78
Average family education						
Primary	1340.58	1952.43	145.64	1155.36	1753.83	151.80
Middle	1929.09	1945.56	100.85	1719.70	1731.26	100.67
Middle-High	2307.72	2072.13	89.79	2083.60	1816.95	87.20
High	2260.24	2077.00	91.89	2012.14	1846.66	91.78
University	2293.16	2031.11	88.57	2062.34	1762.85	85.48
Household income						
1st quintile	1181.89	1481.45	125.35	1010.90	1235.49	122.22
2nd quintile	1593.65	1663.30	104.37	1436.04	1494.91	104.10
3rd quintile	1914.77	1931.64	100.88	1716.19	1730.43	100.83
4th quintile	2286.43	2247.94	98.32	2073.73	2015.75	97.20
5th quintile	2747.46	2581.01	93.94	2464.87	2332.84	94.64

association is between $yz|x$. When the intervals are statistically close, then the common variables of interest are suitable for matching.

In the present estimation of the Fréchet bounds, we consider the set of common variables used in the propensity score estimation. We first estimated these bounds, setting the intervals equal to income quintiles as used in our model specification. In order to analyze the influence of the width of the classes on the measure of uncertainty, we computed the same analysis also setting the intervals equal to income tertiles, eight fixed classes, as defined in Donatiello et al. (2014) that also match HBS and EUSILC, and income deciles. Consumption information was aggregated using the same classes defined for the income distribution. As reported in Table 9, the width of uncertainty is remarkably reduced from 20.3% to 5.9%, moving from tertiles to deciles. Donatiello et al. (2014) report an average width of the uncertainty bound

Table 9. Average width of uncertainty bounds conditioning on common variables by different classes.

Classes	*Average width of uncertainty bounds*
Income tertile	*0.203*
Income quintile	*0.125*
Eight classes defined by Donatiello et al. (2014)*	*0.069*
Income decile	*0.059*

**Donatiello et al. (2014) defined the following classes: "Under EUR 1000", "EUR 1000–1500", "EUR 1500–2000", "EUR 2000–2600", "EUR 2600–3100", "EUR 3100–3600", "EUR 3600–5200" and "EUR 5200 or more".*

equal to 7.8%, setting eight classes equal for income and consumption, which is comparable with our estimated range of 6.9% using the same intervals, though the comparison should be taken with caution because the number of conditioning variables is larger. If we take as a reference class definition the partition in deciles, we may consider an average width of 5.9% as a sound indication of a valid inference, though there still seems to be a good margin for improvement if, for example, auxiliary information was available. Inspection of Table 10 shows that, conditioning on the common variables, all cell probabilities for the eight selected classes are between the lower and upper bounds.

In the next Subsubsection, we study the economic robustness of the matching by investigating the Engel relationship linking the food share, an approximate indicator of well-being (Perali 2003, 2008), and the logarithm of total expenditure. This is a fundamental empirical relation that is stable independently of the society analyzed and the time period considered.

4.1.1. Economic Robustness of the Matched Data: The Engel Relationship and Material Well-Being

An immediate check of the economic robustness of the matched data is the comparison of income in the recipient EUSILC database and consumption from the HBS donor data set. Table 11 shows the number of households per income-expenditure and row frequencies of quintiles of household income and total expenditure grouped by the same classes of income quintiles. The marginal column of Table 11 shows that in the lowest quintiles, total expenditure exceeds income for almost 72% of the families, suggesting under-reporting of income (Meyer and Sullivan 2011). On the other hand, as is reasonable to expect, most families in the upper income quintiles have positive savings.

In Table 12, we focus on the relationship between total expenditure and specific expenditure items in the fused and donor data set. As shown in Table 12, all budget shares have a similar magnitude and pattern in both data sets. Food, clothing and housing shares decrease as total expenditure increases, as is typical for necessity goods. On the other hand, the budget share of transport and communication and recreation and education increase as total consumption increases.

Table 10. Uncertainty bounds for total household income and consumption.

Income classes	Consumption classes	Low.cx	CIA	Up.cx
1	1	0.00021	0.03565	0.10796
2	1	0.00010	0.03799	0.11970
3	1	0.00010	0.02918	0.10007
4	1	0.00000	0.02174	0.08065
5	1	0.00010	0.01085	0.05241
6	1	0.00005	0.00724	0.04055
7	1	0.00000	0.01094	0.04958
8	1	0.00000	0.00460	0.02827
1	2	0.00015	0.03701	0.12627
2	2	0.00008	0.04148	0.14724
3	2	0.00004	0.03939	0.14689
4	2	0.00000	0.03474	0.13315
5	2	0.00003	0.02051	0.08488
6	2	0.00005	0.01559	0.06708
7	2	0.00006	0.02669	0.10232
8	2	0.00001	0.01299	0.05510
1	3	0.00027	0.02766	0.10379
2	3	0.00016	0.03195	0.11913
3	3	0.00008	0.03486	0.14274
4	3	0.00008	0.03415	0.14402
5	3	0.00007	0.02252	0.09490
6	3	0.00000	0.01784	0.07493
7	3	0.00006	0.03193	0.12102
8	3	0.00012	0.01651	0.06482
1	4	0.00012	0.01876	0.07947
2	4	0.00016	0.02276	0.09596
3	4	0.00015	0.02689	0.11786
4	4	0.00005	0.02843	0.12807
5	4	0.00002	0.02015	0.09459
6	4	0.00000	0.01636	0.07610
7	4	0.00018	0.03097	0.12593
8	4	0.00016	0.01671	0.06658
1	5	0.00003	0.00760	0.04745
2	5	0.00010	0.00991	0.05767
3	5	0.00010	0.01228	0.06665
4	5	0.00000	0.01353	0.07412
5	5	0.00013	0.01039	0.07086
6	5	0.00005	0.00864	0.06341
7	5	0.00003	0.01595	0.07668
8	5	0.00018	0.00911	0.05257
1	6	0.00007	0.00408	0.03124
2	6	0.00006	0.00538	0.03740
3	6	0.00018	0.00699	0.04104
4	6	0.00005	0.00748	0.04426
5	6	0.00003	0.00584	0.04441
6	6	0.00005	0.00520	0.04258
7	6	0.00000	0.00980	0.04697
8	6	0.00002	0.00597	0.03840

Table 10. Continued.

Income classes	Consumption classes	Low.cx	CIA	Up.cx
1	7	0.00001	0.00439	0.03289
2	7	0.00003	0.00599	0.03973
3	7	0.00010	0.00763	0.04482
4	7	0.00000	0.00858	0.04950
5	7	0.00005	0.00655	0.04814
6	7	0.00007	0.00614	0.04676
7	7	0.00017	0.01207	0.05559
8	7	0.00027	0.00789	0.04547
1	8	0.00000	0.00111	0.01156
2	8	0.00000	0.00154	0.01298
3	8	0.00000	0.00191	0.01388
4	8	0.00000	0.00223	0.01471
5	8	0.00002	0.00184	0.01446
6	8	0.00000	0.00175	0.01459
7	8	0.00006	0.00346	0.01559
8	8	0.00019	0.00251	0.01455

Notes: Classes are coded as: 1 = "Under EUR 1000"; 2 = "EUR 1000–1500"; 3 = "EUR 1500–2000"; 4 = "EUR 2000–2600"; 5 = "EUR 2600–3100"; 6 = "EUR 3100–3600"; 7 = "EUR 3600–5200"; 8 = "EUR 5200 or more".
Low.cx: The estimated lower bounds for the relative frequencies when conditioning on the common variables.
CIA: The estimated relative frequencies under the Conditional Independence Assumption (CIA).
Up.cx: The estimated upper bounds for the relative frequencies when conditioning on the common variables.

We further concentrate on the relationship between the food category and total expenditure because it is a robust relation whose main features should be maintained in the integrated database. Food expenditure and total expenditure have a similar distribution pattern in the original and fused data set both in the bottom and upper tail (Figure 5). This aggregate picture may hide significant differences, especially in the bottom and top five

Table 11. Conditional frequencies and percentages by income quintiles (in euros).

Quintiles of household income (Y)	Total expenditure (X)						Average savings (Y-X)
	<= 1199	1199 ⊣ 1788	1788 ⊣ 2545	2545 ⊣ 3662	> 3662	Total	
<= 1199	1060	1026	912	501	237	3736	−1098.86
	28.37	27.46	24.41	13.41	6.34	100.00	
1199 ⊣ 1788	937	1068	895	558	272	3730	−450.12
	25.12	28.63	23.99	14.96	7.29	100.00	
1788 ⊣ 2545	754	1095	1098	616	299	3862	94.42
	19.52	28.35	28.43	15.95	7.74	100.00	
2545 ⊣ 3662	807	1058	1055	608	273	3801	1046.40
	21.23	27.83	27.76	16.00	7.18	100.00	
> 3662	702	1092	1165	703	356	4018	3330.52
	17.47	27.18	28.99	17.50	8.86	100.00	
Total	4260	5339	5125	2986	1437	19147	
	22.25	27.88	26.77	15.60	7.51	100.00	

Table 12. Average budget share by quintile group of total expenditure.

Expenditure category		Quintiles of total expenditure				
		1	2	3	4	5
Food	IILS	0.293	0.278	0.260	0.252	0.225
	HBS	0.301	0.280	0.272	0.257	0.226
Clothing	IILS	0.108	0.098	0.092	0.080	0.059
	HBS	0.107	0.098	0.092	0.079	0.058
Housing	IILS	0.327	0.293	0.274	0.255	0.212
	HBS	0.326	0.290	0.266	0.253	0.211
Transport and communication	IILS	0.103	0.151	0.166	0.178	0.185
	HBS	0.098	0.151	0.168	0.177	0.184
Recreation and education	IILS	0.058	0.083	0.107	0.134	0.198
	HBS	0.058	0.084	0.105	0.133	0.199
Health	IILS	0.110	0.096	0.101	0.102	0.121
	HBS	0.111	0.098	0.097	0.101	0.123

percent of the distribution. This is apparent when we compare the quantiles of the synthetic data set against the quantiles of the donor data set, as shown in the Q-Q Plot of Figure 6 referring to the whole sample. If the two groups belong to a population with the same distribution, the point should fall along the 45-degree reference line. Figure 6 shows a different pattern between the two samples for both food and total expenditure, only in the upper tail of the distribution. However, if we zoom in to the bottom and top five percent of

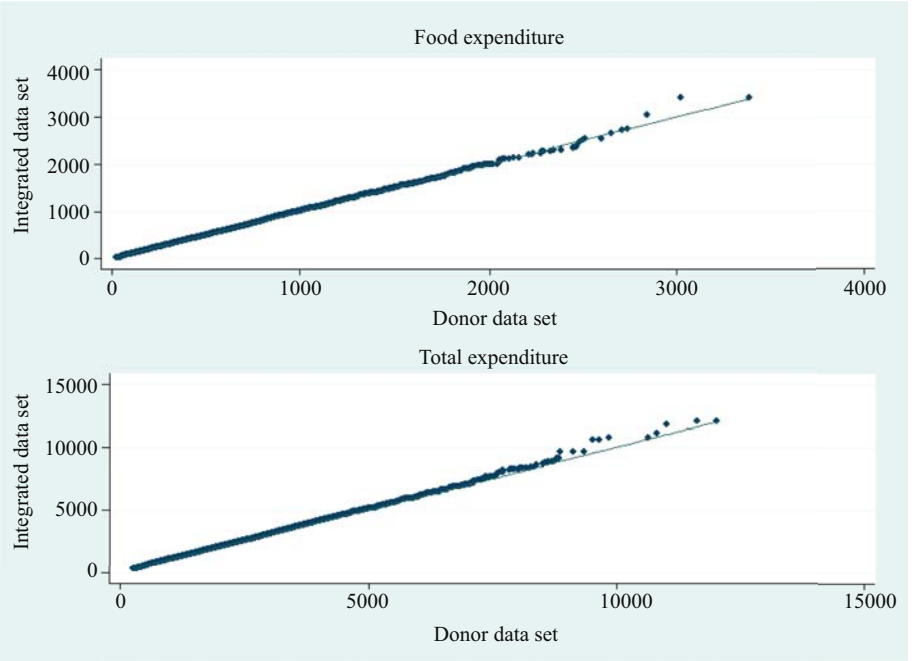

Fig. 6. Q-Q plot of food expenditure and total expenditure.

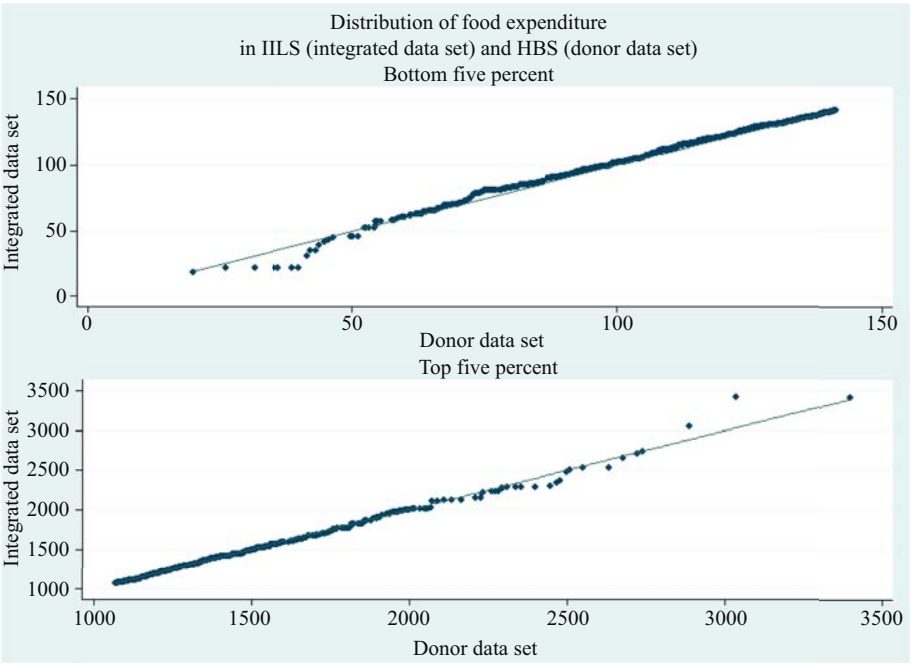

Fig. 7. Q-Q plot of food expenditure: focus on the tails.

the distribution, as shown in Figures 7 and 8, a similar departure in the lower tails can be seen, representing less than five percent of the sample.

In order to describe the shape of the food and total expenditure distributions at the tail, as shown in Tables 13 and 14, we test the statistical difference of the computed ratios of the 90th and 10th percentile describing the extent to which food or total consumption is larger at the top compared to the bottom of both the donor (HBS) and matched (IILS) population. As shown in Tables 13 and 14, we also summarize the dispersion of food and total expenditure with the Gini inequality index and test their difference. Table 14 also illustrates the Foster-Greer-Thorbecke (FGT) poverty measures and the associated statistics testing for the difference of the poverty measures in the donor and fused samples. The Foster-Greer-Thorbecke (Foster et al. 1984) indices are computed by substituting different values of the parameter α in the equation

$$\mathrm{FGT}_{\alpha} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{H}} \left(\frac{\mathrm{z} - \mathrm{y_i}}{\mathrm{z}} \right)^{\alpha},$$

where Z is the poverty threshold equal to 60% of the median of total expenditure respectively in the IILS integrated data (EUR 1082.944) and in the original HBS sample (EUR 1040.557), N is the sample size, H is the number of poor (those with total expenditure at or below z) and y_i is total expenditure of each individual i. With $\alpha = 0$, FGT_0 is the headcount ratio, the proportion of the population below the poverty line. With $\alpha = 1$ FGT_1 represents the poverty gap index, which summarizes the extent to which individuals fall below the poverty line. With $\alpha = 2$ FGT_2 measures the squared poverty gap ("poverty

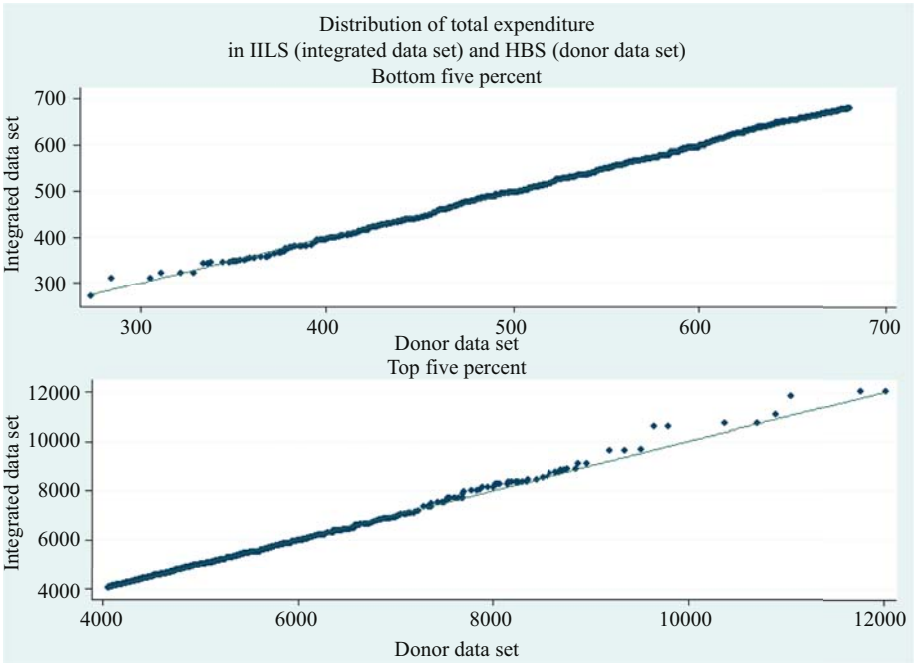

Fig. 8. Q-Q plot of total expenditure: focus on the tails.

severity") index, which places stronger emphasis on the poverty of the poorest individuals. With the exception of the percentile ratio for total expenditure, for all other comparisons we do not reject the null hypothesis that the estimates in the donor and matched data sets are the same at the .01 significance level. On the basis of this evidence, we conclude that the outcome of the matching is both statistically and economically robust.

To further verify the economic robustness of the matched distribution in a welfare measurement context, we estimated the Engel relationship linking the food share, a reliable proxy for well-being (Perali 2003, 2008), and the logarithm of total expenditure, as shown in Figure 9, which plots the inverse relationship between the food share and the logarithm of total expenditure. As the level of total expenditure increases, the food share, and the associated level of household well-being, decreases in a similar fashion in both the recipient and donor distribution.

To also investigate the shape of the conditional distribution of food expenditure with the logarithm of total expenditure in the lower and upper tails where there is higher statistical

Table 13. Dispersion indexes for food expenditure.

		p90/p10	Gini coefficient
IILS (integrated data set)		4.7902	0.3189
HBS (donor data set)		4.7309	0.3206
DIFFERENCE		−0.0592	0.0017
	std. err.	*0.0339*	*0.0022*
	p-value	*0.0802*	*0.4474*

Table 14. Inequality and poverty indexes for total expenditure.

				FGT$_\alpha$ poverty index*		
		p90/p10	Gini coefficient	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$
IILS (integrated data set)		3.6310	0.2766	0.1568	0.0334	0.0101
HBS (donor data set)		3.7593	0.2816	0.1658	0.0354	0.0106
DIFFERENCE		0.1283	0.0050	0.0090	0.0020	0.0005
	std. err.	*0.0218*	*0.0021*	*0.0035*	*0.0012*	*0.0006*
	p-value	*0.0000*	*0.0175*	*0.0103*	*0.1004*	*0.3832*

*$\alpha = 0$: headcount ratio, $\alpha = 1$: poverty gap index, $\alpha = 2$: squared poverty gap index.

noise, we estimated the Engel relation by using also a quantile regression for each distribution quantile not influenced by extreme values. We estimated five quantile regressions for the quantiles 0.10, 0.25, 0.50, 0.75, and 0.90. Figure 10 shows the estimated quantile coefficients with the associated confidence intervals (solid line) and the least squares coefficients (dashed line) that, by construction, do not vary by quantile. OLS estimates underestimate, especially in the lower tails, both the matched IILS data set and the donor HBS dataset. The underestimation is larger in the integrated data set. Figure 11 shows the estimated quantile and OLS coefficients in the same graph. The distance between the estimated OLS coefficients in the integrated and donor data set and by quantile is not economically significant, although it is slightly larger in the lowest quintiles. The difference between quantile regression coefficients at the level of the second quintile is .005. This means that even if the estimated parameter is statistically significant,

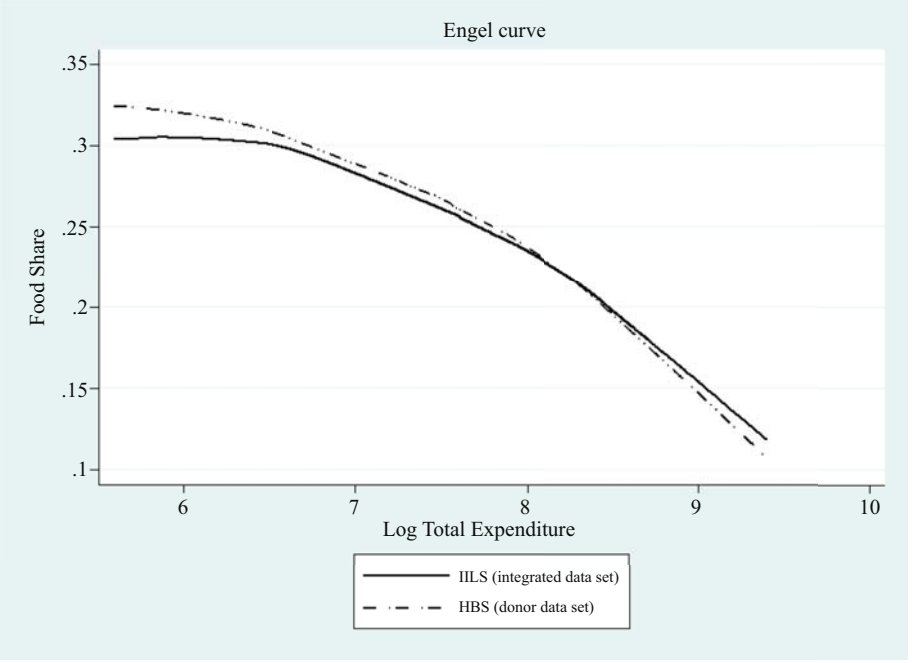

Fig. 9. Engel curve in integrated and donor data sets.

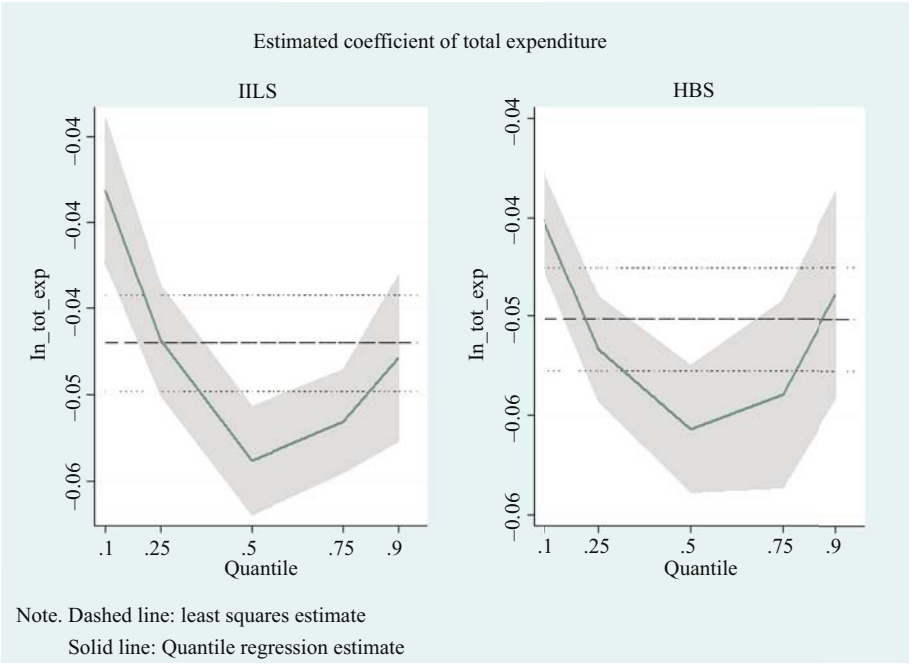

Fig. 10. Estimated coefficient of total expenditure with OLS regression and quantile regression.

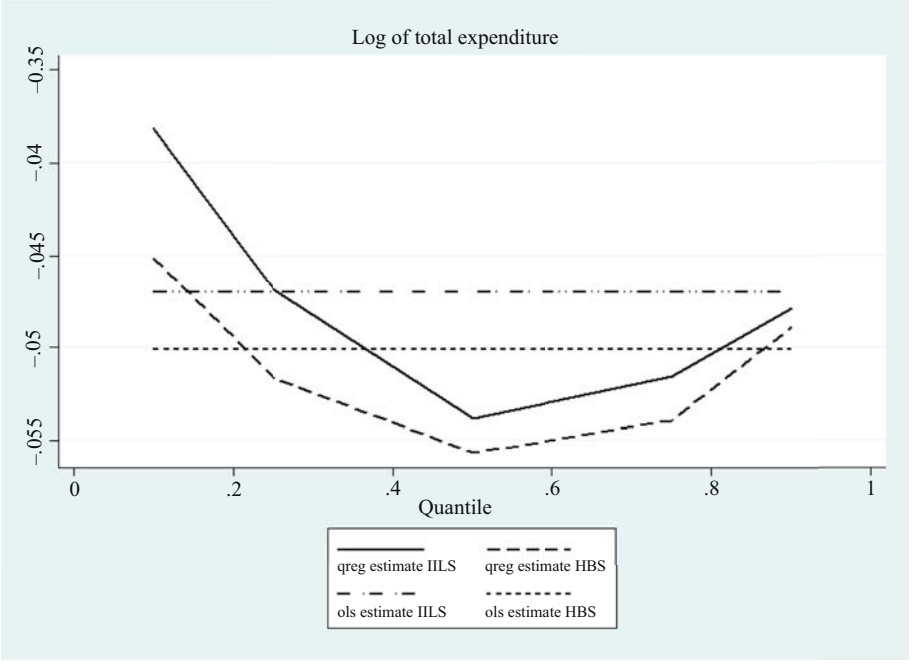

Fig. 11. Estimated coefficient of total expenditure with OLS regression and quantile regression.

the variable's impact is economically negligible (Goldberger 1991). This evidence shows that in the fused data set the economic information is robustly maintained along all the relevant portions of the income distribution.

4.2. Data Fusion Between EUSILC and TUS

The EUSILC survey does not collect information about how Italians spend their time. These detailed data are traditionally gathered within specifically designed time use surveys. Both samples constructed by ISTAT are drawn from the same population with the same sampling design. They share a large set of common variables. Both fundamental conditions are satisfied, so that we could reliably perform the statistical matching technique.

We used the same covariates to match the activities on a weekday, Saturday, and Sunday. We obtained the same conclusion for the time spent on a weekday, on Saturday, and on Sunday. Consequently, as supplementary data we only report the results for the time spent on main activities during a weekday.

The common covariates used in the specification of the propensity score model included the region of residence (three dummies coded as North, Center, South), age (nine dummies for the age classes 3–5, 6–14, 15–19, 20–26, 27–36, 37–46, 47–56, 57–66, and older than 66), gender (1 if the individual is male), the presence in the family of a worker (1 if there was at least one working member), the presence of students (1 if there was at least one student in the family), the presence of children by age classes (0–5, 6–13 and 14–18 years of age), single-parent family (1 if there was a mother or father without partner, 0 otherwise) and the educational level attained (1 if the highest education level was high school or more, 0 otherwise). This set of common variables is the same as the common set used for the EUSILC and HBS, except the income variable that is not present in the TUS.

The distributions of these variables do not show any significant relationship between the two samples (Table A4). The largest absolute differences are recorded for geographical area but, as highlighted by Cramer's V test, these differences are not statistically significant. We also tested the equal distribution of the covariates before and after matching (Table A5). The largest standardized differences before matching are observed between the categories that refer to the geographical area. These differences disappear after matching. The p-values highlight the equality of means of covariates after matching. The covariates that reject the null hypothesis of equality of means before matching are the same covariates recording higher standardized difference.

The estimated propensity score shows a similar density distribution and its values show a common support in the recipient and donor databases. Observations have the same probability of belonging to one of the two samples (Figure A1) and we can be confident about obtaining unbiased results after implementing the matching algorithm to impute the missing values in the recipient database. To lend further support to this assertion, we investigated the matching quality for the variables in which we are most interested, such as rest, work, study and mobility (Tables A6–A9). These figures describe the differences in the original and matched database and in the ratio of the means by each covariate used in the propensity score specification. Almost all ratios are close to 100. This implies that the average in the two groups is similar. Marked deviations from 100 are explained by the

presence of some outliers in the donor data set that are not used to "impute" the missing values in the integrated database as shown by the heavy upper tails (Figures A2 and A3). This problem can be solved by computing the ratio of medians that gives statistical values not influenced by outliers. Note that it is not possible to use the ratio of medians because in most cases the median is equal to 0 and therefore the ratio cannot be calculated. In fact, the time spent on a particular activity does not depend only on one socio-demographic variable as represented in the tables, that is, work time should be compared jointly in relation to age and occupational status.

4.3. Data Fusion Between EUSILC and CISF Surveys

This matching involves the EUSILC survey, which does not present information about social capital, and the CISF survey, which collects detailed information on both bridging and bonding social capital and relational well-being (Menon et al. 2015). The set of common variables is the same as the common set used for the EUSILC and HBS, and EUSILC and TUS with the addition of the occupational status of women. Here, the income variable is not part of the set because it did not pass the balancing procedure.

To link these data sets, we implemented two different propensity score specifications because some variables about family relationships are pertinent only for some types of family. One propensity score specification concerned questions about family relations and the relationship with children. As a consequence, this specification related to a subsample of the EUSILC and CISF data set that does not include singles. We also excluded the families defined as "other types of family" because this typology is not defined in the same way in the two questionnaires and comparison is impossible with the available information. The other specification, on the other hand, analyzed the whole sample because the questions of interest are not related to family composition.

Statistical matching between these two questionnaires can be applied because the surveys refer to the same target population and share a set of common covariates with the same definition. Some variables are used in both specifications. We describe both because the sample size differs and this may affect the shape of the distribution.

4.3.1. Propensity Score Specification Excluding Singles and Other Family Types

In this propensity score specification, which excludes singles and other family types, we included the following variables: region of residence (five dummies coded as North-West, North-East, Center, South, Islands), age of the household head (three dummies coded as less than 35, 35–64, older than 64), dummies for the presence of children by age class (0–5, 6–13 and 14–18 years of age), main activity of the head of the household (four dummies coded as Employee, Unemployed, Retired, Inactive person), woman's occupational status (dummy equal to 1 if the household's wife/partner works), single-parent family (1 if there is a mother or father without partner, 0 otherwise) and education level attained by the household head (four dummies coded as Primary, Middle, High, University).

The distribution of these variables after their harmonization and aggregation is reported in the Table A10. Only the different levels of education have relatively higher values of absolute differences, although they are not statistically different in the two groups as measured by Cramer's V test.

The specification used in the propensity score model achieved the balance in observed covariates (Table A11). Almost all values of standardized differences are reduced after matching, and the p-values show that the means of the recipient and the donor database are not statistically different. The propensity score distribution is similar in the same common support region, so we conclude that the observations have the same probability of being assigned to one of the two samples (Figure A4).

The quality of the matching outcome is high. The ratios of mean are close or very close to 100, revealing that the two databases have similar distributions of the extra information (Tables A10–A14, Figures A5 and A6).

4.3.2. Propensity Score Specification for the Whole Sample

This specification involved the whole sample because the extra information was not related to family type but concerns the attitude to participation in social life and social framework that pertains to singles and families as well. The specification also included variables regarding family composition, because the time spent on social events and voluntary activities also depends on family characteristics. We considered the region of residence (three dummies coded as North, Center, South), three dummies for the presence of children by age class (0–5, 6–13 and 14–18 years of age), two dummies describing man and woman's occupational status (1 if the man/woman was an employee), single-parent family (1 if there was only the mother or the father without partner) and level of education of the head of the household (four categories coded as Primary, Middle, High, University).

The frequency distribution of these variables shows a similar trend in the two samples (Table A15). The level of education of the head of the household displays the largest absolute differences between categories, but these differences are not statistically different, as pointed out by the result of Cramer's V test.

This specification proves that the observed variables are balanced between the recipient and the donor database. After matching, the standardized differences of all covariates are close to 0 and the p-values of the t-tests do not reject the null hypothesis of equality of means in the two samples (Table A16). The distribution of the propensity score value shows that the observations with the same characteristics have the same probability of extraction from both the synthetic and original data set (Figure A7). For simplicity's sake, we show the matching outcome for the variable "Take part in social activities or voluntary work", which is one of the variables of keenest interest in the present matching design because of its relevance to the measurement of well-being. The distribution is similar in the donor and integrated data set. Its ratios of mean are close to 100 (Figure A8 and Table A17).

5. An Example of an Empirical Application to the Measurement of Multidimensional Poverty

To better communicate at least some of the insights that can be obtained using the fused living standard data, we propose some salient results, also from a policy point of view, from an empirical exercise related to the multidimensional measurement of poverty. The monetary dimension of poverty is not sufficient to capture the multifaceted reality of poverty. A person with a relatively low standard of living may suffer from multiple

deprivations. A person in poverty may be jobless and houseless, a single parent, lacking good health, sufficient education or time to invest in the family. It could also be a person poor in the relation or social capital dimensions. Some of these dimensions are not strongly associated with income and can be highly informative about non-material dimensions of well-being. In our analysis, the monetary dimension can take the traditional form of disposable (after-tax) household income, may include the current income derived from the property's net worth (Brandolini et al. 2010), or may additionally include the evaluation of time invested in household production to form an extended notion of income.

In general, an individual receives income Y from labour, pensions, and other transfers and may hold a certain level of net worth or wealth W. Net worth, obtained as total income minus total liabilities, is thus an indicator of long-run economic security, while access to liquid assets is an indicator of the ability to cope with unanticipated emergencies. Current income CY is then defined as the sum of income Y and property income rW, where r is the average rate of return on assets, $CY = Y + rW$. Current income is an important determinant of the "economic situation" of an individual that depends on the flow of services over which it has command (Brandolini et al. 2010).

Adding to current income the value of time invested in household production gives a measure of extended income. The problem of estimating the value of the production of household services stems from the fact that the household product is not marketable. It is therefore difficult to know the value of the marginal product generated within the family enterprise. Therefore, the value of time devoted to paid market or unpaid domestic activities differ. Household production is a nonmarket activity whose value can be measured by its opportunity or market cost. A reasonable practice is to evaluate the time devoted to children at the market value, that is the wage at which families would pay the person that would substitute parents' care (Sharpe et al. 2011, Caiumi and Perali 2015; Poissonnier and Roy 2017).

Such a comprehensive picture of a deprivation profile can be described only using Integrated Living Standards data sets. In the present case, consumption information comes from the household budget survey, income and wealth from the standard of living survey, household time allocation from the time use survey and information on relational well-being from the social capital survey. A multidimensional measure of poverty counts the different forms of deprivation that a person experiences at the same time in different indicators of poverty that, in the present application, are equally weighted. By convention,

Table 15. Incidence of poverty (headcount ratio – H).

	North	Centre	South	Italy
Italian sample				
Equivalent total expenditure	0.1076	0.0893	0.2151	0.1356
Equivalente disposable income	0.0654	0.0852	0.1991	0.1100
Equivalent current income	0.0679	0.0871	0.2007	0.1121
Equivalent extended income	0.0388	0.0505	0.0870	0.0559
Subsample of Italian families with children				
Equivalent total expenditure	0.1983	0.1450	0.2957	0.2200
Equivalente disposable income	0.0709	0.0960	0.2194	0.1276
Equivalent current income	0.0773	0.1067	0.2416	0.1406
Equivalent extended income	0.0462	0.0710	0.0836	0.0647

Table 16. Comparison of poverty measures by number of deprivations and first dimension (total expenditure, disposable/current/extended income).

	North		Centre		South		Italy	
First dimension	H	MPI	H	MPI	H	MPI	H	MPI
6 dimensions (cutoffs 3)								
Equivalent total expenditure	0.087	0.046	0.115	0.062	0.149	0.083	0.115	0.062
Equivalent disposable income	0.084	0.046	0.133	0.074	0.154	0.089	0.119	0.067
Equivalent current income	0.086	0.048	0.138	0.077	0.161	0.093	0.124	0.070
Equivalent extended income	0.082	0.045	0.127	0.071	0.121	0.071	0.105	0.060
10 dimensions (cutoffs 6)								
Equivalent total expenditure	0.041	0.022	0.075	0.039	0.116	0.063	0.074	0.040
Equivalent disposable income	0.044	0.024	0.084	0.045	0.118	0.065	0.079	0.043
Equivalent current income	0.047	0.025	0.086	0.046	0.123	0.068	0.082	0.045
Equivalent extended income	0.041	0.023	0.087	0.046	0.104	0.058	0.073	0.040

Notes:

1. Cutoffs: the minimum number of dimensions in which a person is deprived to be considered multidimensionally poor.
2. The dimensional cutoffs are: half the median for total expenditure or income (disposable/current/extended), net worth and women's time use; secondary school as level of parents education, single parenthood, at least one unemployed in the family and a satisfaction level of eight (in a range 0–10) for social capital dimensions.
3. Incomes are equivalized using Italian Equivalent Scales introduced January 1st, 2015. This scale accounts for family composition, parents' working condition and number of parents present in the family.

a household is identified as multidimensionally poor if it is deprived in some combination of indicators whose weighted sum exceeds 30% of all deprivations (Alkire and Foster 2011). The traditional unidimensional approach to measure poverty is to calculate the proportion of the population who are poor, or headcount ratio H, on the basis of disposable income or total household expenditure. We also compute the index H considering the current and extended notion of income. We further calculate the multidimensional poverty index ($MPI = HA$), or adjusted headcount ratio, as the product of the incidence of poverty (H) and the average intensity of deprivation (A) reflecting the proportion of dimensions in which households are, on average, deprived.

Table 15 reports the incidence of poverty H for both the Italian sample and the subsample of Italian families with children also distinguishing the North, Center and South macroregions based on equivalent disposable, current, extended incomes and total household expenditure. Table 16 presents both the H and MPI measures for six and ten deprivation dimensions. These deprivation dimensions are: 1) equivalent household total expenditure or income (disposable/current/extended), 2) net worth, 3) parents education, 4) number of parents, 5) presence in the family of unemployed members, 6) women's time use for child care and household chores, 7) trust in family members, 8) trust on friends or acquaintances, 9) satisfaction of the relationship with children, 10) satisfaction about time spent together. The results are limited to the subsample of Italian families with children, because only in this context these relational variables are observable. Interestingly, the relative contribution of the dimensions "trust on friends" and "satisfaction about time spent together" are the two most important contributions of all deprivation dimensions. The striking result is that the poverty gap between the North and the South reduces increasingly as we integrate deprivation dimensions in terms of both H and MPI. This is a completely new map of poverty of great utility to policy-makers that we have been able to draw thanks to the construction of the Integrated Italian Living Standard data set.

6. Conclusions

This study has described a procedure used to construct a data set integrating Italian consumption, income, time use, and social capital surveys, adopting propensity score matching. The choice of fusing four data sets was motivated by the recommendations of the Fitoussi Commission (Stiglitz et al. 2010) and the interest of the Italian National Institute of Statistics in estimating well-being from an equitable and sustainable point of view. In general, integrated information is crucial for improving the quality of the estimation of household and individuals' well-being and of the comparisons of their standards of living.

Statistical matching can be seen as an imputation procedure for missing values from a donor data to a recipient data set. We used the propensity score value as a synthetic indicator of the common variables used in the specification model. This study gives detailed information on the matching variables and the statistical tests of the independence of the covariates playing special attention to the main data fusion between the EUSILC and HBS surveys, which we evaluated also exploring uncertainty.

We also compared the distributions of the extra information in the original and synthetic database. For the imputed information, we computed the ratio of mean and median between the two databases for the covariates used in the propensity score specification. We

also tested the economic robustness of the related data set by the Engel relationship, often used as a benchmark measure for welfare measurement. The matched data set passed all statistical and economic tests. To illustrate the value of the integrated information about standards of living we describe an example related to the multidimensional measurement of poverty. The noticeable result is that the poverty gap between the North and the South of Italy reduces increasingly as we integrate deprivation dimensions. This approach revealed a novel map of poverty of significant policy interest that we have been able to draw thanks to the construction of the Integrated Italian Living Standard data set.

The objective of this study is undeniably challenging because it deals with independent data sources not designed with integration purposes. Indeed, from a methodological point of view, we share the common hope that the international institutional effort to produce greater harmonization across HBS, EUSILC and TUS of both socio-demographic and other key economic variables will soon generate significant changes in their questionnaire design. As an example, a useful anchoring between HBS and EUSILC for matching purposes may occur if both surveys are record linked with administrative registers on income and wealth. Further, the *ex-ante* collection of auxiliary variables for integration purposes may involve both food consumed at home or away from home and clothing and footwear (not only in EUSILC, but also in HBS as aggregate recall questions), cumulated and short-term savings, housing value and expenses, transport, health conditions and, not last, stylized time use questions. This evolution would provide important auxiliary information and more meaningful logical constraints that can be effective in making the bias due to the conditional independence assumption negligible by reducing uncertainty.

An underexplored empirical issue that seems worth investigating in a systematic fashion is the comparison of the matching quality between propensity score matching and nonparametric matching methods placing especial emphasis on the selection procedure of the best set of matching variables and on the opportunities to deal with complex sample designs through weights' calibration procedures during the execution of the process. Another relevant empirical issue that may be more thoroughly analyzed is the measurement of the impact on the estimated standard errors derived from the fusion of multiple complex sample surveys.

Despite the lack of valuable auxiliary information, the results are satisfactory. Therefore, we can conclude that the integrated database to measure living standards in Italy can be reliably used to implement multidimensional inequality and poverty analysis explicitly assessing the value of time and social capital and, in general, to measure individual, household and social welfare more thoroughly.

7. References

Albayrak, Ö. and T. Masterson. 2017. *Quality of Statistical Match of Household Budget Survey and SILC for Turkey*. Levy Economics Institute (Working Paper no. 885). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2924849 (accessed October 2018).

Alkire, S. and J. Foster. 2011. "Counting and Multidimensional Poverty Measurement." *Journal of Public Economics* 95(7–8): 476–487. Doi: http://dx.doi.org/10.1016/j.jpubeco.2010.11.006.

Andridge, R.R. and R.J.A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78(1): 40–64. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00103.x.

Attanasio, O., E. Hurst, and L. Pistaferri. 2015. "The Evolution of Income, Consumption, and Leisure Inequality in the US, 1980–2010." In *Improving the Measurement of Consumer Expenditures*, edited by C.D. Carroll, T.F. Crossley, and J. Sabelhaus, 100–140. University of Chicago Press.

Augurzy, B. and C.M. Schmidt. 2001. *The Propensity Score: A Means to an End*, IZA Discussion Paper No. 271. Available at: https://ssrn.com/abstract=270919 (accessed September 2015).

Austin, P.C., N. Jembere, and M. Chiu. 2018. "Propensity Score Matching and Complex Surveys." *Statistical Methods in Medical Research* 27(4): 1240–1257. Doi: http://dx.doi.org/10.1177/0962280216658920.

Black, D.A. and J.A. Smith. 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching." *Journal of Econometrics* 121: 99–124. Doi: http://dx.doi.org/10.1016/j.jeconom.2003.10.006.

Blundell, R. and I. Preston. 1995. "Income, Expenditure and the Living Standards of UK Households." *Fiscal Studies* 16(3): 40–54. Doi: https://doi.org/10.1111/j.1475-5890.1995.tb00226.x.

Brandolini, A., S. Magri, and T. Smeeding. 2010. "Asset-based Measurement of Poverty." *Journal of Policy Analysis and Management* 29(2): 267–284. Doi: https://doi.org/10.1002/pam.20491.

Brewer, M. and C. O'Dea. 2012. *Measuring Living Standards with Income and Consumption: Evidence from the UK*. Institute for Social and Economic Research, University of Essex and Institute for Fiscal Studies (Working Paper n. 2012-05). Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2012-05.pdf (accessed April 2015).

Caiumi, A. and F. Perali. 2015. "Who Bears the Full Cost of Children? Evidence From a Collective Demand System." *Empirical Economics* 49: 33–64. Doi: http://dx.doi.org/10.1007/s00181-014-0854-2.

Caliendo, M. and S. Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22(1): 31–72. Doi: https://doi.org/10.1111/j.1467-6419.2007.00527.x.

Chen, J. and J. Shao. 2000. "Nearest Neighbor Imputation for Survey Data." *Journal of Official Statistics* 16(2): 113–131. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/nearest-neighbor-imputation-for-survey-data.pdf (accessed April 2015).

Coli, A., F. Tartamella, G. Sacco, I. Faiella, M. Scanu, M. D'Orazio, M. Di Zio, I. Siciliani, S. Colombini, and A. Masi. 2005. *La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane*. Technical Report, Working Group ISTAT- Bank of Italy, Rome. Available at: https://www.istat.it/it/files/2018/07/2006_12-1.pdf (accessed April 2015).

Conti, P.L., D. Marella, and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28: 69–88.

Conti, P.L., D. Marella, and M. Scanu. 2016. "Statistical Matching Analysis for Complex Survey Data with Applications." *Journal of the American Statistical Association* 111: 1715–1725. Doi: http://dx.doi.org/10.1080/01621459.2015.1112803.

Conti, P.L., D. Marella, and A. Neri. 2017. "Statistical Matching and Uncertainty Analysis in Combining Household Income and Expenditure Data." *Statistical Methods & Applications* 26(3): 485–505. Doi: http://dx.doi.org/10.1007/s10260-016-0374-7.

Dehejia, R.H. and S. Wahba. 1999. "Casual Effects in Nonexperimental Studies: Revaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94: 1053–1062. Doi: http://dx.doi.org/10.1080/01621459.1999.10473858.

Dehejia, R.H. and S. Wahba. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics* 84(1): 151–161. Doi: http://dx.doi.org/10.1162/003465302317331982.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006a. *Statistical Matching: Theory and Practice*. New York: Wiley.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006b. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints." *Journal of Official Statistics* 28: 137–157. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-matching-for-categorical-data-displaying-uncertainty-and-using-logical-constraints.pdf (accessed September 2017).

D'Orazio, M., M. Di Zio, and M. Scan. 2009. *Uncertainty Intervals for Non-Identifiable Parameters in Statistical Matching*. 57th Session of the International Statistical Institute, Durban (South Africa), August 2009. Available at: http://isi.cbs.nl/iamamember/CD8-Durban2009/A5%20Docs/0074.pdf (accessed April 2015).

D'Orazio, M., M. Di Zio, and M. Scanu. 2017. "The Use of Uncertainty to Choose Matching Variables." In *Statistical Matching*, edited by Ferraro et al. Soft Methods for Data Science. New York: Springer.

Donatiello, G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2014. "Statistical Matching of Income and Consumption Expenditures." *International Journal of Economic Science* Vol. III(3): 50–65. Available at: https://www.iises.net/download/Soubory/soubory-puvodni/pp50-65-ijoes-V3N3.pdf (accessed April 2015).

Foster, J., J. Greer, and E. Thorbecke. 1984. "A Class of Decomposable Poverty Measures." *Econometrica* 52(3): 761–766. Doi: http://dx.doi.org/10.2307/1913475.

Garrido, M.M., A.S. Kelly, J. Paris, K. Roza, D.E. Meier, R.S. Morrison, and M.D. Aldridge. 2014. "Methods for Constructing and Assessing Propensity Score." *Health Services Research* 49(5): 1701–1720. Doi: http://dx.doi.org/10.1111/1475-6773.12182.

Goldberger, A. 1991. *A Course in Econometrics*. Cambridge, Massachusetts: Harvard University Press.

ISTAT. 2013. *Rapporto Bes 2013: il Benessere Equo e Sostenibile in Italia*. Rome: ISTAT. Available at: https://www.istat.it/it/archivio/84348 (accessed August 2018).

ISTAT. 2014. *Rapporto Bes 2014: il Benessere Equo e Sostenibile in Italia*. Rome: ISTAT. Available at: https://www.istat.it/it/archivio/126613 (accessed August 2018).

Kiesl, H. and S. Rässler. 2009. "How Useful Are Uncertainty Bounds? Some Recent Theory with an Application to Rubin's Causal Model." 57th Session of the International Statistical Institute, Durban (South Africa), 16–22 August 2009. Available at: http://isi.cbs.nl/iamamember/CD8-Durban2009/A5%20Docs/0169.pdf (accessed April 2015).

Krug, E.G., L.L. Dahlberg, J.A. Mercy, A.B. Zwi, and R. Lozano. 2002. *World Report on Violence and Health*. Geneva: World Health Organization. Available at: https://www.who.int/violence_injury_prevention/violence/world_report/en/ (accessed April 2015).

Kum, H. and T.N. Masterson. 2010. "Statistical Matching Using Propensity Score: Theory and Application to the Analysis of the Distribution of Income and Wealth." *Journal of Economic and Social Measurement* 35: 177–196.

Lechner, M. 2008. "A Note on the Common Support Problem in Applied Evaluation Studies." *Annales d'Economie et de Statistique* 91/92: 217–235. Doi: http://dx.doi.org/10.2307/27917246.

Lee, W.S. 2013. "Propensity Score Matching and Variations on the Balancing Test." *Empirical Economics* 44(1): 47–80. Doi: http://dx.doi.org/10.1007/s00181-011-0481-0.

Leulescu, A. and M. Agafitei. 2013. *Statistical Matching: a Model Based Approach for Data Integration*. Luxembourg: Eurostat, European Commission. Available at: https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF/477dd541-92ee-4259-95d4-1c42fcf2ef34?version=1.0 (accessed April 2015).

Masterson, T. 2010. *Quality of Match for Statistical Matches Used in the 1999 and 2005 LIMEW Estimates for Canada*. Levy Economics Institute (Working Paper n. 615). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1670765 (accessed April 2015).

Masterson, T. 2014. *Quality of Statistical Match and Employment Simulations Used in the Estimation of the Levy Institute Measure of Time and Income Poverty (LIMTIP) for South Korea, 2009*. Levy Economics Institute (Working Paper n. 793). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2416850 (accessed April 2015).

Menon, M., R. Pendakur, and F. Perali. 2015. "All in the Family: How Do Social Capital and Material Wellbeing Affect Relational Wellbeing?" *Social Indicators Research* 124(3): 889–910. Doi: https://doi.org/10.1007/s11205-014-0816-2.

Meyer, B. and J. Sullivan. 2011. "Further Results on Measuring the Well-being of the Poor Using Income and Consumption." *Canadian Journal of Economics* 44(1): 52–87. Doi: http://dx.doi.org/10.1111/j.1540-5982.2010.01623.x.

Mittag, N. 2013. *Imputations: Benefits, Risks and a Method for Missing Data*. Harris School of Public Policy, University of Chicago. Available at: http://home.cerge-ei.cz/mittag/papers/Imputations.pdf (accessed April 2015).

Perali, F. 2003. *The Behavioral and Welfare Analysis of Consumption*. Dordrecht: Kluwer Academic Publishers.

Perali, F. 2008. "The Second Engel Law: Is it a Paradox?" *European Economic Review* 52(8): 1353–1377. Doi: http://dx.doi.org/10.1016/j.euroecorev.2008.01.005.

Poissonnier, A. and D. Roy. 2017. "Household Satellite Accounts for France. Methodological Issues on the Assessment of Domestic Production." *Review of Income and Wealth* 63(2): 353–368. Doi: http://dx.doi.org/10.1111/roiw.12216.

Rässler, S. 2002. *Statistical matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Lecture Notes in Statistics, 168, Springer.

Rässler, S. 2004. "Data Fusion: Identification Problems, Validity, and Multiple Imputation." *Austrian Journal of Statistics* 33: 153–171. Available at: https://www.researchgate.net/publication/228528513_Data_fusion_Identification_problems_validity_and_multiple_imputation (accessed April 2015).

Renssen, R.H. 1998. "Use of Statistical Matching Techniques in Calibration Estimation." *Survey Methodology* 24: 171–183. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X19980024354 (accessed August 2018).

Ridgeway, G., S.A. Kovalchik, B.A. Griffin, and M.U. Kabeto. 2015. "Propensity Score Analysis with Survey Weighted Data." *Journal of Causal Inference* 3(2): 237–249. Doi: http://dx.doi.org/10.1515/jci-2014-0039.

Rios-Avila, F. 2014. *Quality of Match for Statistical Matches Using the American Time Use Survey 2010, the Survey of Consumer Finances 2010, and the Annual Social and Economic Supplement 2011*. Levy Economics Institute (Working Paper no. 798). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2432153 (accessed April 2015).

Rios-Avila, F. 2015. *Quality of Match for Statistical Matches Using the Consumer Expenditure Survey 2011 and Annual Social Economic Supplement 2011*. Levy Economics Institute (Working Paper n. 830). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2554089 (accessed October 2015).

Rios-Avila, F. 2016. *Quality of Match for Statistical Matches Used in the Development of the Levy Institute Measure of Time and Consumption Poverty (LIMTCP) for Ghana and Tanzania*. Levy Economics Institute (Working Paper n. 873). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2838475 (accessed September 2017).

Rodgers, W.L. 1984. "An Evaluation of Statistical Matching." *Journal of Business & Economic Statistics* 2: 91–102. Doi: http://dx.doi.org/10.2307/1391358.

Rosenbaum, P.R. and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika* 70: 41–55. Doi: http://dx.doi.org/10.1093/biomet/70.1.41.

Rosenbaum, P.R. and D.B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39: 33–38. Doi: http://dx.doi.org/10.2307/2683903.

Rubin, D.B. 1986. "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business, Economics and Statistics* 4: 87–94. Doi: http://dx.doi.org/10.2307/1391390.

Rubin, D.B. and N. Schenker. 1986. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81: 366–374. Doi: http://dx.doi.org/10.1080/01621459.1986.10478280.

Sharpe, A., A. Murray, B. Evans, and E. Hazell. 2011. *The Levy Institute Measure of Economic Well-Being: Estimates for Canada, 1999 and 2005*. Levy Economics Institute (Working Paper n. 680). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1895637&rec=1&srcabs=1670765&alg=7&pos=5 (accessed August 2018).

Singh, A.C., H. Mantel, M. Kinack, and G. Rowe. 1990. *On Methods of Statistical Matching with and Without Auxiliary Information*. Methodology Branch, Statistics Canada (Technical Report SSMD-90-016E). Available at: http://publications.gc.ca/collections/collection_2017/statcan/11-613/CS11-613-90-16-eng.pdf (accessed April 2015).

Singh, A.C., H. Mantel, M. Kinack, and G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption."

Survey Methodology 19: 59–79. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1993001/article/14475-eng.pdf (accessed April 2015).

Sisto, A. 2006. *Propensity Score Matching: un'Applicazione per la Creazione di un Database Integrato ISTAT-Banca d'Italia*. Dipartimento di Politiche Pubbliche e Scelte Collettive – POLIS, Università del Piemonte Orientale (Working Paper n. 63). Available at: http://polis.unipmn.it/pubbl/RePEc/uca/ucapdv/sisto63.pdf (accessed April 2015).

Slesnick, D. 1993. "Gaining Ground: Poverty in the Postwar United States." *Journal of Political Economy* 101: 1–38. Doi: http://dx.doi.org/10.1086/261864.

Smith, J. and P. Todd. 2005. "Rejoinder." *Journal of Econometrics* 125: 365–375. Doi: http://dx.doi.org/10.1016/j.jeconom.2004.04.013.

Stiglitz, E., A. Sen, and J.P. Fitoussi. 2010. *Report by the Commission on the Measurement of Economic Performance and Social Progress*, Paris. Available at: http://www.stiglitz-sen-fitoussi.fr/en/index.htm (Accessed August 2018).

Tedeschi, S. and E. Pisano. 2013. *Data Fusion Between Bank of Italy-SHIW and ISTAT-HBS*. Munich Personal RePEc Archive (Working Paper n. 51253). Available at: https://mpra.ub.uni-muenchen.de/51253/ (accessed April 2015).

Webber, D. and R. Tonkin. 2013. *Statistical Matching of EU-SILC and the Household Budget Survey to Compare Poverty Estimates using Income, Expenditures and Material Deprivation*. EUROSTAT, Methodologies & Working papers, European Union. Available at: https://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF/37d4ffcc-e9fc-42bc-8d4f-fc89c65ff6b1 (accessed April 2015).

Wolff, E.N. and A. Zacharias. 2003. *The Levy Institute Measure of Economic Well-Being*. The Levy Economics Institute (Working Paper n. 372). Available at: http://www.levyinstitute.org/pubs/wp/372.pdf (accessed April 2015).

Wolff, E.N., A. Zacharias, T. Masterson, S. Eren, A. Sharpe, and E. Hazell. 2012. *A Comparison of Inequality and Living Standards in Canada and the United States Using an Expanded Measure of Economic Well-Being"*. Levy Economics Institute (Working Paper n. 703). Available at: http://www.levyinstitute.org/pubs/wp_703.pdf (accessed April 2015).

Received October 2016
Revised June 2018
Accepted September 2018

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 577–597, http://dx.doi.org/10.2478/JOS-2019-0024

Connecting Correction Methods for Linkage Error in Capture-Recapture

Peter-Paul de Wolf[1], *Jan van der Laan*[1], *and Daan Zult*[1]

A commonly known problem in population size estimation using registers, is that registers do not necessarily cover the whole population. This may be because they intend to cover part of the population (e.g., students), due to administrative delay or because part of the target population is not registered by default (e.g., illegal persons). One of the methods to estimate the population size in the presence of undercount is the capture-recapture method that combines the information of two or more samples. In the context of census estimation registers are used instead of samples. However, the method assumes that perfect linkage between the registers can be achieved. It is known that this assumption is often violated.

In the setting of evaluating the population coverage of a census using a post-enumeration survey, a correction for linkage error was proposed. That correction was later generalized by relaxing some of the newly introduced conditions. However, the new correction method still implicitly assumed that the two registers are of equal size. We introduce a further generalization that includes both previously mentioned correction methods and at the same time deals with registers of different sizes. Specific parameter settings will correspond to the different correction methods. We show that the parameters of each method can be chosen such that the resulting estimates all equal the traditional Petersen estimate (1896) that would theoretically be obtained under truly perfect linkage.

Key words: Population size estimator; undercoverage; probabilistic record linkage.

1. Introduction

Capture-recapture methodology goes back to at least the ecological setting of estimating the size of fish and wildlife populations. The basic idea is to take a first sample (capture), tag or mark the captured animals, return them to their population and take a second sample (recapture). Among the recaptures, some of the animals will be marked, others not. The relation between the tagged and non-tagged animals in the second sample is used to construct an estimate of the total population size. See for example Petersen (1896) and Lincoln (1930). Since then, it has been used not only to estimate animal population sizes, but also to estimate undercount in traditional censuses (for an overview see e.g., Fienberg 1992). More recently, it was used to estimate the undercoverage of registers used for the Dutch Census (Gerritse et al. 2016a).

[1] Statistics Netherlands, P.O.Box 24500, 2490 HA The Hague, The Netherlands. Emails: pp.dewolf@cbs.nl, dj.vanderlaan@cbs.nl, and db.zult@cbs.nl.

Acknowledgments: The authors like to thank Jeroen Pannekoek for reviewing an earlier version of the article. The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

In the original setting, one of the assumptions is that the units can be classified without error to belong to the first sample only the second sample only or the overlap of the two samples. This assumption was likely to be met, when the marking of the units in the first sample would stick to the animals during the second sampling (no tag-loss). In the setting of estimating the undercount of a register, this assumption is translated to the assumption that units in the two registers can be linked without error, that is, all links are found and no erroneous links are established. With linking two records, we mean deciding that those records represent the same population unit. Whenever the both registers contain the same reliable unique identifiers like a social security number, it is likely that this assumption holds. However, not all registers contain such a uniform unique identifier. Actually, when considering undercoverage of registers, one can not rely on the existence of such unique identifiers only. Indeed, in order to find units that are not properly registered one should also use sources that do not have such a unique identifier for all units.

In case when a unique identifier is not available, one often relies on probabilistic record linkage techniques like the one developed in Fellegi and Sunter (1969). In this setting, the assumption of perfect linkage is not likely to be met in practice. Especially in a large population, two individuals might, for example, have the same name, leading to a false link, or one individual might be known under two different names, leading to a missed link. This last case would be identical to tag-loss in a classical capture-recapture setting, while the first case can only occur when tags or id-codes can be passed around within the population of interest.

In the presence of linkage errors, the standard capture-recapture estimate of the unknown population size can be biased see e.g., Gerritse et al. (2016b). In Ding and Fienberg (1994) the standard capture recapture estimator is adjusted to correct for linkage errors. In that paper, they considered the situation where a post-enumeration survey (PES) is used to estimate the undercoverage of the population census. See for example, Wolter (1986) for an explanation of using a PES. Ding and Fienberg (1994) assume that the false match that affects the population size estimator mostly occurs when a record from the subset of the PES that should not be matched is actually linked to a record from the subset of the census that should not be matched. In other words, they assume a one-way linkage error, linking PES records to census records. Moreover, they assume that all records in the PES will be linked to a record in the census. Cadwell et al. (2005) also correct for linkage errors, but they use the concept of 'potential linkage' in a bootstrap procedure to construct a population size interval. Their method is potentially interesting when something like a PES is not available.

Di Consiglio and Tuoto (2015) argue that in the setting of administrative data sources, a one-way linkage direction is not guaranteed. That is, they allow for the possibility that a population unit residing in one administrative data source, but not in the other, can be (incorrectly) linked to a unit in the other administrative data source irrespective of which data source is called 'the one' and which is called 'the other'. Hence, they propose a two-way correction for linkage error. In their paper, they assume that the probability of a false match is equal in both linkage directions. We will call this the symmetric two-way correction for linkage error. Using the same error probability in both directions is appropriate in case the two administrative data sources are (approximately) of equal size.

When two registers differ considerably in size, a further extension would be to allow for different error probabilities in the two linkage directions. This would be even more evident when (forced) one-to-one linkage is used. One-to-one linkage means that a record from one source is allowed to be linked to one and only one record from the other source. Since the largest source contains units that are not present in the smaller source, in case of one-to-one linkage a subset of those units can never be linked; there are simply not enough 'target' records in the smaller source. Records that will never be linked, will also never be linked incorrectly. In other words, a unit in the largest administrative data source has a smaller chance of being falsely linked with a unit in the smaller administrative data source, compared to the other way around. Thus, in the current article we introduce an asymmetric two-way correction for linkage error. The formulation of this asymmetric two-way corrections has three parameters. Choosing specific values for those parameters, the formula can cover the one-way correction and the symmetric two-way correction as well.

The outline of the article is as follows. We start by explaining the general setting of capture-recapture and probabilistic linkage. In Section 3 we briefly present the non-corrected estimator, the one-way corrected estimator, the symmetric two-way corrected estimator and our asymmetric two-way corrected estimator. The formula of the asymmetric two-way correction can be viewed as a general estimator, in the sense that all introduced estimators can be expressed with this formula. Finally, we unify all estimators by choosing specific 'optimal' parameters. The following section, Section 4, shows some simulation results using publicly-available fictitious data on the UK population census. In Section 5 we draw conclusions and the appendices contain some technical details. Section 6 contains Appendices 1–4.

2. General Setting

Let us first introduce the notation that will be used throughout the remainder of this article. We try to stay close to the notation used in Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015). We also present the general assumptions underlying the capture-recapture methodology when applied with two registers. Note that we will only discuss the situation of two registers that are linked using probabilistic record linkage methods.

2.1. *Capture-Recapture with Two Registers*

Let R_1 and R_2 denote two registers containing units from a common population $\mathcal{X}$ of unknown size $N_{\mathcal{X}}$. Assuming we can identify population units to belong to either one or both of the registers, we get Table 1 and Table 2. In Table 1 the numbers correspond to the unobservable true population counts, whereas the numbers in Table 2 are the observed counts *after* the linkage process has taken place and thus depend on the used linkage procedure.

In the tables, the first subscript denotes whether or not a population unit resides in R_1 and the second subscript denotes whether or not a population unit resides in R_2. So, for example, N_{10} denotes the (unobserved) number of population units that resides in R_1 but not in R_2. Note that, assuming no duplicates in each R_i, $n_{1+} = N_1$ is the size of R_1 and $n_{+1} = N_2$ is the size of R_2. Moreover, $N_{\neg i}$ denotes the number of units in the population

Table 1. Counts based on population.

		Unit in R_2		
		yes	no	
Unit in R_1	yes	N_{11}	N_{10}	N_1
	no	N_{01}	N_{00}	$N_{\neg 1}$
		N_2	$N_{\neg 2}$	$N_{\mathcal{X}}$

Table 2. Counts based on linkage process.

		Unit in R_2		
		yes	no	
Unit in R_1	yes	n_{11}	n_{10}	n_{1+}
	no	n_{01}	0	n_{01}
		n_{+1}	n_{10}	n_{++}

that do not reside in R_i, that is, $N_{\neg 1} = N_{\mathcal{X}} - N_1$ and $N_{\neg 2} = N_{\mathcal{X}} - N_2$. Even after the linkage process has taken place, we still cannot observe population units that are included in neither register (i.e., N_{00}). That means that $N_{\neg 1} \geq n_{01}$ and $N_{\neg 2} \geq n_{10}$.

Using similar notation, we can write the probability that a population unit resides in register R_i as p_i, and decompose those probabilities as follows: $p_1 = p_{11} + p_{10}$ and $p_2 = p_{11} + p_{01}$ where p_{11} denotes the probability that a unit resides in both registers, p_{10} the probability that a unit does reside in R_1 but not in R_2 and p_{01} the probability that a unit resides in R_2 but not in R_1.

The general assumptions in capture-recapture estimation are:

- The population $\mathcal{X}$ is closed, that is, units can neither enter nor leave the population during the capture-recapture experiment.
- There are no erroneous captures, that is, only units from $\mathcal{X}$ can be captured.
- There are no duplicates in either register, that is, units can only be captured once per register.
- The event that a unit resides in R_1 is independent of the event that a unit resides in R_2.
- The probability that a unit resides in R_i is the same for all units in $\mathcal{X}$.
- There is no error in allocating the units to R_1, R_2 or both.

These assumptions imply that $N_{11}/N_1 = N_2/N_{\mathcal{X}}$ or equivalently, $N_{\mathcal{X}} = (N_1 N_2)/N_{11}$. Hence, under perfect conditions a natural estimator would be the one introduced in Petersen (1896): $\hat{N}_{\mathcal{X}} = (n_{1+} n_{+1})/n_{11}$. See Subsection 3.1 as well.

2.2. *Probabilistic Record Linkage*

The probabilistic record linkage technique we will assume in this article is the one described in Fellegi and Sunter (1969). In their approach, they consider the set of all

possible pairs (a, b) of records from R_1 and R_2: $\{(a, b) \mid a \in R_1 \text{ and } b \in R_2\}$. They decompose that set into two disjunct sets. Set $\mathcal{M}$ consisting of all pairs of records of matches and set $\mathcal{U}$ of all pairs of records of nonmatches. Hence, for example, a pair (a, b) in the set $\mathcal{U}$ of nonmatches should consist of a record a from register R_1 and a record b from R_2 where a and b refer to two different population units. See Figure 2 in Appendix 1 (Subsection 6.1) for a graphical representation of the sets $\mathcal{M}$ and $\mathcal{U}$.

Fellegi and Sunter then describe a model to decide whether an observed pair of records should be allocated to $\mathcal{M}$ or to $\mathcal{U}$. To that end they use comparison functions that assign a value to a pair indicating the amount of similarity between the two records. For example, in case of personal data, a comparison function could assign a value zero whenever the name of the person of record a is not exactly equal to the name of the person of record b, and a value of one whenever the names are exactly equal. Obviously, this can be more elaborate assigning a value between zero and one in case of small spelling mistakes. Different comparison functions can be applied to different variables within a record, which would result in a comparison *vector*.

Selecting a pair of records at random from all possible pairs, the comparison function applied to that selected pair is a random variable. They define the m-probability as the probability that a certain value of the comparison function is found among a pair of records that should belong to the set $\mathcal{M}$ of matches and the u-probability as the probability that a certain value of the comparison function is found among a pair of records that should belong to the set $\mathcal{U}$ of nonmatches. Using those probabilities, they assign weights to each possible pair and say that a pair of records is linked whenever the weight is above some threshold and not linked whenever that weight is below that threshold. Since this is defined at the level of *pairs* of records, it is possible that several records from register R_1 are said to be linked to the same record in register R_2; whenever a pair has a weight above the threshold, it will be said to be linked. In practice, often a one-to-one linkage is then enforced. One of those pairs is selected and designated to be a link, while the other pairs are considered to be non-links despite their weight being above the threshold.

In their paper, Fellegi and Sunter (1969) consider two error probabilities; the probability of a false link (assigning a pair of records to $\mathcal{M}$ where it should be assigned to $\mathcal{U}$) and the probability of a false non-link (assigning a pair of records to $\mathcal{U}$ where it should be assigned to $\mathcal{M}$). Note that these probabilities are thus defined at the level of *pairs* of records and not on the level of *individual* records. In the description of the correction methods (see Section 3) error probabilities are defined at the level of individual records. To be able to discuss the correction methods for linkage error, it is convenient to decompose our registers R_i each into two disjunct sets M_i and U_i. Now M_i consists of all unique records from register R_i that should appear in a pair of matches and U_i of all other unique records from register R_i. Figure 2 in Appendix 1 (Subsection 6.1) graphically shows the differences between the sets $\mathcal{M}$, $\mathcal{U}$, M_i and U_i.

Linking registers with many records would lead to a huge number of pairs. Under these circumstances a technique known as blocking is often used to improve efficiency. With blocking, the registers are split into subsets that agree on one or more highly discriminating identifiers and the linkage process is applied within each subset separately. For the sake of simplicity, we will not address the use of blocking in the current article, since blocking would affect the (estimation of the) m-probabilities in a complex way.

3. Estimation of the Population Size

In this section, we will first briefly present the existing estimators for the population size under no linkage error, one-way error correction and symmetric two-way error correction. At the end of this section, we will introduce our new asymmetric two-way error correction estimator.

Using the notation from Subsection 2.1, we assume that the number of individuals that fall in the four interior cells of Table 2 have a multinomial distribution

$$(n_{11}, n_{10}, n_{01}, N_{\mathcal{X}} - n_{++}) \sim \text{Mult}(N_{\mathcal{X}}, p_{11}, p_{10}, p_{01}, p_{00})$$

where $n_{++} = \text{n}_{11} + n_{10} + n_{01}$. Like in Ding and Fienberg (1994), we will derive the estimators using the approach of maximizing the conditional likelihood as described in Sanathanan (1972). In that approach, the likelihood is written as a product of two likelihoods $L_1(\cdot)$ and $L_2(\cdot)$, where $L_1(\cdot)$ is the likelihood of (n_{11}, n_{10}, n_{01}) for fixed n_{++} and $L_2(\cdot)$ the likelihood of n_{++}, given the cell-probabilities p_{11}, p_{10} and p_{01}. In the conditional approach, first $L_1(\cdot)$ is maximized to derive the maximum likelihood (ML) estimates of the cell probabilities, after which the $N_{\mathcal{X}}$ is found that maximizes $L_2(\cdot)$, given the values of p_{11}, p_{10} and p_{01}.

Using that $\mathbb{E}(n_{++}) = \mathbb{E}(n_{1+}) + \mathbb{E}(n_{+1}) - \mathbb{E}(n_{11}) = (p_1 + p_2 - p_{11})N_{\mathcal{X}}$, where $\mathbb{E}$ denotes taking expectation, we derive the following generic formulation of an estimator of the population total

$$\hat{N}_{\mathcal{X}} = \frac{n_{++}}{\hat{p}_1 + \hat{p}_2 - \hat{p}_{11}} \tag{1}$$

In the following subsections we will derive conditional ML estimators of the cell probabilities under different linkage error scenarios.

3.1. No Linkage Error

Under independence and perfect linkage, we would have the following equations for the probabilities of recording population units in the different observed counts n_{ij}:

$$p_{11} = p_1 p_2 \tag{2}$$

$$p_{10} = p_1 - p_{11} = p_1(1 - p_2) \tag{3}$$

$$p_{01} = p_2 - p_{11} = p_2(1 - p_1) \tag{4}$$

Using the conditional ML approach we would get the estimators

$$\hat{p}_1 = \frac{n_{11}}{n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11}}{n_{1+}}$$

Plugging those estimators into (2) and (1), the estimator of the population total then becomes after some straightforward calculations

$$\hat{N}_{\mathcal{X}}^{P} = \frac{n_{1+} n_{+1}}{n_{11}} \tag{5}$$

and this is essentially the estimator as described in for example, Petersen (1896).

3.2. One Way Correction (OC)

In Ding and Fienberg (1994) the situation of linkage error is considered under the assumptions that (using the notation as in Subsection 2.2)

(a) A true match between records from M_1 and M_2 remains a match with probability $0 < \alpha \leq 1$.
(b) A record from M_1 is matched incorrectly with a record in M_2 with negligible probability.
(c) A false match between records from M_1 and U_2 occurs with negligible probability.
(d) A false match between records from U_1 and M_2 occurs with negligible probability.
(e) Each record from U_1 will be linked with a record in U_2 with common probability $0 \leq \beta < 1$.

The reason for assuming negligible probabilities for (b), (c) and (d) is that in those cases, two errors are made; both the correct match is not made and an incorrect match is made. Cases (a) and (e) are each related to making only one error: in case of (a) with probability $1 - \alpha$ only a correct match is missed and in case of (e) only an incorrect match is made.

The just introduced probabilities α and β are defined at *record* level. Note that the probabilities in the Fellegi and Sunter (1969) setting (see Subsection 2.2), sometimes also denoted by α and β, are defined at the level of *pairs* of records and are thus fundamentally different from the ones used in the current article. Moreover, note that a large α implies more missed links (in expectation), which in turn leads to an upward bias in the estimator $\hat{N}_{\mathcal{X}}$. A large S implies more false links (in expectation), which would lead to a downward bias in $\hat{N}_{\mathcal{X}}$.

Under the aforementioned assumptions we get the following relations

$$p_{11} = \alpha p_1 p_2 + \beta p_1 (1 - p_2) \tag{6}$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1 (1 - p_2) \tag{7}$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1 (1 - p_2) \tag{8}$$

Note that the 'one-way' correction is reflected in (6); the second term on the right hand side only shows the probability of falsely linking (β) a unit that resides in R_1 (p_1) but not in R_2 ($1 - p_2$). The probability of falsely linking a unit that resides in R_2 but not in R_1 is not considered, that is, only one linkage direction is considered.

The conditional ML estimators are then given by (Ding and Fienberg 1994)

$$\hat{p}_1 = \frac{n_{11} - \beta n_{1+}}{(\alpha - \beta) n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11} - \beta n_{1+}}{(\alpha - \beta) n_{1+}}$$

Plugging this into (6) and (1), the population total then can be estimated by

$$\hat{N}_{\mathcal{X}}^{OC} = \frac{(\alpha - \beta) n_{11}}{n_{11} - \beta n_{1+}} \frac{n_{1+} n_{+1}}{n_{11}} = \frac{(\alpha - \beta) n_{11}}{n_{11} - \beta n_{1+}} \hat{N}_{\mathcal{X}}^{P} \tag{9}$$

with $\hat{N}_{\mathcal{X}}^{P}$ as defined in (5). Note that this estimator depends on the parameters α and β which are unknown in practice and should therefore be estimated. This will be discussed in Subsection 3.5.

3.3. Symmetric Two-Way Correction (SC)

In Di Consiglio and Tuoto (2015) it is proposed to relax the assumption of the one-way correction and to allow a two-way correction. This means that assumption (e), as given in the description of the one-way correction, is relaxed to allow for a unit in U_1 that is not in U_2 still to be (incorrectly) linked to a unit in U_2 as well as to allow for a unit in U_2 that is not present in U_1 still to be (incorrectly) linked to a unit in U_1. Both events occur with the same probability $0 \leq \beta < 1$.

This results in the following equations

$$p_{11} = \alpha p_1 p_2 + \beta p_1(1-p_2) + \beta p_2(1-p_1) \tag{10}$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1(1-p_2) - \beta p_2(1-p_1) \tag{11}$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1(1-p_2) - \beta p_2(1-p_1) \tag{12}$$

Again, under certain regularity conditions and using the conditional likelihood approach, they derive that the ML estimators are then given by

$$\hat{p}_1 = \frac{n_{11} - \beta(n_{1+} + n_{+1})}{(\alpha - 2\beta)n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11} - \beta(n_{1+} + n_{+1})}{(\alpha - 2\beta)n_{1+}}$$

Plugging this into (10) and (1), the population total can then be estimated by

$$\hat{N}_{\mathcal{X}}^{SC} = \frac{(\alpha - 2\beta)n_{11}}{n_{11} - \beta(n_{1+} + n_{+1})} \frac{n_{1+}n_{+1}}{n_{11}} = \frac{(\alpha - 2\beta)n_{11}}{n_{11} - \beta(n_{1+} + n_{+1})} \hat{N}_{\mathcal{X}}^{P} \tag{13}$$

3.4. Asymmetric Two-Way Correction (AC)

As a further relaxation of the assumptions, we propose to allow for different probabilities of false links. This means that we allow for a unit present in U_1 but not present in U_2 to be linked to a unit in U_2 with probability $0 \leq \beta_1 < 1$ and a unit present in U_2 but not present in U_1 to be linked to a unit in U_1 but with probability $0 \leq \beta_2 < 1$.

Now the equations for the probabilities of recording population units in the different observed counts become

$$p_{11} = \alpha p_1 p_2 + \beta_1 p_1(1-p_2) + \beta_2 p_2(1-p_1) \tag{14}$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta_1 p_1(1-p_2) - \beta_2 p_2(1-p_1) \tag{15}$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta_1 p_1(1-p_2) - \beta_2 p_2(1-p_1) \tag{16}$$

Under certain regularity conditions, we then get the following ML estimators

$$\hat{p}_1 = \frac{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}}{(\alpha - (\beta_1 + \beta_2))n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}}{(\alpha - (\beta_1 + \beta_2))n_{1+}} \tag{17}$$

See Appendix 2 (Subsection 6.2) for a discussion on admissibility to obtain proper values for the probabilities $\hat{p}_1$ and $\hat{p}_2$ in the interval [0,1].

Plugging (17) into (14) and (1), the population total can then be estimated by

$$\hat{N}_{\mathcal{X}}^{AC} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \frac{n_{1+}n_{+1}}{n_{11}} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \hat{N}_{\mathcal{X}}^{P} \tag{18}$$

Note that this formulation covers all previous situations by choosing appropriate α, β_1 and β_2

- Petersen estimator: $\alpha = 1$ and $\beta_1 = \beta_2 = 0$
- One-way correction: $\alpha = \alpha$, $\beta_1 = \beta$ and $\beta_2 = 0$
- Symmetric two-way correction: $\alpha = \alpha$, $\beta_1 = \beta_2 = \beta$

3.5. *Linking the Correction Methods*

We consider the Petersen estimator in case of perfect linkage, that is, knowing the true N_1, N_2 and N_{11}, the 'optimal' estimator and call it the 'true Petersen estimator' (TP)

$$N_{\mathcal{X}}^{TP} = \frac{N_1 N_2}{N_{11}} = \frac{n_{1+}n_{+1}}{N_{11}}$$

Equating the AC estimator to the TP estimator, that is, setting $\hat{N}_{\mathcal{X}}^{AC} = N_{\mathcal{X}}^{TP}$, we get the following relationship between the parameters

$$\alpha N_{11} + \beta_1(N_1 - N_{11}) + \beta_2(N_2 - N_{11}) = \alpha N_{11} + \beta_1 N_{10} + \beta_2 N_{01} = n_{11} \tag{19}$$

Note that the left-hand side equals the expected number of links under the model for linkage error.

Let us first explore this relationship under the unrealistic assumption that we know the true N_{11}. A natural choice for the parameter α would then be the fraction of true population matches among the links from the linkage process. We will denote this natural choice by $\breve{\alpha}$. Substituting that natural choice in (19) and setting $\beta_1 = \beta^{OC}$ and $\beta_2 = 0$, we get

$$\alpha^{OC} = \breve{\alpha} = \frac{m_{11}}{N_{11}} \quad \text{and} \quad \beta^{OC} = \frac{n_{11} - m_{11}}{N_1 - N_{11}}$$

where m_{11} is the number of true population matches among the links from the linkage process. We will call this choice of parameters the *optimal OC-parameters*.

In case of the symmetric two-way correction, using the natural choice for α and setting $\beta_1 = \beta_2 = \beta^{SC}$ leads to

$$\alpha^{SC} = \breve{\alpha} = \frac{m_{11}}{N_{11}} \quad \text{and} \quad \beta^{SC} = \frac{n_{11} - m_{11}}{N_1 + N_2 - 2N_{11}}$$

We will call this choice of parameters the *optimal SC-parameters*.

In case of the asymmetric two-way correction, we need an additional constraint to uniquely define 'optimal AC-parameters'. In practice, it is convenient to enforce one-to-one linkage in the process. Under that assumption, we can derive the following relationship between the parameters of the asymmetric two-way estimator (see the Appendix 3, Subsection 6.3, for a derivation)

$$\beta_1 = \frac{(\alpha n_{+1} - n_{11})\beta_2}{(\alpha n_{1+} - n_{11}) - 2\beta_2(n_{1+} - n_{+1})} \tag{20}$$

In case we want to satisfy both (20) and (19) using the natural $\breve{\alpha}$ parameter, we get either

$$\alpha^{AC} = \breve{\alpha} = \frac{m_{11}}{N_{11}}, \quad \beta_1^{AC} = \frac{n_{11} - m_{11}}{2(N_1 - N_{11})} \quad \text{and} \quad \beta_2^{AC} = \frac{n_{11} - m_{11}}{2(N_2 - N_{11})}$$

or

$$\tilde{\alpha}^{AC} = \breve{\alpha} = \frac{m_{11}}{N_{11}}, \quad \tilde{\beta}_1^{AC} = \frac{m_{11}N_2 - n_{11}N_{11}}{m_{11}(N_2 - N_1)} \quad \text{and} \quad \tilde{\beta}_2^{AC} = \frac{m_{11}N_1 - n_{11}N_{11}}{m_{11}(N_1 - N_2)}$$

where m_{11} again is the number of true population matches among the links from the linkage process. For the second set of parameters ($\tilde{\alpha}^{AC}$, $\tilde{\beta}_1^{AC}$ and $\tilde{\beta}_2^{AC}$) it holds that the $\tilde{\beta}$'s will be undefined in case $N_1 = N_2$. Moreover, when $N_1 \neq N_2$, one of them will be negative, which contradicts the fact that the $\tilde{\beta}$'s are probabilities. We will hence call the first set of parameters the *optimal AC-parameters*. Note that, in case register R_1 is the largest and hence under one-to-one linkage $N_1 - N_{11} > N_2 - N_{11}$, we get $\beta_1^{AC} < \beta_2^{AC}$ as expected (see discussion in introduction).

According to the error correction model, a false match between a record from U_1 with a record from U_2 occurs with probability β_1 and, independently, a false match between a record from U_2 with a record from U_1 occurs with probability β_2. Considering these events independently, we would count such a link twice. However, enforcing one-to-one linkage, these two events can only happen at the same time. This is reflected in the factor 1/2 in the *optimal AC-parameters* β_i^{AC}.

Given the true N_{11} and choosing the parameters such that they satisfy Equation (19) would thus lead to the optimal estimator. Indeed, using the *optimal OC-parameters*, the *optimal SC-parameters* or the optimal AC-parameters will all yield the same estimator, that is, the true Petersen estimator TP (with perfect linkage).

Unfortunately, in practice we do not know the true N_{11}. Hence, we need to estimate the α and β_i parameters. As long as the estimated parameters satisfy relation (19), the resulting estimates will be exactly the same for all estimators. This would, for example, be the case when we would estimate the optimal parameters by plugging in some estimate for N_{11}, since N_1, N_2, n_{11} and m_{11} are the same in all settings. Indeed, the resulting estimators would then be given by the simple formula

$$\hat{N}_{\mathcal{X}} = \frac{N_1 N_2}{\hat{N}_{11}} \tag{21}$$

where $\hat{N}_{11}$ is a (consistent) estimator of the 'true' overlap between the two registers.

Another possibility would be to use a sample of one of the registers and determine the true matches for that sample. Dividing that number by the sampling fraction would yield a direct estimate of N_{11}. Similarly, we could obtain direct estimates of n_{11} and m_{11}. Note that a direct estimate of n_{11} is needed instead of the original n_{11} to prevent the estimated m_{11} becoming larger than the original n_{11}.

Yet another approach would be to use expert knowledge on the linkage errors, for example, asking experts to give estimates of the parameters. In that case, these expert guesses would not necessarily satisfy relation (19) and the estimators could thus yield different values.

4. Simulation Study

The evaluation of a linkage error correction method in a capture-recapture setting requires knowledge on the bias introduced by the linkage errors. In real data applications the bias is obviously usually unknown. However, in case of a fictitious population the bias can be determined accurately. Such a fictitious population may be a real or simulated list of individuals that is then defined as the 'true population'. Subsets can then be randomly selected from this list and linked with and without linkage error. This way capture-recapture can be applied with and without linkage error, as well as with and without linkage error correction. Hence, the different linkage error correction methods can be evaluated. Moreover, this procedure can be repeated multiple times with different randomly selected subsets in order to obtain multiple estimates. As a result, the linkage error bias and variance over the repeated estimates can be assessed.

We used the same data set as used by Di Consiglio and Tuoto (2015). This publicly available data set concerns a fictitious population that is based on the UK population census, as created for the ESSnet DI (McLeod et al. 2011). The ESSnet DI was a European project on data integration (Record Linkage, Statistical Matching, Micro Integration Processing), running from 2009 to 2011. The advantage of this data set is twofold. Firstly, it is publicly available, which supports reproducibility of our simulation study. Secondly, it is designed to resemble the situation a researcher might encounter in real life. For instance, the samples contain errors such as typical spelling errors in names or incorrect house numbers. Moreover, it is unclear (to us) how exactly the population samples were generated; for example did some groups have higher probability to be in certain samples or not? The main methodological difference between a real data set and this fictitious population is that the true population size is known, while most other real data issues remain present.

Since ensuring that the parameter estimates satisfy relation (19) will result in the same estimates of the population size for all estimators introduced in Section 3, we will not explicitly compare them. This approach deviates from Di Consiglio and Tuoto (2015), where the estimator based on the SC model is compared to the estimator based on the OC model. They conclude that the SC estimator outperforms both the Petersen and the OC estimator. Their conclusion that the SC estimator, as well as the OC estimator, outperform the Petersen estimator in case of linkage errors is fully justified. However, their result that the SC estimator outperforms the OC estimator seems to contradict our findings in Section 3. This apparent contradiction can be explained by the fact that in Di Consiglio and Tuoto (2015) the parameters α^{SC} and β^{SC} (i.e., the optimal parameters under the SC model) are used in both the SC and OC estimator. We now know that in case of the OC estimator the optimal parameters are α^{OC} and β^{OC} and therefore should have been chosen differently.

In our simulation study, we concentrate on different ways to estimate the parameters instead of comparing the different estimators with one choice of parameter estimates. We will use different methods to estimate N_{11} and m_{11} and plug those estimates into the formulas of our 'optimal' parameters, to show empirically that these estimates indeed lead to the same estimate of the population total and to study the behaviour of that estimate with respect to its bias and variance.

4.1. Setup

From the ESSnet DI data we used the files `Person` (a list of persons, acting as the population), `CIS` (observations from a Customer Information System, being a combination of tax and benefit data) and `PRD` (observations from the Patient Register Data of the National Health Service). The `Person` dataset is comprised of 26,625 individuals, the `CIS` has a coverage probability of that population of $\tau_1 = 0.924$ and the `PRD` of $\tau_2 = 0.930$.

To reduce computation time and to be able to apply the linkage process without blocking, we repeatedly constructed a smaller population and corresponding registers from those files, using the following steps

1. Draw a simple random sample without replacement of size 10,000 from `Person`. This will be our population $\mathcal{X}$ with size $N_{\mathcal{X}}$.
2. Select the records from `CIS` that are present in population $\mathcal{X}$ to get register R_1.
3. Select the records from `PRD` that are present in population $\mathcal{X}$. Randomly select a fraction f_2 of those records to get register R_2.

This way we obtained multiple instances of a population and the corresponding registers, where one of them covers the population for about 92.4% and the other for about f_2 times 93.0%. Note that, for small values of f_2, the two registers differ substantially in size.

In the above mentioned setup, one of the registers is quite close in size to the total population: R_1 covers the population for about 92.4%. This resembles the situation where an NSI has a high quality population register and wants to quantify the quality of that register in terms of estimated undercoverage. A different situation may be that no such high quality register exists. In that case neither register used in the capture-recapture procedure has high coverage. To get an idea how the error correction methods work in that case we included two additional scenarios. We maintained the small coverage of R_2 (using $f_2 = 0.15$) but additionally we reduced the coverage of R_1 by randomly selecting a fraction f_1 of its records, where $f_1 \in \{0.15, 0.5\}$. Table 3 shows the mean coverage of the registers we obtained in our simulations.

In Di Consiglio and Tuoto (2015) several linkage scenarios were mentioned: a bronze, a silver and a gold scenario. In the current paper we will only use their silver scenario, that is, we only use the full date of birth (day (`DB_D`), month (`DB_M`) and year (`DB_Y`)) as key variables in the linkage process. We have chosen the silver scenario because it allows for two types of linkage error. Firstly, two different individuals may have the same date of birth and therefore may be falsely linked. Secondly, due to some measurement errors, an individual that is in both samples may be falsely not linked.

Table 3. Mean coverage $\bar{\tau}_i$ of register R_i over the 100 replications.

f_1	f_2	$\bar{\tau}_1$	$\bar{\tau}_2$
1.00	0.90	92.42%	83.65%
1.00	0.50	92.46%	46.48%
1.00	0.15	92.44%	13.93%
0.50	0.15	46.23%	13.94%
0.15	0.15	13.87%	13.94%

Names and surnames would have been better discriminating identifiers, but in the absence of those variables (e.g., due to privacy restrictions), the full date of birth is still reasonably well discriminating.

For the comparison function of the probabilistic record linkage process (see Subsection 2.2), we simply used 'equality' on all key variables separately. That is, whenever two records a and b are compared, the comparison function for key variable V_i is 1 when $V_i(a) = V_i(b)$ or 0 when $V_i(a) \neq V_i(b)$. Whenever V_i is missing in at least one of the two records, the comparison function is defined to be 0 as well. To perform the probabilistic record linkage as described in Subsection 2.2, we used our own R-code. In that code we also forced one-to-one linkage. (See https://github.com/djvanderlaan/reclin for the R-package `reclin` that we used.)

We implemented four methods to obtain values for the N_{11}, m_{11} and n_{11} needed in the formulas for the 'optimal' parameters.

A Since we use simulated data, we know the true m_{11} and N_{11} by design. The n_{11} follows from the linkage process.

B Using the EM-algorithm (see e.g., Herzog et al. 2007) on the complete registers to estimate the posterior m-probabilities. Those posterior probabilities were used to estimate the m_{11} and N_{11}. The n_{11} follows from the linkage process.

C Using a sample of size 200 from the smallest register of which we determine the true match-status. Using that sample, we fitted a logistic model (see the Appendix 4 (Subsection 6. 4) for more information on the used model) and used that to predict the m-probabilities for the complete registers. Those posterior probabilities were used to estimate the m_{11} and N_{11}. The n_{11} follows from the original linkage process.

D Using a sample of size 200 from the smallest register of which we determine the true match-status. Using that sample we calculated the direct estimates of n_{11}, N_{11} and m_{11} for the complete registers.

In methods B and C, summing the posterior m-probabilities over all linked pairs yields an estimate of m_{11}, whereas summing those probabilities over all possible pairs yields an estimate of N_{11}. For a definition of posterior m-probabilities and why summing them is appropriate, we refer to Fellegi & Sunter (1969). Methods B, C and D serve to illustrate how information available during an actual record linkage process can be used to correct the estimator for linkage errors. As long as the sample used in methods C and D is a representative sample of possible record pairs, these methods should give unbiased estimates of N_{11}, m_{11} and n_{11}. Other methods or refinements of these methods that might give more precise estimates are possible. However, finding such refinements is not the main focus of this article; we want to show that even relatively simple methods can already correct for bias due to linkage errors.

With those estimated sizes, we then used the formulas for the 'optimal' parameters as derived in Subsection 3.5 to get estimates of the population size. As discussed in that section, we should then obtain exactly the same estimates for all approaches (OC, SC and AC).

An additional advantage of using a fictitious population is that, beside knowing the true population size, we can also calculate the TP estimator; the Petersen estimator with truly no linkage error. Since this is the maximum likelihood estimator using population

information, the resulting estimate is the best estimate one could get. Thus, we use the TP estimator as a benchmark for the other estimators in our simulations. The TP estimator is based on the counts in Table 1 and does not equal the Petersen estimator one would get in practice using the counts from Table 2.

4.2. Results

We performed 100 replications of the procedure mentioned in the previous subsection and, as expected, we indeed found that all 'optimal' parameters led to the same estimates in all four methods. In Table 4 the mean, median and standard deviation over the 100 replications is given for the difference between the estimates of the population size and the actual population size $N_{\mathcal{X}} = 10{,}000$, for the estimators TP (method A, the benchmark), P (Petersen, using the counts from the linkage process), EM (method B), model (method C) and sample (method D). Note that TP and P are both based on Petersen's formula (Petersen 1896), but TP is using the (in practice unobservable) true population counts, whereas P uses the observed counts.

The first thing to notice is that the Petersen estimator using the observed counts indeed can lead to a heavily biased estimate of the population size, due to the linkage errors that are present. Moreover, we see that the EM-based estimator (method B) has a very large variance compared to the other estimators and at the same time has a larger bias. This indicates that this method is not well suited to be used for correcting linkage error.

Varying the relative size of the second register (i.e., the f_2) does not really influence the correction for linkage error. Indeed, the bias, as well as the variance of those estimators, seems to be more or less the same in all situations.

Table 4. Mean, median and variance of the difference with $N_{\mathcal{X}} = 10{,}000$ of each estimator over the 100 replications, for different values of f_1 and f_2, and sample size 200.

f_1	f_2		TP $- N_{\mathcal{X}}$	P $- N_{\mathcal{X}}$	EM $- N_{\mathcal{X}}$	model $- N_{\mathcal{X}}$	sample $- N_{\mathcal{X}}$
1.00	0.90	mean	8.0	1399.0	−1730.2	18.7	30.0
		median	9.2	1400.2	−1812.5	23.0	13.4
		st. dev.	10.5	38.1	773.9	261.6	224.9
	0.50	mean	10.6	1193.1	−1810.0	19.1	−5.2
		median	12.6	1187.3	−1891.6	33.8	−28.0
		st. dev.	32.0	66.1	773.9	243.8	205.3
	0.15	mean	22.6	1053.5	−2039.9	22.8	37.5
		median	17.2	1051.8	−2109.6	−1.1	31.8
		st. dev.	75.9	125.3	1326.7	259.1	198.2
0.50	0.15	mean	−17.8	−129.5	−1984.0	−51.4	37.3
		median	−28.5	−147.5	−2015.3	−150.4	−69.0
		st. dev.	269.1	258.6	1609.9	611.4	763.9
0.15	0.15	mean	66.1	−861.3	−2638.6	212.5	450.3
		median	48.4	−869.5	−2683.5	183.7	299.3
		st. dev.	624.3	546.4	2116.4	1177.4	1709.6

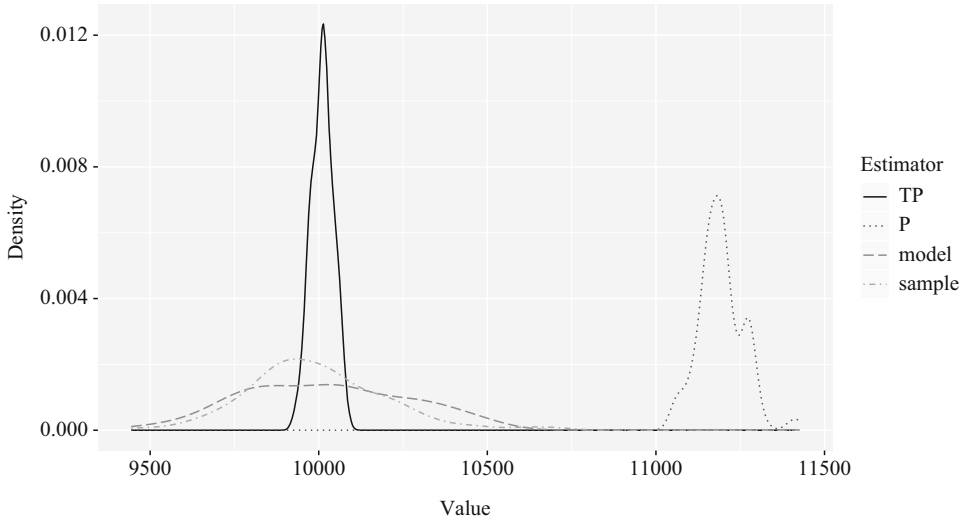

Fig. 1. Distributions of the P, TP, model based and sample based estimators for $N_{\mathcal{X}} = 10{,}000$, $f_1 = 1.0$, $f_2 = 0.5$ and sample size 200.

In case the registers include a unique identifier for some of the records, the identifier could be used as an alternative for taking a sample, under the assumption that the absence of the identifier is not (too) selective. When such a unique identifier is not present, it could, in practice, be quite costly to determine the true match status of pairs. Hence, probably only a small sample would be considered by a National Statistical Institute and that is why we used a relatively small sample from the second register for methods C and D.

Figure 1 shows a smooth estimate of the distribution of the estimators for $f_1 = 1.0$ and $f_2 = 0.5$. For the other values of f_2 the distributions look similar. We did not plot the EM-based estimator in this figure to be able to see more clearly the differences between the other estimation methods.

The figure again shows clearly that the Petersen estimator using the counts from the linkage process has a large bias (due to linkage error) and that the model and sample estimators nicely correct for that. The TP estimates are obviously performing the best, since they use the true knowledge about the number of matches. However, in practice that estimator is not possible.

Finally, we observe that in case both registers have low coverage ($f_1 = f_2 = 0.15$, that is, both registers contain less than 1400 records), the variance of the estimators that correct for linkage error increases considerably. This may not be that surprising when using such relatively small registers. However, the estimates that are supposed to correct for linkage error are still closer to the TP estimate than the Petersen estimate. Hence, we conclude that the correction methods work in all cases, albeit that they work best where at least one of the registers has high coverage.

5. Conclusions

In estimating the population size using capture-recapture, linkage errors (false links and missed links) affect the Petersen estimator. Indeed, the Petersen estimator then becomes

heavily biased. To reduce that bias, some correction methods have been proposed in the literature. These methods introduce some additional parameters that should reflect the probability of occurrence of the two possible types of linkage error. They then model how linkage errors occur and use those error-probabilities to incorporate that model into the estimator. In the current article, we introduced a general formulation for such a correction method. That general formulation incorporates all previously introduced correction methods of that type as special cases.

Looking more closely to the general correction method, it turned out that the parameters could actually be chosen in such a way that the general estimator equals the optimal estimator, that is, the Petersen estimator with a known number of true matches. These optimal parameters can be estimated using different methods. We have shown that for at least two methods, the results improve the traditional Petersen estimator considerably. Those two methods make use of a relatively small sample for which the true match status of the records needs to be determined. More refined methods might even improve more and lead to estimators with smaller variances.

We have shown that it is possible to chose optimal parameters, such that all adjustment methods lead to exactly the same estimates. This reduces the need for making a choice on the error linkage model. However, in case the probabilities are estimated in a different way (e.g., by means of expert opinions), the different linkage error models will lead to different estimates. We have not investigated this further in the current article.

In case it is not possible to make use of a sample to estimate the optimal parameters, the general correction method could still be useful. In that situation, the model for the occurrence of the linkage errors should be assessed to estimate the error probabilities. We would like to note that the model assumes that double errors occur with negligible probabilities. With double errors we mean errors like missing a true match of a record and at the same time linking that record incorrectly to some record in the other register. In estimating the error-probabilities this should be taken into account in some way, because in practice such double errors do occur and would influence the error probabilities.

Using covariates or linking more than two registers would lead to more elaborate methods to estimate the population size in the presence of undercoverage. In these cases, more complex loglinear or Poisson models can be used to obtain a capture-recapture estimate. Similarly, the Fellegi and Sunter based linkage procedure can also be applied more elaborately, for example, by making use of blocking(s). This would affect the (estimates of the) posterior m-probabilities. In our view, the ideas expressed in the current article, as well as the introduced general formulation of the linkage error correction methods, will lead to a better understanding of the implications of such extensions and will be of help in deriving new, linkage error correcting, consistent estimators of the population size.

6. Appendix

6.1. Appendix 1: Sets Defined in the Setting of Probabilistic Record Linkage

Let R_1 be a register with records numbered $\{1, 2, 3, \ldots, 10\}$ and R_2 a register with records numbered $\{1, 2, 3, \ldots, 15\}$. The total number of *pairs* (a, b) that can be constructed from the *records* of those registers is $10 \times 15 = 150$. Figure 2 shows all possible pairs.

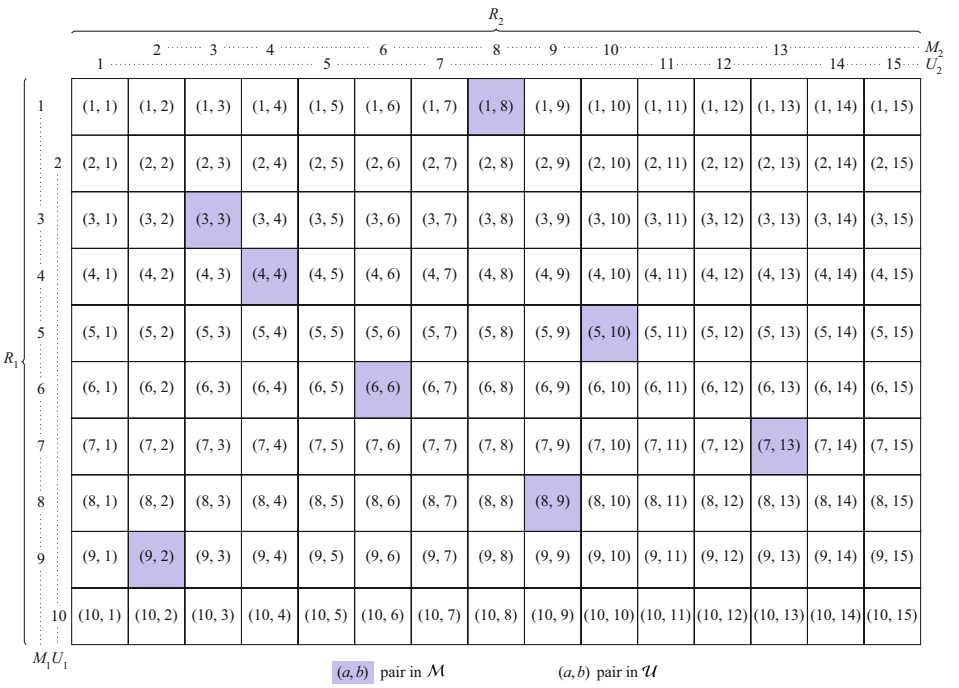

Fig. 2. Graphical representation of the sets of pairs defined in Subsection 2.2.

Moreover, an example of the set $\mathcal{M}$ of pairs of matching records and the set $\mathcal{U}$ of pairs of non-matching records is shown in that figure. In the example, the number of pairs in $\mathcal{M}$ is 8 and the number of pairs in $\mathcal{U}$ is 142.

We can write each register as the union of two disjoint sets, $R_i = M_i \cup U_i$, where the disjoint sets of unique records are given by

$$M_1 = \{1, 3, 4, 5, 6, 7, 8, 9\} \qquad U_1 = \{2, 10\}$$
$$M_2 = \{2, 3, 4, 6, 8, 9, 10, 13\} \qquad U_2 = \{1, 5, 7, 11, 12, 14, 15\}$$

6.2. Appendix 2: Admissibility of Asymmetric Two-way Correction Estimators $\hat{p}_i$

The estimators for the probabilities p_i in case of the asymmetric two-way correction approach should obviously be within [0, 1]. This puts some restrictions on the parameters α, β_1 and β_2.

To ensure that the estimators are non-negative, straightforward calculations lead to the condition that either

$$\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \quad \text{and} \quad \beta_1 + \beta_2 < \alpha \tag{22}$$

or

$$\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \quad \text{and} \quad \beta_1 + \beta_2 > \alpha \tag{23}$$

Additionally, ensuring that both probabilities are not larger than one, leads under (22) to the condition

$$\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \wedge n_{+1}) \tag{24}$$

and under (23) to the condition

$$\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \vee n_{+1}) \tag{25}$$

where $n_{1+} \vee n_{+1}$ equals the maximum of n_{1+} and n_{+1} and $n_{1+} \wedge n_{+1}$ the minimum of n_{1+} and n_{+1}.

Summarizing, we need either

$$\left.\begin{aligned} &\beta_1 + \beta_2 < \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \wedge n_{+1}) \end{aligned}\right\} \tag{26}$$

or

$$\left.\begin{aligned} &\beta_1 + \beta_2 > \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} - (\alpha - (\beta_1 + \beta_2))(n_{1+} \vee n_{+1}) \end{aligned}\right\} \tag{27}$$

Assuming R_1 to be the largest data set, that is, $n_{1+} > n_{+1}$, the set of conditions (26) is equivalent to

$$\left.\begin{aligned} &\beta_1 \geq (n_{11} - \alpha n_{+1})/(n_{1+} - n_{+1}) \\ &\beta_1 + \beta_2 < \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \end{aligned}\right\} \tag{26$'$}$$

and the set of conditions (27) to

$$\left.\begin{aligned} &\beta_2 \geq (\alpha n_{1+} - n_{11})/(n_{1+} - n_{+1}) \\ &\beta_1 + \beta_2 > \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \end{aligned}\right\} \tag{27$'$}$$

Assuming the two data sets to be of equal size, that is, $n_{1+} = n_{+1}$, the set of conditions (26) is equivalent to

$$\left.\begin{aligned} &\alpha \geq n_{11}/n_{1+} \\ &\beta_1 + \beta_2 < \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \end{aligned}\right\} \tag{26$''$}$$

and the set of conditions (27) to

$$\left.\begin{aligned} &\alpha \leq n_{11}/n_{1+} \\ &\beta_1 + \beta_2 < \alpha \\ &\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \end{aligned}\right\} \tag{27$''$}$$

6.3. Appendix 3: Enforcing One-To-One Linkage

In our asymmetric two-way correction method, we have three parameters: α, β_1 and β_2. In case we enforce one-to-one linkage, we can actually do with two, because in that situation we can write β_1 as a function of α and β_2.

In Figure 3 the relation between (expected) counts based on the population and based on linkage are shown in the situation where we potentially would like to apply the asymmetric two-way correction with enforced one-to-one linkage. Under the assumption of one-to-one linkage, it should hold that $\beta_1 N_{10} = \beta_2 N_{01}$, as can be seen in the figure. Noting that $\mathbb{E}N_{10} = p_1(1-p_2)N_{\mathcal{X}}$ and $\mathbb{E}N_{01} = p_2(1-p_1)N_{\mathcal{X}}$ and plugging in the estimators $\hat{p}_1$ and $\hat{p}_2$ from (17), we can derive the following relation

$$\beta_1 = \frac{(\alpha n_{+1} - n_{11})\beta_2}{(\alpha n_{1+} - n_{11}) - 2\beta_2(n_{1+} - n_{+1})} \tag{28}$$

Note that, assuming equal sizes of the two registers, i.e., $n_{1+} = n_{+1}$, Equation (28) yields $\beta_1 = \beta_2$. That is, we would obtain the situation in which the symmetric two-way correction is applicable.

Moreover, from (28) it follows that

$$\alpha > 2\beta_2 \text{ and } n_{1+} > n_{+1} \Rightarrow \beta_1 < \beta_2$$
$$\alpha > 2\beta_2 \text{ and } n_{1+} < n_{+1} \Rightarrow \beta_1 > \beta_2$$

as expected (see discussion in Section 1).

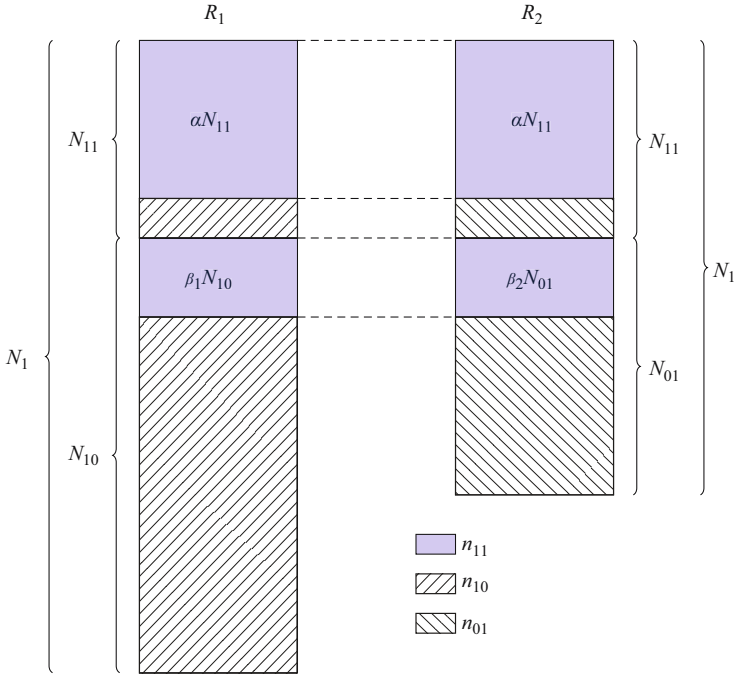

Fig. 3. Relations between counts based on population and based on one-to-one linkage.

6.4. Appendix 4: Estimation of the Matching Probabilities Using Logistic Regression

For a sample of records from the smallest register it is assumed that the true match status can be determined, that is, we assume that it is known whether or not the record should be linked to a record from the larger register and if so, with which record it should be linked. Therefore, for a subset of all pairs generated in the linkage process, the true match status is known. The goal of the logistic regression model is to predict the probability that this pair is a true match, based on properties of the record pair.

In the regression model the following covariates are used

1. The result of the comparison of the linkage variables. In this case the linkage variables are the three elements of the date of birth; day (`DB_D`), month (`DB_M`) and year (`DB_Y`). These variables are binary; both records of the pair agree on the variables (true) or not (false). If in at least one of the records a variable is missing, we consider it a disagreement (false).
2. Whether or not the pair is selected when enforcing one-to-one linkage (`LNK`). This is also a binary variable which is false when there is a more likely match for one or both of the records. This variable is a strong predictor for true matches.

The target variable is the true match status (a binary variable). The model is estimated using the sampled pairs and then used to calculate predictions of the matching probability for all pairs. The main goal of the model is to correct for differences in the population between the sample and the complete set of pairs. This should result in a more accurate estimate of the number of linkage errors. As the model is merely used as an illustration of a method that can be used to estimate the relevant parameters for the correction methods, the model is kept relatively simple. Therefore, all variables are added as main effects and no interactions are used in the model.

To estimate m_{11} the probability that a pair is a true match given that a pair has been linked is needed, and to estimate N_{11} the probability of a true match given that a pair has been linked or has not been linked is needed. Therefore, as long as the sample is representative for the set of pairs, using only `LNK` should be enough to obtain unbiased estimates of N_{11} and m_{11} (n_{11} follows directly from the linkage procedure). Adding additional variables to the regression model, such as `DB_D`, `DB_M` and `DB_Y` in this case, could lead to a reduction of the variance of the estimators when the probability of a false link depends on this variable and when the sample is not representative with respect to these variables. Other variables (sex and age) were investigated, but as these did not affect the outcomes in this study, these are not included in the final results.

7. References

Cadwell, B.L., P.J. Smith, and A.L. Baughman. 2005. "Methods for Capture-Recapture Analysis When Cases Lack Personal Identifiers." *Statistics in Medicine* 24(13): 2041–2051. Doi: https://doi.org/10.1002/sim.2081.

Di Consiglio, L. and T. Tuoto. 2015. "Coverage Evaluation on Probabilistically Linked Data." *Journal of Official Statistics* 31(3): 415–429. Doi: https://doi.org/10.1515/jos-2015-0025.

Ding, Y. and S.E. Fienberg. 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error." *Survey Methodology* 20: 149–158. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994002/article/14422-eng.pdf (accessed June 2019).

Fienberg, S.E. 1992. "Bibliography on Capture-Recapture Modelling with Application to Census Undercount Adjustment." *Survey Methodology* 18: 143–154. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1992001/article/14494-eng.pdf (accessed June 2019).

Fellegi, I.P. and A.B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64: 1183–1210. Doi: https://doi.org/10.1080/01621459.1969.10501049.

Gerritse, S.C., B.F.M. Bakker, P.P. de Wolf, and P.G.M. van der Heijden. 2016a. "Under Coverage of the Population Register in the Netherlands, 2010." *Discussion paper 2016-02 (Centraal Bureau voor de Statistiek, Den Haag/Heerlen)*. Available at: https://www.cbs.nl/nl-nl/achtergrond/2016/08/under-coverage-of-the-population-register-in-the-netherlands-2010 (accessed June 2019).

Gerritse, S.C., B.F.M. Bakker, D. Zult, and P.G.M. van der Heijden. 2016b. The Effects of Imperfect Linkage and Erroneous Captures on the Population Size Estimator, Chapter 3 of PhD thesis *An Application of Population Size Estimation to Official Statistics*, S.C. Gerritse, ISBN 978-94-6233-323-9. Available at: https://dspace.library.uu.nl/bitstream/handle/1874/337476/Gerritse.pdf (accessed June 2019).

Herzog, T.N., F.J. Scheuren, and W.E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. Springer. Doi: https://doi.org/10.1007/0-387-69505-2.

Lincoln, F.C. 1930. "Calculating Waterfowl Abundance on the Basis of Banding Returns." *United States Department of Agriculture Circular* 118: 1–4. Available at: https://archive.org/details/calculatingwater118linc/page/n1.

McLeod, P., D. Heasman, and I. Forbes. 2011. Simulated Data for the on the Job Training, ESSnet DI. Available at: https://ec.europa.eu/eurostat/cros/content/job-training_en (accessed 15 April 2019).

Petersen, C.G. J. 1896. "The Yearly Immigration of Young Plaice into the Limfjord from the German Sea." *Report of the Danish Biological Station (1895)* 6: 5–84. Available at: https://www.biodiversitylibrary.org/ia/reportofdanishbi06dans#page/8/mode/1up (accessed June 2019).

Sanathanan, L. 1972. "Estimating the Size of a Multinomial Population." *The Annals of Mathematical Statistics* 43(1): 142–152. Doi: https://doi.org/10.1214/aoms/1177692709.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: https://doi.org/10.1080/01621459.1986.10478277.

Received July 2018
Revised January 2019
Accepted April 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 599–624, http://dx.doi.org/10.2478/JOS-2019-0025

Imprecise Imputation: A Nonparametric Micro Approach Reflecting the Natural Uncertainty of Statistical Matching with Categorical Data

Eva Endres[1], *Paul Fink*[1], *and Thomas Augustin*[1]

Statistical matching is the term for the integration of two or more data files that share a partially overlapping set of variables. Its aim is to obtain joint information on variables collected in different surveys based on different observation units. This naturally leads to an identification problem, since there is no observation that contains information on all variables of interest.

We develop the first statistical matching micro approach reflecting the natural uncertainty of statistical matching arising from the identification problem in the context of categorical data. A complete synthetic file is obtained by imprecise imputation, replacing missing entries by *sets* of suitable values. Altogether, we discuss three imprecise imputation strategies and propose ideas for potential refinements.

Additionally, we show how the results of imprecise imputation can be embedded into the theory of finite random sets, providing tight lower and upper bounds for probability statements. The results based on a newly developed simulation design – which is customised to the specific requirements for assessing the quality of a statistical matching procedure for categorical data – corroborate that the narrowness of these bounds is practically relevant and that these bounds almost always cover the true parameters.

Key words: Data fusion; data integration; finite random sets; hot deck imputation; (partial) identification.

1. Introduction

Nowadays, a tremendous amount of data is readily accessible, as generated by researchers, companies, and governments. Thus, instead of collecting new data to answer research questions, it is a more convenient alternative to use already available data sources. However, there is often no single data source that includes all information of interest. Statistical matching (also called data integration or data fusion) furnishes a method with which researchers can integrate data collected in different surveys. For example, it was applied by Serafino and Tonkin (2017) to statistically match the data of the *EU Statistics on Income and Living Conditions* and the *Household Budget Survey*.

Assume that we are interested in three blocks of variables, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, while there are two data files, A and B, available. Data file A contains n_{A} observations of $(\boldsymbol{X}, \boldsymbol{Y})$, and data

[1] Ludwig-Maximilians-University of Munich, Ludwigstrasse 33, 80539 Munich, Germany. Emails: eva.endres@stat.uni-muenchen.de, paul.fink@stat.uni-muenchen.de, and augustin@stat.uni-muenchen.de.

Acknowledgments: We thank the associate editor for valuable remarks. We also want to express our gratitude to all participants of the research seminar of the Foundations of Statistics Group at LMU Munich for valuable discussions on this work. The first author also thanks the LMU Mentoring program for financial support.

file B contains n_B observations of ($\boldsymbol{X}$, $\boldsymbol{Z}$). The observations in B come from the same population but are disjoint from the observations in A. The aim of statistical matching, namely the gain of joint information about variables not jointly observed, is twofold (e.g., D'Orazio et al. 2006b, 2):

(i) the estimation of the joint distribution of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ or any of its characteristics (*macro approach*), and/or
(ii) the creation of a synthetic data file with complete observations on $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ (*micro approach*).

As the schematic representation in Figure 1 suggests, statistical matching can be interpreted as a missing data problem. The observations of the *specific variables* $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are missing in a special block-wise pattern in A ∪ B, which denotes the union of the two available data files. Following, for example, D'Orazio et al. (2006b, 6), the missingness is induced by the given allocation to a certain data file, and thus the missing data mechanism in the framework of statistical matching can convincingly be assumed to be missing completely at random. However, this absence of joint information on all variables leads to A severe identification problem: the parameters that concern the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are not directly estimable from A ∪ B. Throughout the article, we use the term *parameter* to refer to a component of the (joint) probability distribution.

For instance, D'Orazio et al. (2006b) show various ways to remedy the issue of non-identifiability. On the basis of their underlying concepts, these methods can be allocated into three basic groups: Approaches which

(i) assume the conditional independence of the specific variables given the *common variables* $\boldsymbol{X}$, in order to achieve a factorisation of the joint distribution whose components are estimable on A ∪ B,
(ii) require auxiliary information in terms of a third file or other external information about parameters concerning the relationship of $\boldsymbol{Y}$ and $\boldsymbol{Z}$,
(iii) refrain from aiming at precise point estimates and account for the uncertainty of the statistical matching problem by estimating a set of plausible parameters, resulting in lower and upper bounds for the parameters concerning the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. These estimates can be interpreted as set-valued point estimates, not to be confused with confidence regions.

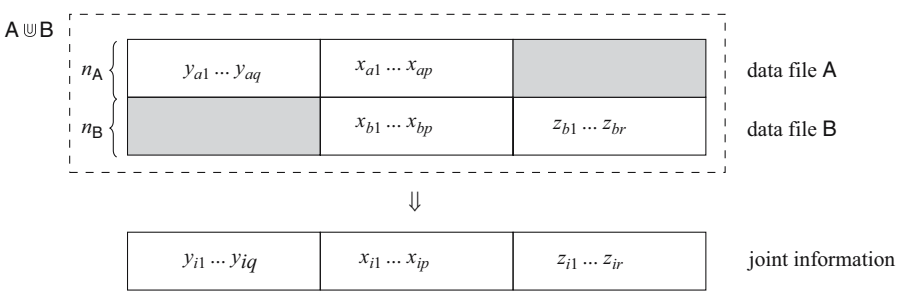

Fig. 1. Schematic representation of the statistical matching problem (See D'Orazio et al. 2006b, 5 (modified)).

In practice, it is not testable whether the conditional independence assumption holds, and in most applications it might be contested. Manski's *Law of Decreasing Credibility* (Manski, 2007, 3), which states that the maintenance of unjustified assumptions reduces the credibility of analyses, makes a very strong argument against the first group of approaches. Auxiliary information, which is the basis of the second group of approaches, is often not available for a certain statistical matching task. Hence, applying statistical matching, taking the underlying uncertainty credibly into account, is the means of choice in these situations.

In the context of statistical matching, typically the term *uncertainty* refers to the previously mentioned identification problem. It points to the fact that even if we have complete information on the marginal distributions of $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Z})$, the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ cannot uniquely be determined (e.g., D'Orazio et al. 2006a). Thus, lower and upper bounds on the parameters (i.e., probability components) are the best that can be obtained without relying on strong untestable assumptions or external information.

Elaborating the concept of uncertainty and how to measure it formed the central focus of the papers by Conti et al. (2012, 2017). Much of the current literature on uncertainty regarding the statistical matching task pays attention to the continuous case, especially to normally distributed variables (e.g., D'Orazio et al. 2006b; Rässler 2002; Ahfock et al. 2016). However, there is also a relatively small body of literature that is concerned with categorical data. For instance, D'Orazio et al. (2006a), Vantaggi (2008), and Di Zio and Vantaggi (2017) deal with statistical matching of categorical data considering different circumstances.

As emphasised by Conti et al. (2012, 70), the "third group of techniques" reflecting the natural uncertainty of statistical matching, does not [usually] "directly aim at reconstructing a complete data set". In the present article, we introduce imprecise (single) imputation as the first micro approach for categorical data that directly accounts for the natural uncertainty of statistical matching. It is based on the imputation of *sets* of plausible values, which leads to a complete synthetic data file with partially set-valued observations. Furthermore, embedding imprecise imputation into the framework of *finite random sets* will allow us to derive lower and upper bounds for the parameters of the joint distribution. As we will highlight, imprecise imputation can be interpreted as a generalisation of multiple hot deck imputation (e.g., Little and Rubin 2002) and fractional hot deck imputation (e.g., Kim and Fuller 2004). The bounds, which we obtain by our imprecise imputation procedure, envelop the results from multiple hot deck imputation and fractional hot deck imputation.

The article is structured as follows. Section 2 recalls the background of our work by giving a brief overview of the basic setting of statistical matching, its interpretation as a missing data problem, and hot deck imputation in this context. Section 3 describes the idea of imprecise imputation and introduces three imputation procedures. Subsequently, in Section 4, we embed imprecise imputation into the theory of finite disjunctive random sets and show how it can be utilised to estimate lower and upper bounds for the parameters of interest from our imputed data file. After providing the setting and results of a simulation study in Section 5, we conclude with a summary and outlook in Section 6. The appendix (Section 7) contains a more detailed description and

justification of the design of the simulation study and graphics on the results of the simulation study.

2. Statistical Matching

2.1. The Basic Setting and its Missing Data Interpretation

Let us assume that we have two data files, A and B, indexed by $\mathcal{I}_A$ and $\mathcal{I}_B$, respectively, with n_A and n_B disjoint observation units. Without loss of generality, we assume that the index sets are disjoint: $\mathcal{I}_A = \{1, \ldots, n_A\}$ and $\mathcal{I}_B = \{n_A+1, \ldots, n_A + n_B\}$. Furthermore, let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be the vector of common variables, and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_r)$ be the vectors of specific variables. Denote the domains of the possible values of X_l, $l = 1, \ldots, p$, by $\mathcal{X}_l$, their corresponding Cartesian product by $\mathcal{X}$, and proceed analogously for the specific variables, defining $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$, $\mathcal{Z}_1, \ldots, \mathcal{Z}_r$, as well as $\mathcal{Y}$ and $\mathcal{Z}$.

As displayed in Figure 1, data file A exclusively contains information on $(\boldsymbol{X}, \boldsymbol{Y})$ as observations $(\boldsymbol{x}_a, \boldsymbol{y}_a)_{a \in \mathcal{I}_A}$, while data file B comprises information on $(\boldsymbol{X}, \boldsymbol{Z})$ only, as observations $(\boldsymbol{x}_b, \boldsymbol{z}_b)_{b \in \mathcal{I}_B}$. Consequently, there is no observation that contains simultaneous information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. In the following, the available information will be consolidated in the incomplete sample A ⋃ B, representing the union of files A and B (see Figure 1) with $n := n_A + n_B$ observations, indexed by $\mathcal{I} = \mathcal{I}_A \cup \mathcal{I}_B$.

Furthermore, we assume that all observations are independently and identically distributed, each following the joint probability distribution $P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z})$, where the realisations for a certain observation $i \in \mathcal{I}$ are depicted as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})$, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ir})$. By collecting all probability components of the underlying distribution, we derive the parameter vector consisting of the probability entries of the multidimensional probability table of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$.

As previously mentioned, statistical matching may be regarded as a missing data problem. Hence, a natural strategy to solve the statistical matching task is imputation, that is, the substitution of the missing entries with suitable real or artificial values to derive a complete (but partially synthetic) data file. To prepare our method, in the following section we focus on *hot deck imputation*, where the missing entries of an observation (*recipient*) are replaced by records from a similar observation (*donor*) of the same sample. Hot deck imputation ensures that only *live* values, that is, actually observed and no artificial values, are substituted, and that the marginal and conditional distributions are preserved well for large samples (e.g., Conti et al. 2008). Hot deck imputation methods are frequently used in practice, comparatively easy to apply, and nonparametric (e.g., Andridge and Little 2010); for a general missing data case, see, for example, Little and Rubin (2002, 66).

2.2. Hot Deck Imputation for Statistical Matching

In the context of statistical matching, hot deck imputation belongs to the group of nonparametric micro approaches. In the following, we will recall and formalise an example for four variables (X_1, X_2, Y_1, Z_1) from D'Orazio et al. (2006b, Chap. 2.4) and also explain our notation. The data samples A and B are assigned to the roles of *recipient file* and *donor file*. Since it is a symmetric problem, D'Orazio et al. (2006b) only describe the

case where A is the recipient file and B the donor file. The reverse case works analogously. The choice of whether only A, only B, or A ⋃ B should be imputed depends on many factors. In this article, we impute A ⋃ B without loss of generality. See, for instance, D'Orazio et al. (2006b, 35–36) for a discussion on this issue.

Random hot deck imputation means that for each missing entry in the recipient file, a donor record from the donor file is randomly chosen by simple random sampling and its corresponding values are used to replace the missing entries in the recipient file. Every missing entry of the specific variable Z_1 in the recipient file A, that is, z_{a1}, $a \in \mathcal{I}_\mathsf{A}$, is replaced by the synthetic value $\tilde{z}_{a1} := z_{b1}$, $b \in \mathcal{I}_\mathsf{B}$, where b is the randomly chosen observation unit from the index set $\mathcal{I}_\mathsf{B}$ of data file B and, hence, $\tilde{z}_{a1} \in \{z_{b1} : b \in \mathcal{I}_\mathsf{B}\}$. The a-th observation of complete, synthetic data file A is composed of $(x_{a1}, x_{a2}, y_{a1}, \tilde{z}_{a1})$, where the tilde marks the imputed and thus synthetic value.

However, simple random sampling gives all observation units in the donor file the same probability of being selected. Thus, it implicitly induces the independence of both the common and specific variables.

A more promising procedure is the assignment of donor and recipient records within groups of similar (homogeneous) records that are created by exploiting the information of the common variables. The realisations of selected categorical common variables are used to generate groups of similar records in both the recipient file and the donor file. Little and Rubin (2002) call these groups *adjustment cells*. Following D'Orazio et al. (2006b), we will call them *donation classes*. The choice of the common variables that are actually used to perform statistical matching (the so-called *matching variables*) has a high impact on the resulting matching quality. It is desirable that the common variables are highly correlated with, or good predictors for the specific variables (Rässler 2002, 10). See, for instance, D'Orazio et al. (2017) on how to choose the *matching variables*.

Consider again data file A as the recipient. The first step of hot deck imputation within homogenous groups is the assignment of all observations in A ⋃ B to donation classes. For this purpose, we partition the index set $\mathcal{I}$ into $D \leq |\mathcal{X}|$ index sets $\mathcal{I}^d$, $d = 1, \ldots, D$, such that for any d, all observation units in $\mathcal{I}^d$ have the same realisations of $\boldsymbol{X}$. Moreover, define $\mathcal{I}^d_\mathsf{A} := \mathcal{I}^d \cap \mathcal{I}_\mathsf{A}$ and $\mathcal{I}^d_\mathsf{B} := \mathcal{I}^d \cap \mathcal{I}_\mathsf{B}$. Every missing entry for the specific variable Z_1 of an observation unit from A in the d-th donation class, that is, $z_{a1}, a \in \mathcal{I}^d_\mathsf{A}$, is replaced by $\tilde{z}_{a1} := z_{b1}$, $b \in \mathcal{I}^d_\mathsf{B}$, which is the corresponding value of a randomly chosen observation from the donation class $\mathcal{I}^d_\mathsf{B}$, and hence $\tilde{z}_{a1} \in \left\{z_{b1} : b \in \mathcal{I}^d_\mathsf{B}\right\}$ for all $a \in \mathcal{I}^d_\mathsf{A}$.

Using donation classes, the imputation of Z is conditional on $\boldsymbol{X}$, thus reproducing the empirical conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ in A. Since there are no joint observations of all variables, additionally conditioning on $\boldsymbol{Y}$ is not possible. Thus, a conditional independence – between the imputed values of $\boldsymbol{Z}$ and the values of $\boldsymbol{Y}$, given $\boldsymbol{X}$ – is implicitly (empirically) established in the synthetic parts of the resulting complete file (see Rässler 2002, 200–204).

Every complete synthetic data file that consists of observations $(\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\boldsymbol{z}}_a)_{a \in \mathcal{I}_\mathsf{A}}$ and $(\boldsymbol{x}_b, \tilde{\boldsymbol{y}}_b, \boldsymbol{z}_b)_{b \in \mathcal{I}_\mathsf{B}}$ straightforwardly delivers estimates of the underlying joint distribution by evaluating the observed relative frequencies. Written in a form preparing for the generalisation developed in Subsection 4.3, we obtain for an event $\mathcal{E} = \mathcal{E}_\mathcal{X} \times \mathcal{E}_\mathcal{Y} \times \mathcal{E}_\mathcal{Z}$ with $\mathcal{E}_\mathcal{X} \subseteq \mathcal{X}$, $\mathcal{E}_\mathcal{Y} \subseteq \mathcal{Y}$ and $\mathcal{E}_\mathcal{Z} \subseteq \mathcal{Z}$,

$$\begin{aligned}\hat{P}(\mathcal{E}) := \hat{P}((\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \in \mathcal{E}) &= \frac{1}{n}\left|\{a \in \mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a,\boldsymbol{y}_a,\tilde{\boldsymbol{z}}_a) \in \mathcal{E}\} \cup \{b \in \mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b,\tilde{\boldsymbol{y}}_b,\boldsymbol{z}_b) \in \mathcal{E}\}\right| \\ &= \frac{1}{n}\left|\{a \in \mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a \in \mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a \in \mathcal{E}_y, \tilde{\boldsymbol{z}}_a \in \mathcal{E}_{\mathcal{Z}}\}\right| \\ &\quad + \frac{1}{n}\left|\{b \in \mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b \in \mathcal{E}_{\mathcal{X}}, \tilde{\boldsymbol{y}}_b \in \mathcal{E}_y, \boldsymbol{z}_b \in \mathcal{E}_{\mathcal{Z}}\}\right|.\end{aligned} \tag{1}$$

Any event which is not directly representable as a Cartesian product can be decomposed into the union of disjoint events of the previous form.

In the context of missing data, it is a well-known problem that single imputations are not able to reflect the uncertainty that arises from the missingness of joint information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Therefore, it is commonly recommended to apply *multiple imputation* techniques (e.g., Little and Rubin 2002, chap. 5.4), where the replacement of missing entries is performed several times. The obtained complete data files are then analysed by common methods for complete data and the results are subsequently pooled to achieve point estimates. Such multiple imputation techniques have been further developed by Rässler (2002, chap. 4) for application in statistical matching with the intention to estimate lower and upper bounds for the parameters of interest in the spirit of Manski (1995). However, Rässler (2002) only considers normally distributed data and, as stated in Ahfock et al. (2016, 82), by applying multiple imputation "there is no guarantee that the range of imputed datasets fully captures the uncertainty over the partially identified parameters".

3. Imprecise Imputation

3.1. Basic Idea and Terminology

Based on these considerations, we will now develop the concept of imprecise imputation, where we suggest imputing a *set* of plausible values for a missing entry. This leads to precise observations $(\boldsymbol{x}_a, \boldsymbol{y}_a)_{a\in\mathcal{I}_{\mathsf{A}}}$ in A and $(\boldsymbol{x}_b, \boldsymbol{z}_b)_{a\in\mathcal{I}_{\mathsf{B}}}$ in B, and to *imprecise*, that is, set-valued, synthetic observations $(\tilde{\boldsymbol{\jmath}}_a)_{a\in\mathcal{I}_{\mathsf{A}}}$ in A and $(\tilde{\boldsymbol{\eta}}_b)_{b\in\mathcal{I}_{\mathsf{B}}}$ in B. Please note that our aim is *not* to identify a single element of these imprecise observations for the purpose of precise single imputation, but rather to regard the whole set as the final piece of indivisible information. In Subsection 4.3 we show how the set-valued imprecise observations can be directly used to obtain estimates for the probability components of the joint distribution.

The following subsections detail and illustrate imprecise imputation. Three different ways of determining the sets of plausible values to be imputed are introduced, each taking into account the variations in how strong and trustworthy the underlying relationship between the common and specific variables is. Without loss of generality, again let A be the recipient and B the donor file, and let the donor classes be defined as in Subsection 2.2.

- **D** *Domain imputation* replaces every missing entry z_{al}, $a \in \mathcal{I}_{\mathsf{A}}$, of a variable Z_l, $l = 1, \ldots, r$, with its domain, that is,

$$\tilde{\jmath}_{al} := \mathcal{Z}_l, \quad \forall a \in \mathcal{I}_{\mathsf{A}}, \quad l = 1, \ldots, r. \tag{2}$$

- **VW** *Variable-wise imputation* on the basis of donation classes replaces every missing entry $z_{al}, a \in \mathcal{I}^d_{\mathsf{A}}$, of a variable Z_l, $l = 1, \ldots, r$, with the set of live values of

Z_l within the corresponding class $\mathcal{I}^d_{\mathsf{B}}$. Thus,

$$\tilde{\mathfrak{z}}_{al} := \left\{z_{bl} : b \in \mathcal{I}^d_{\mathsf{B}}\right\}, \quad \forall_a \in \mathcal{I}^d_{\mathsf{A}}, \quad d = 1, \ldots, D, \quad l = 1, \ldots, r. \tag{3}$$

- **CW** *Case-wise imputation*, that is, the simultaneous imputation of all missing entries of an observation a in $\mathcal{I}^d_{\mathsf{A}}$, where every tuple $\boldsymbol{z}_a = (z_{a1}, \ldots, z_{ar})$, $a \in \mathcal{I}^d_{\mathsf{A}}$ is replaced with the set of live tuples in the corresponding class $\mathcal{I}^d_{\mathsf{B}}$. Consequently,

$$\tilde{\mathfrak{z}}_a := \left\{(z_{bl}, \ldots, z_{br}) : b \in \mathcal{I}^d_{\mathsf{B}}\right\}, \quad \forall a \in \mathcal{I}^d_{\mathsf{A}}, \quad d = 1, \ldots, D. \tag{4}$$

3.2. Illustration and Discussion of the Different Types of Imprecise Imputation

3.2.1. Domain Imputation

The most conservative way to determine the set of plausible values that are candidate values for the substitution of a missing entry is to use the whole domain of the corresponding variable. Concretely, this means that every missing entry z_{al}, $a \in \mathcal{I}_{\mathsf{A}}$, $l = 1, \ldots, r$ is substituted by the set of all possible realisations of Z_l, that is, its domain $\mathcal{Z}_l$. Hence, $\tilde{\mathfrak{z}}_{al} := \mathcal{Z}_l$, $\forall a \in \mathcal{I}_{\mathsf{A}}$ becomes a set-valued entry in data file A, where all elements of the set are treated as equally plausible, but without a further reduction in the complexity by some (arbitrary) weighting or aggregation of the elements. The imputed sets for one variable are equal for all observations. This procedure is briefly illustrated in the following running toy example.

Minimal Example 1 *Consider two data files,* A *and* B*, which consist of* $n_{\mathsf{A}} = 2$ *observations of* (Y_1, Y_2, X_1, X_2) *and* $n_{\mathsf{B}} = 3$ *observations of* (X_1, X_2, Z_1, Z_2)*, respectively. The corresponding domains of the variables are* $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Z}_1 = \{0, 1\}$ *and* $\mathcal{Y}_2 = \mathcal{Z}_2 = \{0, 1, 2\}$*. Domain imputation results in the following completed data file.*

Table 1. Minimal example 1.

Y_1	Y_2	X_1	X_2	Z_1	Z_2
1	**2**	**1**	**0**	{0; 1}	{0; 1; 2}
0	**2**	**0**	**0**	{0; 1}	{0; 1; 2}
{0; 1}	{0; 1; 2}	**1**	**0**	**0**	**0**
{0; 1}	{0; 1; 2}	**1**	**0**	**1**	**1**
{0; 1}	{0; 1; 2}	**0**	**0**	**1**	**2**

Numbers in bold represent the original data. The files A and B are visually divided by the dashed line. The numbers in curly brackets depict the sets of possible realisations of the corresponding variables, that is, the domains, which are here the replacements for the previously missing entries.

This imputation procedure resembles the approach of Ramoni and Sebastiani (2001), who use an incomplete sample to estimate bounds for the parameters of conditional probability distributions in the context of Bayesian networks.

Applying domain imputation, it is guaranteed that the true (but missing) value is always an element of the imputed set. As previously mentioned, domain imputation is very

conservative, and thus it can also be applied if the common variables are not good predictors for the specific variables. However, it neglects any available dependence structure between the common and specific variables in the available data. In the following, we will introduce two other methods to determine the set of values for imputation that both take these dependencies into account, albeit to a different extent.

3.2.2. Variable-Wise Imputation

If $q \geq 2$ or $r \geq 2$, with due regard to the association between the common and specific variables, imputation can be performed on two different levels, either by treating each of the specific variables separately or by treating the specific variables within each of the two blocks simultaneously (see, e.g., Joenssen 2015, chap. 3, for precise imputation). In this section, we describe imprecise imputation on the separate level, while the simultaneous level will be addressed in the next section.

The imputation of live values only within donation classes ensures that associations between the common and specific variables are incorporated. As a consequence, the preservation of the dependence structure is improved and the estimated bounds for the parameters of interest become more narrow.

Without loss of generality, again let A be the recipient file and B the donor file. All observations $i \in \mathcal{I}_{\mathsf{A}} \cup \mathcal{I}_{\mathsf{B}}$ are allocated into donation classes depending on their realisations of the matching variables selected from the common variables $\boldsymbol{X}$, following the notation as introduced in Subsection 2.2. For every observation $a \in \mathcal{I}_{\mathsf{A}}^{d}$, the missing entry z_{al} of the variable Z_l, $l = 1, \ldots, r$ is substituted by the set of all live values of this variable from the same donation class in the donor file B, resulting in Equation (3).

Minimal Example 2 *Consider the same data situation as in Example 1. Now we will illustrate the application of the just-described variable-wise imputation. The different backgrounds display the different donation classes based on the combinations of the realisations of X_1 and X_2. Both common variables are used as matching variables in this example.*

Table 2. Minimal example 2.

Y_1	Y_2	X_1	X_2	Z_1	Z_2
1	**2**	**1**	**0**	{0; 1}	{0; 1}
0	**2**	**0**	**0**	{1}	{2}
{1}	{2}	**1**	**0**	**0**	**0**
{1}	{2}	**1**	**0**	**1**	**1**
{0}	{2}	**0**	**0**	**1**	**2**

This procedure preserves the dependencies between the common and the specific variables; however, the successive imputation of single variables breaks the dependence structure among the specific variables. Little and Rubin (2002, 72), for instance, have already stated that imputation should be multivariate to preserve the dependencies between the variables. If one attaches high value to this requirement, the imputation should be performed simultaneously for all variables in the data file as described in the following section. Nevertheless, variable-wise imputation is a good compromise between

the very conservative domain imputation and the more data-driven case-wise imputation procedure detailed in the following section.

3.2.3. Case-Wise Imputation

For case-wise imputation, we interpret the missing entries of one observation $a \in \mathcal{I}_{\mathsf{A}}^{d}$ out of the d-th donation class in the recipient file as tuple of the form $(z_{a1}, \ldots, z_{ar})$. This tuple of missing entries is replaced by the set of tuples $\tilde{\mathfrak{z}}_a$, which have been observed in the donor file B and the same donation class d, as in Equation (4). This strategy ensures that also the dependencies among the specific variables $\boldsymbol{Z}$ remain unchanged. The following example illustrates this imputation procedure.

Minimal Example 3 *Consider again the situation of Example 1 as a starting point. Interpret the empty cells z_{a1} and z_{a2} as tuples (z_{a1}, z_{a2}), $a = 1, 2$, and analogously y_{b1} and y_{b2} as tuples (y_{b1}, y_{b2}), $b = 3, 4, 5$. The result of case-wise imputation in this example is displayed in the following.*

Table 3. Minimal example 3.

(Y_1, Y_2)	X_1	X_2	(Z_1, Z_2)
(1, 2)	**1**	**0**	{(0, 0); (1, 1)}
(0, 2)	**0**	**0**	{(1, 2)}
{(1, 2)}	**1**	**0**	**(0, 0)**
{(1, 2)}	**1**	**0**	**(1, 1)**
{(0, 2)}	**0**	**0**	**(1, 2)**

3.2.4. General Remarks

A potential issue arises if at least one donation class in the donor file is empty. If so, variable-wise and case-wise imputation cannot directly be applied and we then recommend imputing the domains $\mathcal{Z}_1, \ldots, \mathcal{Z}_r$ or the Cartesian product of the domains $\mathcal{Z}$.

The partially set-valued data files produced by imprecise imputation can be interpreted as a set of underlying precise data files. On closer inspection, the sets produced by the three imputation procedures are nested: the largest set of underlying precise data files is obtained by domain imputation, while case-wise imputation yields the smallest set. Equation (15) shows this relationship formally.

Fractional hot deck imputation (e.g., Kim and Fuller 2004), which is also an imputation approach that is based on set-valued imputations, produces precise results that are contained in the sets obtained by imprecise imputation. It uses a weighting scheme, which is transferred onto the set of values to impute. This strategy reduces complexity by circumventing the direct handling of the imputed set-valued observation by creating a single completed data file with accordingly down-weighted precise pseudo-observations. This kind of precise data allows the direct use of common statistical models and methods. The variability, introduced by having multiple values to be imputed, is accounted for, in the situation of the fractional hot deck imputation, in the variance estimation of the precise estimator. However, variance estimation in the context of fractional hot deck imputation may be argued to be more complex yet more reliable in comparison to multiple imputation (e.g., Yang and Kim 2016).

During the imprecise imputation process, variable-wise and, in particular, domain imputation may create combinations of variable realisations which are contextually unjustified. For instance. D'Orazio et al. (2006b) distinguishes between two types of *logical constraints* to exclude impossible or unlikely combinations in the synthetic categorical data:

(i) *existence of some quantities* on the basis of the individual observation unit, and
(ii) *inequality constraints* on the level of the estimated probability distributions.

Especially the first case can easily be incorporated into the imputation step. Single, implausible values or tuples of values containing the unjustified combinations can easily be removed from the synthetic file. As an extension to both types of constraints, the set of values to be imputed can be restricted further removing not only contextually impossible values but also combinations of values that showed to be very rare within the data file or the population, motivated by the approach of Cattaneo (2013), developed in a decision-theoretic context. This means that the set of (variable-wise or case-wise) live values is restricted to the set of all values whose relative frequencies exceed a certain threshold δ, which may be dependent on the donation class. Increasing δ would gradually eclipse our conservative perspective, resulting, in the extreme case, in a precise single-valued imputation.

We propose to build upon the set-valued data directly, without reducing their complexity via a weighting scheme. In contrast to widely adopted imputation procedures yielding single-valued data, we are now in the situation of statistical analysis of partially set-valued data. To frame imprecise imputation formally, it will be embedded into the concept of finite disjunctive random sets, which allows the estimation of tight lower and upper bounds for the parameters.

In order to allow for a concise description in the following sections, we will take the observation-wise perspective on the imputed sets (i.e., the notation in terms of tuples), which corresponds to the perspective taken by the case-wise imputation. The imputation results of the other procedures can be transferred by taking the Cartesian product, e.g., $\tilde{\mathfrak{d}}_a = \tilde{\mathfrak{d}}_{a1} \times \ldots \times \tilde{\mathfrak{d}}_{ar}$.

4. Imprecision Imputation and Finite Disjunctive Random Sets

Imprecise imputation provides us with partially set-valued data. To prepare a well-founded statistical analysis, we have to formalise imprecise imputation probabilistically. For this purpose, the direct formalisation of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ as collections of random variables and corresponding realisations is no longer sufficient. Starting from an applied point of view, two types of generalisations, which will indeed prove compatible among each other, could be imagined. Firstly, we could abstractly look for a concept of set-valued variables with corresponding set-valued realisations. Secondly, we could assume that every set represents outcomes of various random variables, one of which is the true underlying, yet not precisely observable, random variable. (Throughout this article, we use the term *random variable* to refer to a mapping to the real numbers as well as to some nonnumerical finite space. In the context of the latter, the term *random element* is sometimes used for the sake of distinction e.g., Nguyen 2006).

In this section it will be shown how set-valued observations, and thus the resulting data files of the three imprecise imputation procedures in particular, are covered by the concept of *disjunctive random sets*, also known as *ill-perceived random variables* (Couso et al. 2014; Nguyen 2006). This embedding allows for the assessment of probability statements and the construction of corresponding estimates from the partially set-valued synthetic file derived from imprecise imputation. The interpretation of the set-valued quantities as disjunctive random sets corresponds to the view of Dempster (1967), on which the Dempster-Shafer theory of belief functions (Shafer 1976) is built, which has become very popular in artificial intelligence (see, for example Denoeux 2016).

4.1. Random Set Formulation of Imprecise Imputation

The true random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ map from the underlying population space, denoted by Ω in the sequel, into the domains $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, yielding realisations $\boldsymbol{x}_i$, $\boldsymbol{y}_i$, $\boldsymbol{z}_i$ with $i \in \mathcal{I}$, respectively. Now, neither $\boldsymbol{y}_b$ nor $\boldsymbol{z}_a$ are available, but are replaced by synthetic observations $\tilde{\mathfrak{y}}_b$ and $\tilde{\mathfrak{z}}_a$, respectively, according to either Equation (2), (3), or (4), depending on the chosen imprecise imputation procedure. To formalise this situation, we follow the common practice in statistical matching, treating $\mathcal{I}_\mathsf{A}$ and $\mathcal{I}_\mathsf{B}$ as fixed. This allows us to globally replace $\boldsymbol{Y}$ and $\boldsymbol{Z}$ by the set-valued variables $\mathfrak{Y}$ and $\mathfrak{Z}$ (with realisations $\mathfrak{y}_i$ and $\mathfrak{z}_i$, $i \in \mathcal{I}$). The imputed values are already sets, so they fit in nicely, but in order to deal with the already observed realisations, we regard them now as singletons containing only the observed value, for example $\mathfrak{z}_{bl} = \{z_{bl}\}$, $\forall b \in \mathcal{I}_\mathsf{B}, l = 1, \ldots, r$. The variables $\mathfrak{Y}$ and $\mathfrak{Z}$ map into the corresponding power sets $2^{\mathcal{Y}}$ and $2^{\mathcal{Z}}$, whereby mapping into the empty set is excluded.

If we collect the random variables of interest in a variable $\boldsymbol{\Gamma}$ and define $\mathcal{W} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then

$$\boldsymbol{\Gamma} := (\boldsymbol{X}, \mathfrak{Y}, \mathfrak{Z}) : \Omega \to 2^{\mathcal{W}} \setminus \{\emptyset\} \tag{5}$$

is a finite nonempty random set (see Definition 3.1 in Nguyen 2006, 35), satisfying the required measurability condition by equipping $2^{\mathcal{W}}\setminus\{\emptyset\}$ with its power set. Since in our setting the imputed (synthetic) set-valued entries of the specific variables are understood as the collection of possible underlying true values, this random set has to be interpreted in the disjunctive way (see, for example Couso et al. 2014; Couso and Dubois 2014).

In general, any disjunctive random set $\boldsymbol{\Gamma}$ induces an upper inverse $\boldsymbol{\Gamma}^*$ and a lower inverse $\boldsymbol{\Gamma}_*$. When considering an event of interest $\mathcal{E} \subseteq \mathcal{W}$, which is now a singleton in the considered space $2^{\mathcal{W}}$, the upper inverse contains all the elements of the population whose image overlaps with $\mathcal{E}$, while the lower inverse contains only those elements of the population whose (nonempty) image is entirely contained within $\mathcal{E}$:

$$\boldsymbol{\Gamma}^*(\mathcal{E}) := \{\omega \in \Omega : \boldsymbol{\Gamma}(\omega) \cap \mathcal{E} \neq \emptyset\} \tag{6}$$

and

$$\boldsymbol{\Gamma}_*(\mathcal{E}) := \{\omega \in \Omega : \boldsymbol{\Gamma}(\omega) \subseteq \mathcal{E}\}. \tag{7}$$

In a heuristic formulation, the upper inverse considers all aspects that do not entirely contradict $\mathcal{E}$, while the lower inverse collects all aspects that necessarily imply $\mathcal{E}$. By using

the probability measure $\mathbb{P}$ defined on the original probability space involving Ω, the upper and lower probabilities are then defined in terms of the upper and lower inverse, respectively:

$$P^*(\mathcal{E}) = \mathbb{P}(\mathbf{\Gamma}^*(\mathcal{E})) \quad \text{and} \quad P_*(\mathcal{E}) = \mathbb{P}(\mathbf{\Gamma}_*(\mathcal{E})) \quad \forall \mathcal{E} \subseteq \mathcal{W}. \tag{8}$$

In order to improve readability we have not marked the image probability measure induced by the random set $\mathbf{\Gamma}$, i.e., $P_{\mathbf{\Gamma}} = P$, and we proceed analogously with the corresponding set functions P^* and P_*. If we refer to a different image measure, the random quantity inducing this image measure, will be set as subscript to P. If we look at an underlying, ill-perceived random variable $\boldsymbol{W}_0 : \Omega \rightarrow \mathcal{W}$, only knowing that the unobserved true value $\boldsymbol{W}_0(\omega)$ lies (with probability one) within the observed set $\mathbf{\Gamma}(\omega)$, it can be shown (see, for example Couso et al. 2014) that for every event $\mathcal{E} \subseteq \mathcal{W}$ the upper and lower probabilities induced by the random set enclose the probability of $\boldsymbol{W}_0$:

$$P_*(\mathcal{E}) \leq P_{\boldsymbol{W}_0}(\mathcal{E}) \leq P^*(\mathcal{E}) \quad \forall \mathcal{E} \subseteq \mathcal{W}.$$

This leads to another way of interpreting a random set, namely as producing a family of compatible, precise probability measures $\mathcal{P}(\mathbf{\Gamma})$, which is a subset of the set $\mathcal{P}$ of all probability measures on $(2^{\mathcal{W}}, 2^{2^{\mathcal{W}}})$. Nguyen (1978) showed that if $\mathcal{W}$ is finite, the probability distribution induced by $\mathbf{\Gamma}$ corresponds to the basic probability assignment in Dempster-Shafer theory and thus makes the belief function mathematically equivalent to P_*. Consequently, the technical results from that area may be used as well.

In the present special case of finite $\mathcal{W}$, the set $\mathcal{P}(\mathbf{\Gamma})$ coincides with the credal set $\mathcal{M}(P^*)$, that is, those precise probability measures that respect the upper and lower bounds defined by P^* and P_* event-wise (see Miranda et al. 2010), which also embeds the situation considered here into the framework of imprecise probabilities (e.g., Walley 1991; Augustin et al. 2014).

In particular, P_* and P^* are lower and upper probabilities that are envelopes of all probability measures P in $\mathcal{M}(P^*)$:

$$P_*(\mathcal{E}) = \inf_{P \in \mathcal{M}(P^*)} P(\mathcal{E}) \quad \text{and} \quad P^*(\mathcal{E}) = \sup_{P \in \mathcal{M}(P^*)} P(\mathcal{E}).$$

Indeed, P^*, P_* and $\mathcal{M}(P^*)$ are three mathematically equivalent formulations that can be transferred into each other. Therefore, from an applied point of view, each of them can be seen as the core result of a probabilistic description of imprecise imputation. For any possibly true probability distribution $P_{\boldsymbol{W}_0}$, our embedding into random sets provides us with a set $\mathcal{M}(P^*)$ of distributions induced by $P_{\boldsymbol{W}_0}$ such that $\mathcal{M}(P^*)$ contains $P_{\boldsymbol{W}_0}$. By construction, this is the smallest set that is deducible from the concrete imputation procedure without adding further assumptions or knowledge. Dually, $P^*(\mathcal{E})$ and $P_*(\mathcal{E})$ are the narrowest bounds, deducible on the probabilities of an $\mathcal{E}$.

4.2. *Conditioning Disjunctive Random Sets*

The representation via the set $\mathcal{M}(P^*)$ of compatible probability distributions including the embedding into the framework of imprecise probabilities guides the further probabilistic analysis of the partially set-valued data file achieved by imprecise imputation. For

instance, if the elements of $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ eventually get associated with real-valued outcomes, then a generalised expectation is logically defined via the infimum and supremum of all compatible traditional expectations based on image measures of elements of $\mathcal{M}(P^*)$.

A similar procedure suggests itself for conditioning, namely an element-wise application of conditioning for all $P \in \mathcal{M}(P^*)$, provided $P(\mathcal{C}) > 0$ for a conditioning event $\mathcal{C}$ (see, for example Dubois and Prade (1992) or Fagin and Halpern (1991) for a discussion and a comparison to an alternative). It can be shown (e.g., De Campos et al. (1990), Couso et al. (2014), and Fagin and Halpern (1991)) that this leads to the following closed-form results for the upper conditional probability

$$P^*(\mathcal{S}|\mathcal{C}) = \sup_{P\in\mathcal{M}(P^*)} P(\mathcal{S}|\mathcal{C}) = \frac{P^*(\mathcal{S}\cap\mathcal{C})}{P^*(\mathcal{S}\cap\mathcal{C}) + P_*(\bar{\mathcal{S}}\cap\mathcal{C})} \tag{9}$$

and the lower conditional probability

$$P_*(\mathcal{S}|\mathcal{C}) = \inf_{P\in\mathcal{M}(P^*)} P(\mathcal{S}|\mathcal{C}) = \frac{P_*(\mathcal{S}\cap\mathcal{C})}{P_*(\mathcal{S}\cap\mathcal{C}) + P^*(\bar{\mathcal{S}}\cap\mathcal{C})}, \tag{10}$$

where $\bar{\mathcal{S}}$ denotes the complement of $\mathcal{S}$.

4.3. Parameter Estimation by Means of Disjunctive Random Sets Based on Imprecise Imputation

So far, this approach has been described in a probabilistic setting, where every entity involved is known (besides the true hidden/ill-perceived random variable). In the following, the statistical perspective will be taken in which the probabilities that correspond to the random set need to be estimated from a finite sample. Consequently, we take our synthetic data file derived from imprecise imputation as consisting of $n = n_{\mathsf{A}} + n_{\mathsf{B}}$ realisations γ_i, $i \in \mathcal{I}$, of the corresponding generic random set $\mathbf{\Gamma}$ from Equation (5). Referring to Equation (8), with Equations (6) and (7), we obtain, in a generalisation of Equation (1), for our event $\mathcal{E} = \mathcal{E}_{\mathcal{X}}\times\mathcal{E}_{\mathcal{Y}}\times\mathcal{E}_{\mathcal{Z}}$:

$$\begin{aligned}
\widehat{P^*}(\mathcal{E}) &= \frac{1}{n}\left|\{i\in\mathcal{I} : \gamma_i \cap \mathcal{E} \neq \emptyset\}\right| \\
&= \frac{1}{n}\left(\left|\{a\in\mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\mathfrak{z}}_a)\cap\mathcal{E}\neq\emptyset\}\right| + \left|\{b\in\mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b, \tilde{\mathfrak{y}}_b, \boldsymbol{z}_b)\cap\mathcal{E}\neq\emptyset\}\right|\right) \\
&= \frac{1}{n}\left|\{a\in\mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a\in\mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a\in\mathcal{E}_{\mathcal{Y}}, \tilde{\mathfrak{z}}_a\cap\mathcal{E}_{\mathcal{Z}}\neq\emptyset\}\right| \\
&\quad + \frac{1}{n}\left|\{b\in\mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b\in\mathcal{E}_{\mathcal{X}}, \tilde{\mathfrak{y}}_b\cap\mathcal{E}_{\mathcal{Y}}\neq\emptyset, \boldsymbol{z}_b\in\mathcal{E}_{\mathcal{Z}}\}\right|
\end{aligned} \tag{11}$$

and

$$\begin{aligned}\widehat{P_*}(\mathcal{E}) &= \frac{1}{n}|\{i \in \mathcal{I} : \gamma_i \subseteq \mathcal{E}, \gamma_i \neq \emptyset\}| \\ &= \frac{1}{n}\left(|\{a \in \mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\mathfrak{z}}_a) \subseteq \mathcal{E}\}| + |\{b \in \mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b, \tilde{\mathfrak{y}}_b, z_b) \subseteq \mathcal{E}\}|\right) \\ &= \frac{1}{n}\left|\{a \in \mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a \in \mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a \in \mathcal{E}_{\mathcal{Y}}, \tilde{\mathfrak{z}}_a \subseteq \mathcal{E}_{\mathcal{Z}}\}\right| \\ &\quad + \frac{1}{n}\left|\{b \in \mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b \in \mathcal{E}_{\mathcal{X}}, \tilde{\mathfrak{y}}_b \subseteq \mathcal{E}_{\mathcal{Y}}, z_b \in \mathcal{E}_{\mathcal{Z}}\}\right|.\end{aligned} \tag{12}$$

From $\widehat{P^*}(\mathcal{E})$ and $\widehat{P_*}(\mathcal{E})$ also an estimate of the induced underlying set of probability measures can be derived:

$$\widehat{\mathcal{M}}(P^*) = \{P \in \mathcal{P} : \widehat{P_*}(\mathcal{E}) \leq P(\mathcal{E}) \leq \widehat{P^*}(\mathcal{E}), \quad \forall \mathcal{E} \subseteq \mathcal{W}\}. \tag{13}$$

In comparing the estimates resulting from the different types of imputation procedures, it is essential to recall that the different set-valued data files are nested, by construction, with respect to all compatible underlying precise data files. The set resulting from domain imputation is a (nonstrict) superset of the set obtained from variable-wise imprecise imputation, which contains the set produced by case-wise imprecise imputation. Therefore, with the abbreviations introduced in Subsection 3.1, it holds that

$$\widehat{\mathcal{M}}\left(P^{*^{CW}}\right) \subseteq \widehat{\mathcal{M}}\left(P^{*^{VW}}\right) \subseteq \widehat{\mathcal{M}}\left(P^{*^{D}}\right) \tag{14}$$

and, for every event $\mathcal{E} \subseteq \mathcal{W}$,

$$\widehat{P_*}^{D}(\mathcal{E}) \leq \widehat{P_*}^{VW}(\mathcal{E}) \leq \widehat{P_*}^{CW}(\mathcal{E}) \leq \widehat{P^*}^{CW}(\mathcal{E}) \leq \widehat{P^*}^{VW}(\mathcal{E}) \leq \widehat{P^*}^{D}(\mathcal{E}). \tag{15}$$

This allows us to compare the results obtained through the different imputation approaches to the result under conditional independence, which yields a single precise probability distribution. It can be argued that the probability distribution under conditional independence is contained in any of the estimated sets. Furthermore, as can be seen from the relations between the different sets of probabilities in Equation (14), the set induced by case-wise imputation can be regarded as containing probability distributions neighbouring the one under conditional independence. The other sets can be interpreted to deviate even more from conditional independence, where domain imputation has the largest deviation. Domain imputation demonstrably neglects any conditional dependence structure in the construction of its bounds. Therefore, the bounds are maximal, but not vacuous, thus constraining the parameter space.

In addition to logical constraints on the imputation level (see Subsubsection 3.2.4), constraints on the level of the estimated probability distribution can be regarded as a refinement of the estimated set $\widehat{\mathcal{M}}(P^*)$ of probabilities derived from our imprecise imputation (see Equation (13)). Since by construction $\widehat{\mathcal{M}}(P^*)$ is representable as a convex polyhedron in $\mathbb{R}^{|\mathcal{W}|-1}$, especially linear constraints can be incorporated very conveniently.

Minimal Example 4 *For demonstration purposes, let us estimate the bounds of conditional probabilities* $P(Y_1 = 1|Z_1 = 1)$ *for the case-wise imputed data of our toy*

example from Example 3. For the upper conditional probability we need to estimate $P^*(Y_1 = 1, Z_1 = 1)$ *and* $P_*(Y_1 \neq 1, Z_1 = 1)$ *in accordance to Equation* (9). *We estimate the upper joint probability with Equation* (11) *by counting how many observations have or could have realisation with* $y_1 = 1$ *and* $z_1 = 1$. *This holds for observations 1 and 4:* $\widehat{P^*}(Y_1 = 1, Z_1 = 1) = \frac{1}{5} \cdot 2 = 0.4$. *The lower joint probability is obtained by Equation* (12) *by counting how many observations only have realisations with* $Y_1 \neq 1$ *and* $Z_1 = 1$. *This holds for observations 2 and 5, and hence* $\widehat{P_*}(Y_1 \neq 1, Z_1 = 1) = \frac{1}{5} \cdot 2 = 0.4$ *and thus the upper conditional probability is* $\widehat{P^*}(Y_1 = 1|Z_1 = 1) = \frac{0.4}{0.4+0.4} = 0.5$. *Similarly, the lower and upper joint probabilities are estimated, occurring in Equation* (10): $\widehat{P_*}(Y_1 = 1, Z_1 = 1) = 0.2$ and $\widehat{P^*}(Y_1 \neq 1, Z_1 = 1) = 0.4$, *resulting in the lower conditional probability* $\widehat{P_*}(Y_1 = 1|Z_1 = 1) = \frac{0.2}{0.4+0.2} = \frac{1}{3}$. *Thus,* $\hat{P}(Y_1 = 1|Z_1 = 1)$ *is within the interval* $\left[\frac{1}{3}; \frac{1}{2}\right]$.

5. Simulation Study of Imprecise Imputation

To investigate the quality of imprecise imputation, we have performed a simulation study. It would have been possible to also match real data, but in a real-data application the true underlying distribution is unknown and assessing the statistical matching quality is possible only by checking whether the marginal distributions are preserved. Since this is clearly not sufficient as a sole quality criterion, we have simulated data. With the aid of a simulation study we have also been able to cover various data scenarios which make the results of our investigation of the quality criteria more credible. Moreover, the noise arising from the sampling procedure in the context of real-data applications is neutralised.

We simulated a complete categorical data file $\mathsf{A} \cup \mathsf{B}$ with i.i.d. observations and split it into two separate files, A and B, with $n_\mathsf{A} = n_\mathsf{B}$. Subsequently, the observations of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are deleted from A and B, respectively, and the two files are statistically matched by imprecise imputation. To assess the statistical matching quality, we analysed, on the one hand, whether the true parameters of the marginal distributions and the joint distributions are within their respective estimated bounds, and, on the other hand, the distance between the upper and lower bounds. This distance, which we will call *interval width* in the following, is an appropriate performance measure since the true parameters would always lie within the estimated bounds if we chose the unit interval as a trivial estimator of a probability component. Thus, the narrower the interval that covers the component of the true parameter, the better the procedure performs. In the following, we will detail the simulation design, parameters, and results. All simulations and analyses are conducted in R (R Core Team 2018). The specific task presented in this paper is implemented in a published R-package *impimp* (Fink et al. 2019), which was also utilised in the simulation, but is in the same way usable for real-data applications.

5.1. Simulation Design

The starting point of our simulation analysis is two categorical data files, A and B. Both of them contain information on four common variables $\boldsymbol{X} = \{X_1, X_2, X_3, X_4\}$ and four specific variables $\boldsymbol{Y} = \{Y_1, Y_2, Y_3, Y_4\}$ or $\boldsymbol{Z} = \{Z_1, Z_2, Z_3, Z_4\}$, respectively, with domains $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{Z}_1 = \mathcal{Z}_2 = \{0, 1\}$ and $\mathcal{X}_3 = \mathcal{X}_4 = \mathcal{Y}_3 = \mathcal{Y}_4 = \mathcal{Z}_3 = \mathcal{Z}_4 = \{0, 1, 2\}$.

Altogether, we modify the following four simulation parameters:

1. The strength of the bivariate associations in terms of the corrected contingency coefficient C, also known as Sakoda's adjusted Pearson's C: $C \in [0, 0.2)$, $C \in [0.2, 0.6)$, or $C \in [0.6, 1)$;
2. The Jensen-Shannon divergence JSD (e.g., Lin 1991) from the marginal distribution of the common variables to the discrete uniform distribution: $JSD > 0.15$ or $JSD \leq 0.015$;
3. The numbers of observations $n_A = n_B \in \{50, 100, 250\}$; and
4. the dependence structure among the variables (see Figure 2).

Altogether, we obtain 72 simulation scenarios. An explanation of the choice of the simulation parameters follows in the next section. An exhaustive justification and description of the simulation design can be found in Appendix A and Appendix B, respectively.

5.2. *Simulation Parameters*

As already stated by Rässler (2002, 10), the common variables should be good predictors for the specific variables. This ensures that the donation classes are suitable for generating homogeneous groups of observations that lead to proper donor values for a missing entry. Taking this fact into account, we vary the dependence structure within a simulated data file in terms of its bivariate associations.

Figure 2 shows four different dependence structures that are covered by our simulation design. The upper six variables of each design represent the binary variables, and the six variables below the dashed line represent the variables with three categories. The connecting lines between the variables display the bivariate dependencies among these variables. For example, in the top line of Structure 1, the variable X_1 is connected to variable Y_1 and also to variable Z_1. The strengths of these bivariate associations are controlled by the corrected contingency coefficient $C \in [0, 1]$. This association measure for categorical variables is based on the χ^2-coefficient for contingency tables, but is corrected for the number of observations, as well as for the number of categories.

At first sight, the number of observations plays a counterintuitive role in this simulation study. We expect that the distances between the lower and upper bounds of the parameters of interest increase in situations with a higher number of observations. This is due to the

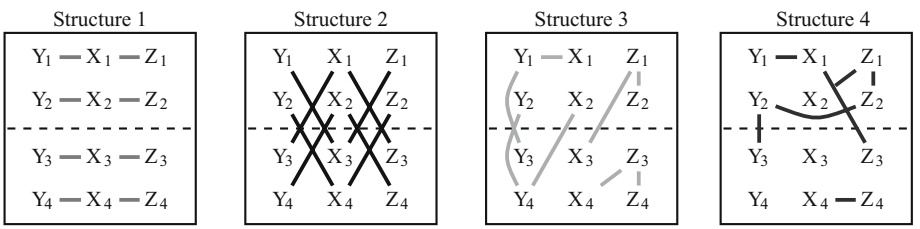

Fig. 2. Four different dependence structures among the variables in the simulation study. A line between two variables indicates dependence between them.

fact that a growth of the number of observations also causes an increase in the number of missing entries, which, in turn, leads to less precise estimations.

The Jensen-Shannon divergence from the marginal distributions of the common variables to the discrete uniform distribution is expected to have an indirect effect on the statistical matching quality. If one or more of these marginals are far away from the discrete uniform distribution, we obtain rare realisations of our matching variables, which induce rare donation classes. This circumstance may likely lead to situations where certain rare donation classes of the recipient file do not exist in the donor file. In these cases, we impute, in accordance with the recommendation in Subsubsection 3.2.4, the domain for the missing entries that corresponds to a minimum of information which, in turn, leads to bounds that are (slightly) further apart.

5.3. Simulation Results

As discussed, we use two measures of quality. Firstly, we investigate whether the true parameters of our simulation distributions lie within the corresponding lower and upper bounds estimated on the synthetic and partially set-valued data. Secondly, we report the mean interval widths that equal the mean distances between the upper and lower bounds. An interval width of 0 corresponds to a precise estimation.

Table 4 shows that the true values of the components of the marginal and the joint distributions almost always lie inside the estimated bounds. When considering the coverage of the marginal distributions (upper part of Table 4), the only visible difference is between the domain and donation-based approaches with respect to the coverage of the true probability: while the intervals for domain imputation are always wide enough to cover the true probability, for variable-wise and to the same extent for case-wise imputation the estimated intervals are sometimes too narrow. Regarding the joint distribution (lower part of Table 4), the intervals estimated on the domain-imputed data still always cover the true probability, but there is now also a slight difference between

Table 4. Relative number of probability table components for which the true parameter of the marginal distributions (top) / joint distributions (bottom) lies inside the estimated bounds, aggregated over all repetitions. The presented summary lists the result when pooling all simulation scenarios. The absence of decimal places for domain imputation highlights the numerically exact values.

Imputation procedure	Min.	1st quartile	Median	3rd quartile	Max.	Mean
Domain	1	1	1	1	1	1
Variable-wise	0.9250	0.9613	0.9867	0.9967	1.0000	0.9792
Case-wise	0.9250	0.9613	0.9867	0.9967	1.0000	0.9792

Imputation procedure	Min.	1st quartile	Median	3rd quartile	Max.	Mean
Domain	1	1	1	1	1	1
Variable-wise	0.9975	0.9989	0.9994	0.9996	0.9998	0.9992
Case-wise	0.9944	0.9985	0.9990	0.9993	0.9997	0.9987

case-wise and variable-wise imputation, showing the hierarchy of the intervals as given in Equation (15). Nonetheless, the estimated intervals of the donation-based imputation approaches still almost always cover the true probability. The difference between marginal and joint coverage is mostly due to the fact that, using the simulation design, the joint distribution had more components (46,656) than observations in the data file, which means that most of the underlying probability entries were zero. The marginal distributions, in contrast, consisted of only two to three entries, which made it harder to distinguish on the estimated level between the different imputation approaches. By and large, the results show a desirable output and also demonstrate the power of our method, which achieves high average coverage even across the diverse simulation scenarios.

The interval width was separately analysed for the components of the marginal distributions and joint distributions within the simulation. The aggregated results are displayed in the figures in Appendix C and summarised in the following.

The mean and maximal interval widths of the estimated intervals for the marginal distributions using domain imputation are always 0.5. This is the maximum interval width which can be achieved if we impute $\mathsf{A} \cup \mathsf{B}$ under the constraint that $n_{\mathsf{A}} = n_{\mathsf{B}}$. Both variable-wise imputation and case-wise imputation yield intervals that are, in most of the cases, smaller than the intervals obtained by domain imputation. This also holds for the components of the joint distributions.

The interval widths of the marginals are conspicuously affected by the divergence of the marginal distributions to the discrete uniform distribution. If the marginals are close to the uniform distribution, the intervals are narrow. However, this effect decreases if there are few direct connections between the specific variables and the common variables. For the interval widths of the components of the joint distribution, we can observe a slightly contrary effect regarding the combination of marginals that are close to the uniform distribution and few direct connections between the specific variables and common variables. For the simulation designs with a higher divergence to the uniform distribution, the variation of the interval widths is considerably smaller. Moreover, in these cases, the median of the interval widths lies below the median of the design, with a smaller divergence to the uniform distribution. At first sight, this result appears somewhat counterintuitive, but can be explained as follows. Given a fixed value for the corrected contingency coefficient C, with marginal distributions of the common variables which are far away from the discrete uniform distribution, we obtain a probability table which has fewer combinatorial possibilities for each cell than with marginals close to the uniform distribution. This circumstance makes the estimation more precise in some cases, which in turn leads to smaller interval widths.

Furthermore, the results show that with a growing number of observations, the interval widths of the marginal distributions slightly increase. The interval widths also show higher variations in these cases. The interval widths for the components of the joint distribution show the same behaviour with respect to the number of observations.

The strengths of the bivariate associations in terms of the corrected contingency table also affect the widths of the intervals concerning the marginal distributions. In particular, the first dependence structure shows that the interval width decreases with a higher C. Nevertheless, the difference between low and high associations is, in few cases, (especially for marginals close to the uniform distribution) opposite, or only visible in the

variations. Considering the interval widths for the components of the joint distribution, we can see that high associations improve the estimation.

The simulation results also show that, as expected, the dependence structure among the variables in a data file has an influence on the estimated lower and upper bounds of the parameters of the marginal distributions. The mean interval widths increase if the specific variables and the common variables have only few connections. The last dependence structure where there are only few connections between the common variables and the specific variables tends to lead to intervals with higher widths for the components of the joint distribution.

To sum up, all imputation procedures yield lower and upper bounds that almost always cover the components of the true parameter value. The number of cases where a component of the true parameter lies outside of the estimated interval is negligible. Additionally, the width of the intervals decreases the more the dependence structure among the variables in the data file are incorporated in the imputation procedure. This also holds for small associations and for structures where the specific variables only have few connections to the common variables.

6. Concluding Remarks

We have presented the first micro approach for statistical matching of categorical data that reflects the natural uncertainty of statistical matching. Our approach relies on imprecise imputation, that is, the idea to impute sets of plausible values. We suggested three types of imputation strategies: domain, variable-wise, and case-wise imprecise imputation. They can be distinguished by their ability to reproduce the available dependence structure between the common and the observed specific variables in the originals files A and B into the synthetic file. They also differ in the amount of data constellations produced beyond those obtained by single or multiple imputation under the conditional independence assumption. Imprecise imputation can be seen as a set-valued generalisation of multiple (hot deck) imputation on the one hand, and fractional hot deck imputation on the other hand.

The most conservative approach, domain imputation, does not take any dependencies in the original data into account. Essentially, the dependencies present in the original files are diluted in the resulting complete synthetic file. This approach is suitable especially when there is little dependence between the common and specific variables. On the other hand, imprecise imputation based on donation classes is able to utilise the observed dependencies between the common and specific variables, and even, in the example of the case-wise variant, within the specific variables.

Embedding imprecise imputation into the framework of finite random sets allows us to derive set-valued estimates of the underlying true parameters. These estimates – possibly after their refinement by external information, see, for example, Subsubsection 3.2.4 – reflect the uncertainty inherent in the identification problem of statistical matching. The estimation procedure utilises the set-valued information to full extent without artificially reducing the complexity of the imputed sets. Simulation results, based on a new simulation technique for dependent categorical data, corroborate that the true parameter values lie almost always inside the respective estimated bounds.

Imprecise imputation is an intuitive statistical matching micro approach which can easily be extended for more than two data files. In a strongly unbalanced statistical matching situation where, for example $n_{\mathsf{A}} \ll n_{\mathsf{B}}$, imprecise imputation can be applied straightforwardly to impute only the smaller file. If so, A takes the role of the recipient and the larger file, B, the role of the donor. In this special situation, the estimates for the specific variables $\boldsymbol{Y}$ are precise.

Moreover, the imprecise imputed data file with synthetic set-valued observations can be used as a starting point to derive one or multiple data files of the usual form. This would bring back the opportunity to use statistical procedures for the analysis of these now entirely single-valued data and to combine the results obtained from those data files by common multiple imputation techniques. However, one would then lose sight, to a considerable extent, of the conviction of this work, which is to produce a credible analysis by taking the full uncertainty into account.

Further studies need to be carried out to validate the performance of imprecise imputation. On the one hand, additional simulation parameters and dependence structures should be investigated in simulation studies. On the other hand, the performance of imprecise imputation should also be assessed by real-data applications. However, considerably more work will need to be done to find a definition of appropriate statistical matching quality criteria, since the true joint distribution is not available for comparisons. A further natural progression of this work is the comparison of imprecise imputation to existing statistical matching macro approaches that also address the identification problem. For this purpose, a comparison of the uncertainty measures introduced in Conti et al. (2012) or Conti et al. (2017) is desirable.

Finally, we should stress that imprecise imputation is not restricted to the block-wise missing pattern in the statistical matching framework: it is also applicable to general missing data problems. All three types of imprecise imputation promise considerable potential for a credible analysis of (non)randomly missing data far beyond statistical matching and are worthwhile to be elaborated upon and evaluated in detail.

7. Appendix

7.1. Appendix A. Why we need a new simulation procedure

To generate simulated categorical data that meet all the desired properties, we propose a new procedure which we detail in the following section. However, first we want to elucidate why conventional simulation approaches are not suitable for our requirements. The key aspects are listed as follows:

(i) One way to generate categorical data with predefined properties is to draw random observations from a multidimensional probability table, which, on the one hand, fulfils all of these properties that, on the other hand, represents the probability entries of the joint distribution of all variables. The main disadvantage of this procedure is that it can be very difficult to find a suitable joint distribution that fulfils all the desired properties. Furthermore, we would argue that it is necessary to consider several joint distributions in order to draw valid conclusions about the performance

of imprecise imputation, which in turn makes the problem of finding suitable distributions even harder.

(ii) Another option would be the simulation of categorical data based on a multidimensional (logit) regression model. However, a regression model cannot be used to control for the dependence structure and strength within the set of variables in the detail we wish to have.

(iii) The simulation of categorical data which imply a certain dependence structure can also be realised using a probabilistic graphical model such as a Bayesian network. The major problem with this way of proceeding is the resulting conditional independence among parts of our variables. If the – in real-world applications potentially unjustified – conditional independence assumption holds in our simulated data, statistical matching techniques directly utilising this assumption would unfairly outperform, making a fair comparison of procedures impossible.

(iv) A further feasible way to generate dependent categorical data is to employ a multivariate normal distribution with a predefined correlation matrix and discretise the data drawn from it. Nevertheless, the resulting simulated data have an ordinal scale instead of a nominal scale and we have no direct control on the strengths of the dependencies in terms of the corrected contingency coefficient. The same problems hold for simulation techniques that are based on a Gaussian copulas, such as the one suggested by Barbiero and Ferrari (2017).

To sum up, our goal is to use a simulation technique that takes all of our desired properties into account and avoid the problems described previously.

7.2. *Appendix B. Simulation procedure*

For this purpose, we invented a new simulation procedure that is directly based on two-way tables of relative frequencies and a suitable association measure. The bivariate associations within the simulated data can be expressed by this association measure on

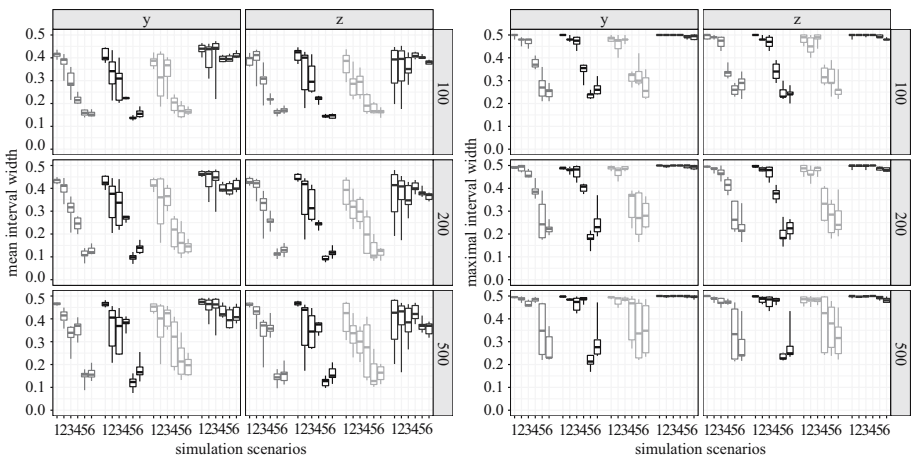

Fig. 3. Mean and maximal interval widths of the components of the marginal distributions of the specific variables for variable-wise imputation. The two columns display the pooled results for the marginals of the specific variables ***Y*** *and* ***Z****, respectively.*

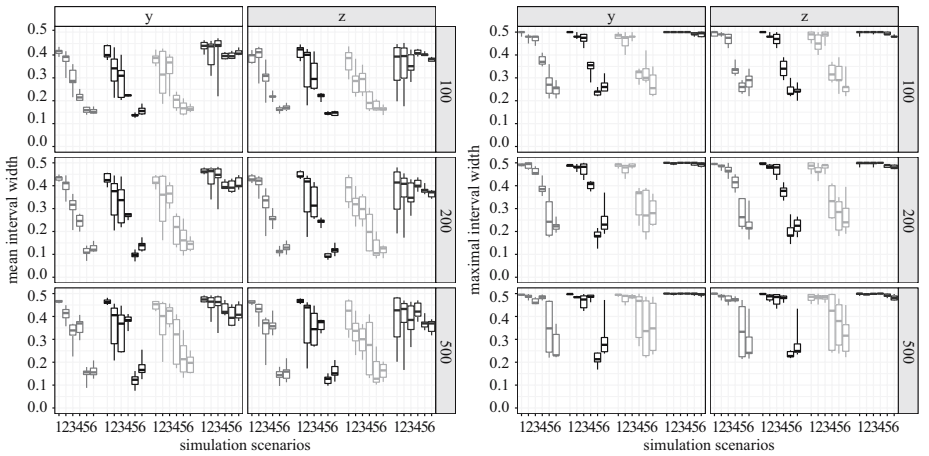

Fig. 4. Mean and maximal interval widths of the components of the marginal distributions of the specific variables for variable-wise imputation. The two columns display the pooled results for the marginals of the specific variables ***Y*** *and* ***Z****, respectively.*

bivariate frequency tables of sizes 2×2, 2×3, and 3×3 reflecting the domains listed in Section 5. As also mentioned therein, we use the corrected contingency coefficient to express the strength of associations. Since – for a fixed and known number of observations – the absolute frequencies can be directly derived by the relative frequencies, and vice versa, this association measure is also suitable for tables of relative frequencies and leads to the same results.

In a first step, we generate a set S of relative frequency tables that represents the set of all possible frequency tables of above-mentioned sizes. S is created by taking all combinations of two discrete (marginal) probability distributions, whose event probabilities are strictly positive and on a one-percent grid. This strict positivity is needed because zero entries in the marginal distributions lead to zero entries in the table under independence. This entails that the $\mathcal{X}^2$ coefficient and all association measures based

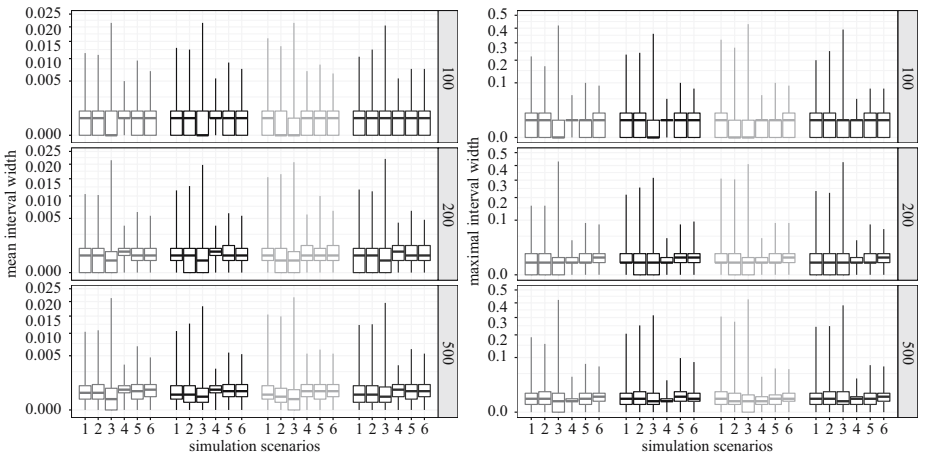

Fig. 5. Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of ***X****,* ***Y****,* ***Z*** *for domain imputation.*

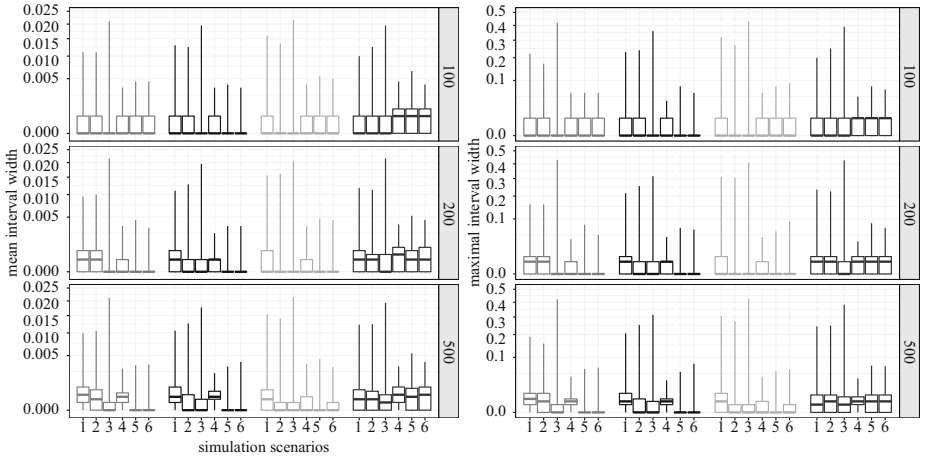

*Fig. 6. Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of **X**, **Y**, **Z** for variable-wise imputation.*

on it are not defined. S covers a large variety of marginal distributions and association measures ($|S| = 48\ 044\ 502$).

In a second step, we randomly draw one frequency table from S^* for each bivariate association depicted in Figure 2, where $S* \subseteq S$ denotes the set of probability tables that meets all predefined requirements for a specific simulation setting. Afterwards, we multiply the selected tables of relative frequencies with the desired number of observations and create a data file with complete observations $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. To meet the challenges of a statistical matching framework, we split this data file into two parts that represent the files A and B with $n_{\mathsf{A}} = n_{\mathsf{B}}$, and remove the observations $\boldsymbol{z}$ from A and $\boldsymbol{y}$ from B, respectively.

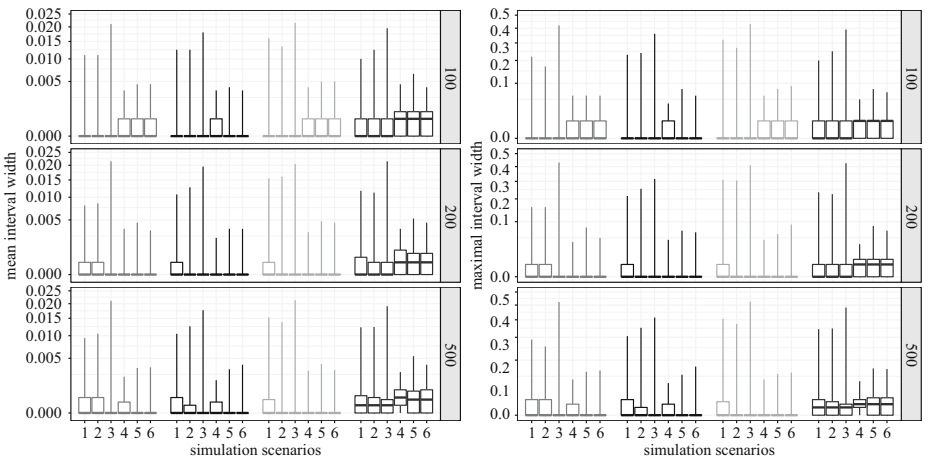

*Fig. 7. Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of **X**, **Y**, **Z** for case-wise imputation.*

7.3. Appendix C. Simulation results

Figures 3–7 show the interval widths of the parameter estimates on the partially set-valued synthetic data, aggregated for 20 simulation runs. The graphics are grouped by the different dependence designs (see Figure 2) and the numbers of observations. The results are displayed separately for the parameters of the marginal distributions and the parameters of the joint distributions. The whiskers range from the minimum to the maximum to ensure better readability. Please note that while the interval widths for the components of the joint distribution are reported on a square root scale to spread the values and make the different results more visible, the values themselves are not transformed.

The figure showing the mean and maximal interval widths of the components of the marginal distributions of the specific variables for domain imputation is not shown here since the interval widths are 0.5 for all simulation scenarios. This is no coincidence and results deterministically from the numbers of observations n_A and n_B.

8. References

Ahfock, D., S. Pyne, S.X. Lee, and G.J. McLachlan. 2016. "Partial Identification in the Statistical Matching Problem." *Computational Statistics & Data Analysis* 104: 79–90. Doi: https://doi.org/10.1016/j.csda.2016.06.005.

Andridge, R.R. and R.J.A. Little. 2010. "A Review of Hot Deck Imputation for Survey Nonresponse." *International Statistical Review* 78: 40–64. Doi: https://doi.org/10.1111/j.1751-5823.2010.00103.x.

Augustin, T., Coolen, F.P.A., de Cooman, G., and Troffaes, M.C.M. (Eds.). 2014. *Introduction to Imprecise Probabilities*. Chichester: Wiley. Doi: https://doi.org/10.1002/9781118763117.

Barbiero, A. and P.A. Ferrari. 2017. "An R Package for the Simulation of Correlated Discrete Variables." *Communications in Statistics – Simulation and Computation* 46: 5123–5140. Doi: https://doi.org/10.1080/03610918.2016.1146758.

Cattaneo, M. 2013. "Likelihood Decision Functions." *Electronic Journal of Statistics* 7: 2924–2946. Doi: https://doi.org/10.1214/13-EJS869.

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: 354–365. Doi: https://doi.org/10.1016/j.csda.2008.07.041.

Conti, P.L., D. Marella, and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28: 69–88. Available at: http://www.scb.se/dokumentation/statistiska-metoder/JOS-archive/ (accessed July 2019).

Conti, P.L., D. Marella, and M. Scanu. 2017. "How Far from Identifiability? A Systematic Overview of the Statistical Matching Problem in a Non Parametric Framework." *Communications in Statistics Theory and Methods* 46: 967–994. Doi: https://doi.org/10.1080/03610926.2015.1010005.

Couso, I. and D. Dubois. 2014. "Statistical Reasoning with Set-valued Information: Ontic vs. Epistemic Views." *International Journal of Approximate Reasoning* 55: 1502–1518. Doi: https://doi.org/10.1016/j.ijar.2013.07.002.

Couso, I., D. Dubois, and L. Sánchez. 2014. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Cham: Springer. Doi: https://doi.org/10.1007/978-3-319-08611-8.

De Campos, L.M., M.T. Lamata, and S. Moral. 1990. "The Concept of Conditional Fuzzy Measure." *International Journal of Intelligent Systems* 5: 237–246. Doi: https://doi.org/10.1002/int.4550050302.

Dempster, A.P. 1967. "Upper and Lower Probabilities Induced By a Multivalued Mapping." *The Annals of Mathematical Statistics* 38: 325–339. Doi: https://doi.org/10.1214/aoms/1177698950.

Denoeux, T. 2016. "40 Years of Dempster-Shafer Theory." *International Journal of Approximate Reasoning* 79: 1–6. Doi: https://doi.org/10.1016/j.ijar.2016.07.010.

Di Zio, M. and B. Vantaggi. 2017. "Partial Identification in Statistical Matching with Misclassification." *International Journal of Approximate Reasoning* 82: 227–241. Doi: https://doi.org/10.1016/j.ijar.2016.12.015.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006a. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints." *Journal of Official Statistics* 22: 137–157. Available at: http://www.scb.se/dokumentation/statistiska-metoder/JOS-archive/ (accessed July 2019).

D'Orazio, M., M. Di Zio, and M. Scanu. 2006b. *Statistical Matching: Theory and Practice*. Chichester: Wiley. Doi: https://doi.org/10.1002/0470023554.

D'Orazio, M., M. Di Zio, and M. Scanu. 2017. "The Use of Uncertainty to Choose Matching Variables in Statistical Matching." *International Journal of Approximate Reasoning* 90: 433–440. Doi: https://doi.org/10.1016/j.ijar.2017.08.015.

Dubois, D. and H. Prade. 1992. "Evidence, Knowledge, and Belief Functions." *International Journal of Approximate Reasoning* 6: 295–319. Doi: https://doi.org/10.1016/0888-613X(92)90027-W.

Fagin, R. and J.Y. Halpern. 1991. "A New Approach to Updating Beliefs." In *Uncertainty in Artificial Intelligence*, edited by P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, 347–374. New York: Elsevier.

Fink, P., E. Endres, and M. Schmoll. 2019. *impimp: Imprecise Imputation for Statistical Matching*. https://CRAN.R-project.org/package=impimp. (accessed July 2019).

Joenssen, D.W.H. 2015. *Hot-Deck-Verfahren zur Imputation fehlender Daten – Auswirkungen des Donor-Limits [Hot-Deck Procedures for the Imputation of Missing Data: Effects of the Donor Limit, translation by the authors]*. Ph. D. thesis, Technische Universität Ilmenau. Available at: https://www.db-thueringen.de/receive/dbt_mods_00026076. (accessed July 2019).

Kim, J.K. and W. Fuller. 2004. "Fractional Hot Deck Imputation." *Biometrika* 91: 559–578. Doi: https://doi.org/10.1093/biomet/91.3.559.

Lin, J. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37: 145–151. Doi: https://doi.org/10.1109/18.61115.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data* (2nd ed.). Hoboken: Wiley. Doi: https://doi.org/10.1002/9781119013563.

Manski, C.F. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, C.F. 2007. *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

Miranda, E., I. Couso, and P. Gil. 2010. "Approximations of Upper and Lower Probabilities By Measurable Selections." *Information Sciences* 180: 1407–1417. Doi: https://doi.org/10.1016/j.ins.2009.12.005.

Nguyen, H.T. 1978. "On Random Sets and Belief Functions." *Journal of Mathematical Analysis and Applications* 65: 531–542. Doi: https://doi.org/10.1016/0022-247X(78)90161-0.

Nguyen, H.T. 2006. *An Introduction to Random Sets*. Boca Raton: Chapman & Hall/CRC.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/. (accessed July 2019).

Ramoni, M. and P. Sebastiani. 2001. "Robust Learning with Missing Data." *Machine Learning* 45: 147–170. Doi: https://doi.Org/10.1023/A:1010968702992.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer.

Serafino, P. and R. Tonkin. 2017. "Statistical Matching of European Union Statistics on Income and Living Conditions (EU-SILC) and the Household Budget Survey." In *Eurostat: Statistical Working Papers*. Luxembourg: Publications Office of the European Union. Doi: https://doi.org/10.2785/933460.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

Vantaggi, B. 2008. "Statistical Matching of Multiple Sources: A Look Through Coherence." *International Journal of Approximate Reasoning* 49: 701–711. Doi: https://doi.org/10.1016/j.ijar.2008.07.005.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Yang, S. and J.K. Kim. 2016. "Fractional Imputation in Survey Sampling: A Comparative Review." *Statistical Science* 31: 415–432. Doi: https://doi.org/10.1214/16-STS569.

Received May 2018
Revised December 2018
Accepted April 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 625–651, http://dx.doi.org/10.2478/JOS-2019-0026

A Lexical Approach to Estimating Environmental Goods and Services Output in the Construction Sector via Soft Classification of Enterprise Activity Descriptions Using Latent Dirichlet Allocation

Gerard Keogh[1]

The research question addressed here is whether the semantic value implicit in environmental terms in an activity description text string, can be translated into economic value for firms in the construction sector. We address this question using a relatively new applied statistical method called Latent Dirichlet Allocation (LDA). We first identify a satellite register of firms in construction sector that engage in some form of environmental work. From these we construct a vocabulary of meaningful words. Then, for each firm in turn on this satellite register we take its activity description text string and process this string with LDA. This *softly-classifies* the descriptions on the satellite register into just seven environmentally relevant topics. With this seven-topic classification we proceed to extract a statistically meaningful weight of evidence associated with environmental terms in each activity description. This weight is applied to the associated firm's overall output value recorded on our national Business Register to arrive at a supply side estimate of the firm's EGSS value. On this basis we find the EGSS estimate for construction in Ireland in 2013 is about EURO 229m. We contrast this estimate with estimates from other countries obtained by demand side methods and show it compares satisfactorily, thereby enhancing its credibility. Our method also has the advantage that it provides a breakdown of EGSS output by EU environmental classifications (CEPA/CReMA) as these align closely to discovered topics. We stress the success of this application of LDA relies greatly on our small vocabulary which is constructed directly from the satellite register.

Key words: Latent dirichlet allocation (LDA); environmental goods and services (EGSS); satellite register; lexical analysis; supply side estimates.

1. Introduction

Whether it is carbon emissions, increasing global temperatures or the depletion of natural resources such as woodland or water, it is evident that human activity affects the environment. This development has led to an increasing focus on man-made factors that impact the environment and a consequential need to measure and monitor those factors. Interestingly, the natural tendency is to highlight harmful effects such as pollution or increasing global temperatures while efforts that enhance or sustain the environment, such as insulating our homes or producing energy from renewable sources, tend to be given somewhat less prominence. Evidently, from a policy perspective it is important to be able to combine measures of harmful effects with enhancing and sustainable effects, to gain a

[1] Central Statistics Office, Ardee Road, Rathmines, Dublin 6, Ireland. Email: Gerard.Keogh@cso.ie

fuller picture of human impact on the environment. From a statistical point of view EU Regulation 691/2011 (EU-691 2011) on European Environmental Economic Accounts (EEEA) is a framework to build a fuller picture. It provides for the collection of national level data on harmful factors, such as air emissions and material balances that monitor the use of natural resources, as well as mitigating effects such as environmental taxes (e.g., carbon tax) that dis-incentivise harmful means of production.

Regulation 691/2011 (EU-691 2011) also incorporates a module on Environmental Goods and Services Sector (EGSS). This measures the economic value (gross output) of 'eco-industries' (i.e., 'the green economy'). Under this module, member states in the EU are obliged annually from 2017 onward to report data on output value, exports, employment and gross value added in the production of goods and services that mitigate environmental damage, or manage natural resources in a sustainable way. Accordingly, estimating EGSS is now a looming obligation for National Statistics Institutes (NSIs) within the EU. In this article we set out a completely novel supply side approach to estimate EGSS output. Our approach is based on *lexical analysis* of the textual activity description of each firm held on our national Business Register (BR). We use a relatively new applied statistical tool called Latent Dirichlet Allocation (LDA). We apply this to each firm's activity description in turn, with a view to extracting the 'weight of evidence' of the environmental component from the activity description. The resulting weight is multiplied by the firm's total output to estimate the portion of output that is likely to be purely environmental in origin.

Both the OECD manual (OECD 1999) and Eurostat's Practical Guide for Completion of EGSS Accounts (Eurostat 2015) suggest two approaches to estimate the output value of EGSS. First, the demand side approach is based on National Accounts (NA) expenditure aggregates. Often this can be relatively straightforward in that output is largely the NA expenditure aggregate, or some part thereof, for a particular environmental sector; examples include water services, waste water treatment and waste collection and disposal. Second, in contrast to the NA based demand side methodology, the supply side approach, where practicable, typically takes two forms:

a) Using Structural Businesses Survey (SBS) data from primary suppliers of environmental goods and services (e.g., Prodcom), possibly supplemented by a small survey focussed on businesses in specific industry sectors.
b) Conducting a full specialist survey of businesses active in the green economy; an example is the Green Goods and Services Survey (BLS 2011) conducted by the Bureau of Labor Statistics (BLS) in the United States. We note this survey was only run in the years 2010 and 2011 before being cancelled as a result of budget cuts.

Evidently, both demand and supply side approaches have their strengths and weaknesses. Indeed, while the demand side can provide accurate aggregate values, it cannot readily distinguish between different products or services. Accordingly, it can be difficult to identify the purely environmental component of output in this approach. Meanwhile, even though the SBS based supply side approach can differentiate between different products or services, its coverage of environmentally specific products and services can be limited. To overcome this limitation some NSIs will supplement SBS sources with a small survey focussed on specific industry sectors. Of course, a full specialist survey such as the GGSS is likely to yield the most robust estimates of green

output. However, this approach tends to be avoided by smaller NSIs due to attendant costs, and burdens it imposes on respondent businesses.

Estimating EGSS output for a complex sector like construction is particularly problematic, as even well run surveys of the sector are likely to elicit low response levels. Resulting estimates of overall green output are likely to be of poor quality, while CEPA (Classification of Environmental Protection Activities) and CReMA (Classification of Resource Management Activities) breakdowns of EGSS output required under the regulation will be even poorer still. In this situation, a fairly common work-around involves applying appropriate industry specific factors gleaned from experts in the field to existing NA output aggregates (e.g., Statistics Estonia 2015). For example, in the construction sector an appropriate factor might be arrived at by Quantity Surveyors pooling their knowledge of construction costs across a variety of 'standard' construction projects, such as building a typical three-bed home (e.g., RICS 2016). While this approach is sensible it takes little or no account of the specific emphasis of individual firms within the sector. Consequently, without a specific satellite construction register being in place, a firm that is involved in the construction of bridges is treated similarly to one that installs attic insulation, leading to poor estimates and potentially biased breakdowns by CEPA or CReMA.

For a complex sector like construction this seems anomalous and suggests a prerequisite for accurate measurement is the development of an appropriate satellite register, in our case an environmentally specific construction register. If, moreover, we categorise this satellite register by type/class of environmental activity and obtain an expert factor for each class, then we should be able to arrive at a fairly sensible estimate of EGSS output. Clearly, a natural way to determine appropriate classes of environmental activity in construction is to identify meaningful common themes or topics, and use these as a basis for computing an EGSS output estimate. The purpose of this article is to describe how these topics can be learned directly based on a lexical analysis of activity descriptions on our satellite construction register. Furthermore, we show how this knowledge may then be used to arrive at an estimate of EGSS output. We accomplish this using a relatively new applied statistical tool called Latent Dirichlet Allocation (LDA). We show that LDA can learn meaningful environmental topics latent within activity descriptions on our satellite construction register; we note this is a novel application of LDA. Based on the relative importance of these environmental topics within a business's activity description text, we are able to compute an EGSS 'weight of evidence' factor for that business. Importantly, this firm/business level evidence weight reflects the *semantic emphasis* a particular construction business places on those environmental goods and services it supplies. We multiply this weight by the most recent overall supply side output value for that business, as recorded on the BR, to compute an estimate of the value of EGSS output. Summing this across all construction businesses on the satellite register we arrive at an EGSS value for the whole construction sector. We emphasise the 'weight of evidence' that we compute is chiefly a novel by-product of using LDA and therefore renders LDA useful in areas well beyond the field of pure text processing.

The remainder of this article is organised as follows. In Section 2 we discuss the rationale behind our approach, we feel this is necessary because the basis of our approach is quite different to the traditional supply side methodology and justification is therefore needed. Section 3 describes the salient features of LDA's statistical model and Bayesian

inference for this model. This section is a little technical and may be browsed by a reader interested primarily in applications. In Section 4 we describe how we arrive at our satellite construction register and identify a vocabulary of environmentally relevant words based on this register. Section 5 addresses the important issue of model evaluation for topic models. In Section 6 we use the best topic model identified in Section 5 to arrive at an estimate of EGSS output, based on the emphasis of environmentally relevant words in activity descriptions on the satellite register. We compare the estimates with those of other countries and find our overall estimate of EGSS output to be marginally on the low side. We also provide a statistically meaningful further breakdown of our overall estimate by CEPA and CReMA. Section 7 concludes.

2. The Basis of Our Approach

Estimating the weight or portion of a construction business's output that is environmental using LDA is the key novel contribution of this article. The rationale behind this approach is that a firm's description of its activity, as recorded in the activity description itself, stresses the main types of work it carries out. Relative (semantic) weights computed from text analysis of activity descriptions, therefore reflect the relative weight firm's place on different types of work they carry out. Clearly, it is to be expected that the types of work described are also the principle sources of the firm's revenue. Accordingly, it is reasonable to assert that the distribution of relative weights is also reflective of the relative contribution of each type of work (mentioned in the activity description) to the value of economic output. So, for example, a business installing insulation and constructing internal partition walls (i.e., dry walling) has an environmental weight and a pure construction weight. In this article, we show how we can use LDA to compute the relative weight of these two components. Sensibly, we assume these are its principle sources of revenue. The relative weight for insulation computed via LDA reflects the prevalence of insulation in all activity descriptions, in each individual business's particular description, as well as the prevalence of insulation in each topic and across all topics latent in the satellite construction register. Importantly, activity descriptions are classified probabilistically (i.e., a soft classification) according to topic. The amount of total probability that LDA finds in an activity description that is attributable to environmental terms (i.e., words in a description) such as insulation, is a measure of the weight of evidence associated with environmental topics in a particular description and across all descriptions.

Interestingly, the essence of our rationale is that it seeks to mimic the process that an official in a statistics office would apply when forming an impression of the key activities undertaken by a business. Based solely on that business's description of itself and descriptions from similar businesses in that sector, an official would form an impression of the main types of work carried out and their relative importance. Naturally, they would also seek out sensible relevant features (i.e., topics), and use these to arrive at a *refined* sense of the relative importance of the types of work of the business. Further, and in the absence of other sources, if they were required to estimate the distribution of the business's main sources of revenue, with good reason they might adopt weights derived via *refined* relative importance. Of course, the business's self-description may be an activity description as used here, or it may be a description obtained for example from googling the business's website.

We note two particular strengths of our approach that should be highlighted. First, in our implementation of LDA we construct the word vocabulary needed by LDA directly from the satellite construction register itself. This contributes greatly to LDA's success in finding meaningful topics from the activity descriptions on our satellite register. Accordingly, we avoid a common pitfall of using LDA as a *black box* to extract topics from a sea of documents, only to find the topics found have little relevance to real meaningful concepts. Second, as a consequence of finding meaningful topics we are able to map these topics very closely to CEPA and CReMA classifications. Thus, LDA provides us with a statistically meaningful set of relative weights (i.e., distribution) at the activity description level, and therefore at the firm level also, for these classifications. We use these firm level weights to allocate the overall EGSS output estimate according to CEPA and CReMA at the firm level. This is a particularly valuable bonus to adopting our approach.

Significantly, we stress that we do not hold the view that this lexical approach should replace existing demand or supply side methods, but rather provide a complementary method of estimating output from the supply side. Moreover, with this in mind we have programmed the core LDA method in SAS/IML. We feel this may facilitate its wider availability to the official statistics community and in other applied areas such as biostatistics, which often rely on proprietary software systems for statistical analysis. It is also worth mentioning that R has two packages for topic model analysis, one called '*lda'* (Chang and Dai 2015) and the other called '*topicmodels'* (Hornik and Grün 2011). Some of the methodology and analysis conducted here could also be accomplished with these implementations. Equally, proprietary software called MALLET (McCallum 2002) is also available for topic model analysis. As these implementations do not quite fit our needs we have found it expedient to re-work LDA from scratch.

3. Latent Dirichlet Allocation (LDA)

Coding a large data set of natural language textual descriptions (i.e., a corpus of documents) such as business activity, occupation or morbidity is a commonplace job within NSIs. Typically, a coder will *hard code* the description to a single class. Unfortunately, more often than not coding is inexact. In this situation the coder arrives at the appropriate class, via an initial *soft assignment* of two or more classes, based on similarity or relevance judgements and domain specific expert knowledge. The coder then picks the appropriate class from the soft classes based on their relative probabilities or evidence. An important feature of this expert *soft-coding* is its reliance on making an informed choice based on the most sensible combination of relevant themes or topics in the text description. LDA (Blei et al. 2003) is a fully Bayesian procedure that seeks to replicate a first-order approximation to the soft-coding processes of domain experts. Accordingly, it is likely to be of interest to official statisticians and prove beneficial where register development and analysis is needed, as is the case in the realm of EGSS.

The statistical model underpinning LDA relies on a generative model that links a document, labelled by d, in a corpus of D documents, to a set of W unique words in a vocabulary via a latent or hidden set of relevant topics. In LDA topics are typically labelled by the random variable z, the overall number of topics K is constant and assumed *a priori*. The generative model postulates that each document in a corpus is generated

by first picking a multinomial distribution with a K-vector of (topic) parameters $\theta_{jd}(j = 1, \ldots, K)$, where $\sum_{j=1}^{K} \theta_{jd} = 1$. This means the vector of topic parameters θ_{jd} is fixed for the specific document, but varies from document to document in the corpus. For the current document, the K vector of parameters θ_{jd} is generated from a Dirichlet prior distribution with hyperparameter α (which by the way can also be a vector of size K), this, of course, ensures the topic parameters sum to 1. In other words, a prior Dirichlet is used to generate a set of multinomial probabilities θ_{jd} across K topics for the d^{th} document in the corpus. We call the resulting multinomial distribution with parameters $\theta_{jd}(j = 1, \ldots, K)$ generated in this way, the topic multinomial distribution for the document.

Within each topic, LDA's statistical model also specifies a separate multinomial distribution with a vector of parameters φ over all W unique words in the vocabulary. Each individual word in document d is then generated by picking a specific topic $z = j$ from the topic multinomial distribution. This fixes the multinomial distribution with parameter set φ_j for unique words from the vocabulary that occur in topic j, we call this the word multinomial distribution. The individual word in the document is then picked at random from the vocabulary based on the probabilities in this word multinomial distribution. This generative step determines the conditional probability $P(w|z = j) = \varphi_{wj}$ of choosing the word w under the word multinomial distribution for the j^{th} topic. Using the theorem of total probability, we can combine the marginal and conditional probabilities in a mixture model to compute the probability of a specific vocabulary word w as

$$P(w) = \sum_{j=1}^{K} P(w|z_j = j)P(z_j = j) = \sum_{j=1}^{K} \varphi_{wj}\, \theta_{jd} \tag{1}$$

with θ_{jd} the document specific probability (associated with vocabulary word w) in topic j. The full LDA statistical model also posits a Dirichlet prior with hyperparameter β on the (topic specific) word multinomial distribution φ_{wj}, this is used to generate the multinomial word distribution for that specific topic. We mention that the Dirichlet prior distributions are chosen because they are conjugate to the multinomial.

The generative model outlined above may seem somewhat elaborate but in practice it is quite straightforward. Unique words in the vocabulary are assigned probabilistically to a specific topic. Starting with two hyperparameters α and β, we generate a word in a document by first drawing a set of parameters θ_{jd} for topics from a specified $Dirichlet(\alpha)$ distribution; using a Dirichlet prior ensures $\sum_{j=1}^{K} \theta_{jd} = 1$. We then draw (i.e., sample) a specific topic $z = j$ from a multinomial distribution with parameters θ_{jd}. Separately, we draw a set of multinomial parameters φ_{wj} for words in topic j from a $Dirichlet(\beta)$ distribution; once again using a Dirichlet prior ensures $\sum_{w=1}^{W} \varphi_{wj} = 1$. The word is then drawn from the unique vocabulary of words by sampling from the multinomial distribution with parameters φ_{wj}. Repeating this procedure N times generates a document with N words. Further repeating the whole process D times, generates a corpus of D documents where each document is based on K topics.

Consider the following hypothetical example, in the realm of EGSS there might be three topics, *energy saving, renewables and recycling*. For document (i.e., activity description) d, the trinomial topic distribution is generated from a $Dirichlet(\alpha)$ with (probability) parameters $\theta_{d,energy\ saving} = 0.1$, $\theta_{d,renewables} = 0.7$ and $\theta_{d,recycling} = 0.2$. Then, topic $z =$

$2 = renewables$ might be selected based on sampling from this topic distribution. Assuming a 10 word vocabulary, we then generate the multinomial word distribution with (probability) parameters $\varphi_{1,2} = 0.82, \varphi_{2,2} = 0.02, \ldots, \varphi_{10,2} = 0.02$ from a $Dirichlet(\beta)$ for these ten words. Assuming *solar* is the first word in the vocabulary, we then might select it based on these probabilities. This process associates the word *solar* with the topic *renewables* assigned in the d^{th} document. Note, this association of word with topic is purely probabilistic as no other/external information is incorporated. Repeating this procedure N times generates a document with N words from the vocabulary having a trinomial topic distribution and repeating this document process D times generates a corpus based on three topics.

The above process describes how to generate a corpus based on a statistical model. However, in practice, interest centres on using this model as a basis for discovering the set of topics, from an observed corpus of documents and vocabulary of words. This estimation of a set of topics involves learning the matrix parameter sets φ and θ from the words in the corpus of documents. One clever strategy for doing this estimation, introduced by Griffiths and Steyvers (2004), is based on Gibbs sampling (Geman and Geman 1984). Interestingly, this approach avoids sophisticated approximations to difficult integrals of probability distributions that are functions of the parameter sets φ and θ, such as variational Bayes (Blei et al. 2003) or expectation-propagation (Minka and Lafferty 2002). Instead, it seeks to directly evaluate the posterior distribution over the assignments of words to topics $P(z|w)$ and recover the matrices of parameters φ and θ for the corpus of documents by examining this distribution. From Bayes Theorem we can write

$$P(z|w) = \frac{P(w,z)}{P(w)} = \frac{P(w|z)P(z)}{P(w)} \propto P(w|z)P(z) \tag{2}$$

Based on this form, Griffiths and Steyvers (2004) compute separate expressions for $P(w|z)$ and $P(z)$ that are functions of the word-topic counts (n_{wj}) and document-topic counts (n_{jd}) respectively. Using these quantities they derive the full conditional topic distributions required for Gibbs sampling; expressions for the full conditionals, as well as estimates of the (matrix) parameter sets $\hat{\varphi}$ and $\hat{\theta}$ computed from the respective word-topic and document-topic count matrices are given in the Appendix (Section 8). Full details on the derivation of these equations are also given in a number of articles, including Heinrich (2009), Wang (2008), and Carpenter (2010).

Heinrich (2009) also outlines an algorithm for implementing the Gibbs sampler. This too is straightforward as it relies on maintaining matrices for word-topic counts (n_{wj}) and document-topic counts (n_{jd}). The word-topic count matrix (n_{wj}) gives the number of times word w has been assigned to topic j in the vector of assignments z. Meanwhile, the document-topic count matrix (n_{jd}) gives the number of times a word from document d has been assigned to topic j. For the next word in the document, each Gibbs estimation step simply involves decrementing the current count for the topic assignment for that word in both matrices by 1, followed by resampling from the full conditional multinomial topic distribution (i.e., with the current topic excluded) to generate a new topic assignment. The word-topic and document-topic matrices for this word, new topic, and document combination are then incremented by 1. We mention that we have implemented this algorithm in SAS/IML and verified its performance on a novel test problem given in

Griffiths and Steyvers (2004). Interestingly, this test problem comprises 2,000 images (i.e., documents), each being a 5×5 grid of pixels (pixel = word), with the intensity of a pixel specified by an integer and representing the number of times the word occurs in the document. A set of ten topics is constructed; each topic is a 5×5 grid image with a horizontal or vertical white bar set against a black background. Each document is generated by sampling 100 pixels from these topics. The test of our SAS/IML implementation involved generating 500 documents from a vocabulary of 25 words, word 1 to word 25 laid out on 5×5 grid pattern, with these words assigned to topics mirroring those in Griffiths and Steyvers (2004). We ran our implementation for 200 Gibbs iterations and found it recovered the set of ten topics very well indeed, producing results very similar to those reported by Griffiths and Steyvers (2004). We also note that our implementation has two additional refinements; the first allows the Dirichlet prior hyperparameters α and β to be estimated by maximising the joint log-likelihood, Appendix Equations (A1) and (A2), over these hyperparameters via an additional Newton-Raphson step, while the second allows for the Dirichlet topic parameter α to be a vector of length K. We remark however, that in test runs on our EGSS satellite construction register data these refinements only improved on the estimate of the (joint) log-likelihood generated by the core Gibbs estimation routine by a fraction of one percent. In light of this, our analysis proceeds with a scalar topic parameter α, accordingly the $Dirichlet(\alpha)$ and $Dirichlet(\beta)$ distributions are symmetric.

4. The Satellite Register, Document Corpus and Creating the Vocabulary

In our case, the EGSS satellite construction register is a subset of NACE Divisions 41–43 (construction sector) on the CSO's National Business Register (BR). In all, there are over 28,000 entities in the construction sector that describe themselves using approximately 13,500 unique activity descriptions (*Note: after this research was initially completed, a BR coherence project resulted in a substantial increase of approximately 12,000 new businesses in the construction sector being added on to the BR*.). To create the EGSS satellite construction register we have manually scanned each unique activity description and marked it where it included an environmental phrase, such as, *insulation* or *solar* or *heat pump* etc. This process produced a set of 1,077 unique activity descriptions covering 1,228 businesses in the construction sector; our satellite construction register comprises these 1,228 businesses. Meanwhile, we take the set of 1,077 unique activity descriptions we have identified to be our corpus of unique documents (i.e., we simply take each document to be a single unique activity description in this corpus).

In practice, the performance of LDA depends critically on the relevance of the vocabulary. Firstly, we distinguish between words and terms, a word is a unique entry in the vocabulary while a term is the occurrence of a word in a document. Clearly then a word may appear several times as a term in a document. The vocabulary itself is made up of unigram words only, but with some exceptions, such as *heat pump* taken as *heatpump,* while a hyphenation like *Geo-thermal* is taken as *Geothermal*. We follow the practice used in Information Retrieval (IR) and build our vocabulary of relevant words directly from the corpus itself. Initially, each document is first cleaned of punctuation or other non-alphabetic symbols, misspellings corrected and so-called *stop words,* such as THE, IS,

THAT, HE, removed. An initial basic vocabulary of all the unique words is compiled from the terms in the cleaned corpus. Our *cleaned* corpus of 1,077 activity descriptions comprised 5,565 terms corresponding to 830 unique words. We then applied the so-called *tf-idf* scheme (Spärck 1972), a popular scoring method for documents in a corpus, to reduce this further. For each unique word in the vocabulary and each document we compute

$$tf\text{-}idf_{wd} = tf_{wd} \times idf_{w} \tag{3}$$

where tf_{wd} is the term frequency count for word w in document d, and idf_w is the inverse document frequency count, this measures the number of occurrences of each vocabulary word in the corpus (on the log scale). The end result is a word-by-document matrix whose entries are the *tf-idf* values for each vocabulary word in each document in the corpus. The appealing feature of *tf-idf* is that it identifies a set of words that is discriminative for documents in the corpus. Based on the resulting *tf-idf* values, which ranged from about 1.4 to 14, we selected *tf-idf* values of six or higher. This had the effect of removing about 90% of document word instances from the corpus while reducing the vocabulary to 642 words. We further scrutinised the vocabulary words rejected through *tf-idf* analysis and found the 90% cut-off to be too severe as it rejected some words such as *drywall, reclamation, earth, drain, forestry etc.*, which we felt should be in the vocabulary. Accordingly, we decided to scan the remaining 192 unique words and add back some words based on their relevance to construction or environment activity. Of these, we identified seventy unique words that we felt were relevant based on our knowledge of the construction and environmental sectors. Note, while we preferred higher *tf-idf* value words we did not simply select the next highest seventy *tf-idf* values from the 192 unique words. Thus, for example we added back word *UPVC* which has a *tf-idf* value of just 2.845, but is environmentally quite relevant in the fitting of UPVC windows and doors in homes. In any event, this process resulted in a vocabulary comprising 712 unique words that we felt were meaningful construction or environmental words. A full listing of the resulting vocabulary is shown in Appendix Table A2 where we have given a complete list of the *tf-idf* selected vocabulary words and those seventy words added back based on relevance. It is clear from the listing that words added back are relevant and should indeed contribute to improving classification with LDA on our corpus. Moreover, we highlight that the process of compiling the vocabulary was done while assembling a dictionary of environmental terms for EGSS and occurred well in advance of our implementation of LDA. Thus the vocabulary was not selected for LDA so as to specifically fit this corpus, accordingly we stress the results described here are not a consequence of over-fitting using LDA with this vocabulary on our corpus of unique activity descriptions. Interestingly, this vocabulary includes general words like *construction, house* and *system*, as well as more environmentally specific words such as *energy, solar* and *insulate*. This combination of general and specific words in the vocabulary is important, as these combine together probabilistically to generate meaning, and it is topic-specific meaning we are attempting to uncover using LDA. So, having both types of words present in documents will serve to enhance topic learning via LDA. We feed both the corpus of 1,077 unique activity descriptions and the set of 712 unique words in our vocabulary, into our LDA routine with a view to learning or extracting the set of EGSS relevant topics.

5. Identifying the Number of Topics (Model Selection/Evaluation) and Visualising Topics

LDA requires the number of topics K to be given as an input. Accordingly, it makes sense to find an optimum value for K. One method of model selection commonly used to measure performance in IR is to compute the *perplexity* for a subset of held out documents (Heinrich 2009). Roughly speaking, *perplexity* is a cross-validation type measure, found by updating the LDA word-topic and document topic count matrices, via running the Gibbs sampler on an unseen document.

However, *perplexity* has not been adopted widely by statisticians because it does not directly measure the probability or evidence $P(\tilde{d}|K=k) = \prod_{t=1}^{n} P\left(w_{\tilde{d},t}|K=k\right)$ for an unseen held-out document $\tilde{d}$, comprising n terms $w_1, \cdots, w_t, \cdots, w_n$ for words from the vocabulary. Note, generally we use the index j to label topics, but here the topic notation $K=k$ is adopted to distinguish the fact that the number of topics is fixed at k for each computation of the evidence associated with that value of k.

The LDA model assumes documents are independent and words in each document are also independent. Accordingly, from Bayes Equation (2) the evidence is in fact the normalising (probability) constant $P(w)$. Interestingly, Wallach et al. (2009) set out a number of methods to evaluate this quantity for LDA. Their analysis shows a number of methods including the Harmonic Mean Method used by Griffiths and Steyvers (2004) lead to poor estimates of $P(w)$. They offer two credible alternative methods: a Chib-style estimator and their so-called "left-to-right" algorithm. For our purposes we have re-coded their Mathlab Chib-style estimator (see http://people.cs.umass.edu/~wallach/code/etm/) in SAS/IML, with a view to finding an optimal value of K for the corpus of EGSS activity descriptions. Our procedure for finding the optimal K involved running LDA on 90% of the documents in our corpus and holding back 10%. We fixed the number of topics at k and ran LDA on the 90% corpus to get stable estimates of the word-topic (n_{wj}) and document-topic (n_{jd}) count matrices. These were fed into the Chib routine along with the 10% subset of documents held-out, and the evidence probability $P(\tilde{d}|K=k)$ computed for each held-out document. The overall probability for all held-out documents is simply the product of each document's evidence probability, as documents are assumed independent; this independence assumption is valid here, as our documents are activity descriptions from individual businesses that are independent of one another within the construction sector. We simulated this procedure 30 times with a different randomly chosen set of held out documents. This generated 30 estimates for the overall evidence probability. The mean and standard deviation of these 30 estimates is then computed. For accuracy, all computations are done on the log scale, accordingly, we report the Model Log Evidence probability for each setting of k in Figure 1. We mention that to some extent this is a *belt-and-braces* approach, as the resampling in the Chib estimator is designed to give unbiased estimates based on just one simulation.

The plot in Figure 1 shows the results from running the Chib estimation routine. The Mean Log Evidence for each model initially increases as a function of k, reaches a peak at around seven or eight and decreases thereafter. This kind of profile is often seen when varying the dimensionality of a statistical model, with the optimal model being rich enough to fit the information available in the data, yet simple enough to avoid over-fitting

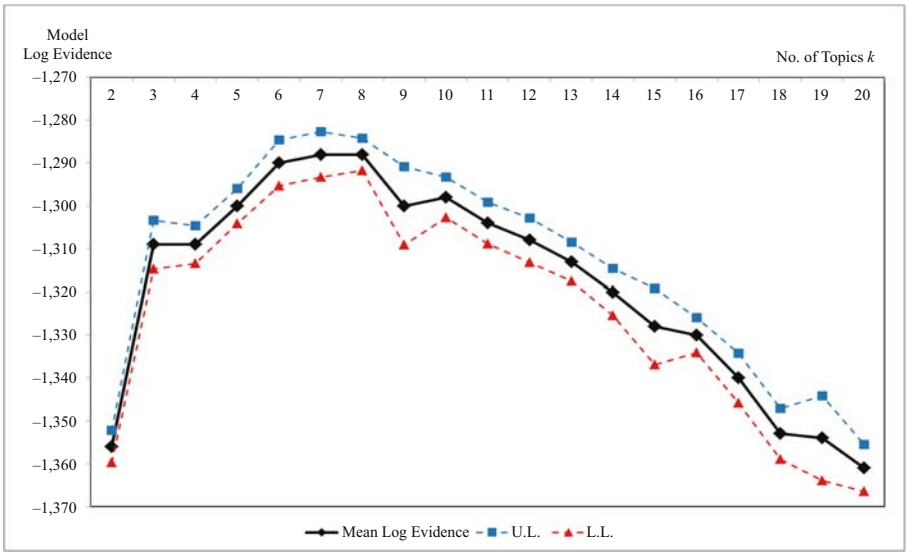

Fig. 1. Model log evidence for topic models.

that data. Typically, in an IR situation, with a corpus of millions of documents and a vocabulary of 10,000 words, finding a small optimal value for k would be unlikely. However, for this dataset the results are very appealing for two reasons. First, as most companies in the construction sector do similar work, we had expected there should only be a small number of topics related to construction within EGSS, and this turns out to be the case. Second, and far more importantly, we expected k to be small because we used a well-defined vocabulary constructed directly from the corpus itself. Accordingly, we expected LDA to find structure based on words having a fairly strong relevance to both EGSS and construction. The plot in Figure 1 also shows the two standard error lower and upper limits, labelled LL and UL respectively, arising from the 30 simulation runs. This band is quite narrow, demonstrating the stability of the Chib estimator and therefore attesting to the quality of the estimated log evidence probability, which is also appealing.

Naturally the value of k found using the Chib procedure depends on the Dirichlet prior hyper-parameters α and β. Each of the 30 runs in our simulation procedure assumed a fixed value k for the number of topics, and α and β initially set equal to 1 and 0.1 respectively. Setting α and β to a fixed constant, is nothing other than a shorthand means of forcing all k parameters in the corresponding Dirichlet distribution to be equal, the resulting distributions are therefore also symmetric. Nonetheless, after running the Gibbs procedure and before running the Chib procedure in each simulation, we also sought optimal values for α and β, given the optimal Gibbs assignments of topics to words in each document. Optimal values for α and β were found by maximising the joint log-likelihood, given in Equations A1 and A2 (Appendix), over these two parameters separately using a Newton-Raphson scheme. The mean values of the resulting estimates of α and β, across each of the 30 simulation runs, is shown in Figure 2 as a function of the number of topics k.

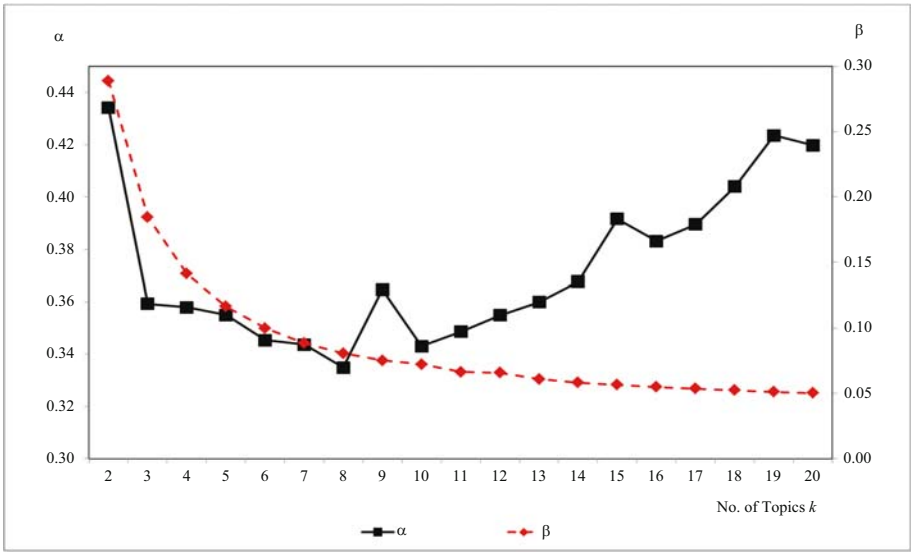

Fig. 2. Mean values for Dirichlet prior hyperparameters α and β for topic models.

It is clear from the plots in Figure 2 that α drops fairly rapidly, reaching a minimum of about 0.33 at about $k = 7$ or 8 topics. Typically, a smaller value for α will favour selecting the same few (i.e., 1 or 2) topic assignments for terms occurring in document d with high probabilities. In practice, this means words in this document can only be assigned to 1 or 2 meaningful topics, and more generally words will therefore tend to cluster strongly according to topic. Thus, as is the case here, when k is relatively low, a small value for α will ensure words cluster into a small number of meaningful topics. Meanwhile, β also drops fairly rapidly with increasing k, but the rate of decent appears to slow significantly at about $k = 7$ or 8 topics with $\beta = 0.07$. This too is appealing, as a small value for β is typical and can be expected to result in a fine-grained decomposition of the corpus into meaningful topics (Griffiths and Steyvers 2004).

Based on this model selection procedure it is clear that sensible settings for the parameters are $K = 7, \alpha = 0.33$ and $\beta = 0.07$, these values are used in all subsequent analysis. The top ten words from the vocabulary associated with this topic model are displayed in Table 1, with the topic titles having been named by us on pragmatic grounds. We *visualise* each topic $k = 1, \ldots, K = 7$, by ranking the words in that topic using their term-score (see Blei and Lafferty 2009)

$$\textit{term-score}_{wk} = \hat{\varphi}_{wk} \times \left(\frac{\hat{\varphi}_{wk}}{\left(\prod_{l=1}^{K} \hat{\varphi}_{wl} \right)^{1/K}} \right) \tag{4}$$

where $\hat{\varphi}_{wk}$ (see Section 8) is the estimated per-topic (vocabulary) word probability. This formula is inspired by the *tf-idf* scheme Equation (3) in that the second term in Equation (4) down-weights words that have high-probability under all topics.

It is clear from Table 1 that topics recovered by running LDA on our activity descriptions are environmentally meaningful. As suggested by the small value found for α

Table 1. Highest ranking term scores for vocabulary words by topic (K = 7).

Windows and doors	Insulation	Agriculture	Pure construction	Water and waste	Alternative energy	Energy saving pollution
WINDOW	INSULATE	BUILDING	TIMBER	WATER	SOLAR	ENERGY
DOOR	ATTIC	CLADDING	HOUSE	CONSTRUCTION	ENERGY	INSTALL
PVC	WALL	CONSTRUCTION	FRAME	WIND	PLUMBING	INSULATE
FITTING	CAVITY	TANK	CONTRACTOR	CIVIL	RENEWABLE	ENVIRONMENT
INSTALL	HOUSE	SLATTED	SITE	TREATMENT	HEAT	AIR
UPVC	BUILDING	SHED	DEMOLITION	HIRE	PUMP	BUILDING ENERGY RATING
GLAZING	ROOF	ROOFING	DRYLINING	SALE	WOOD	ACOUSTIC
ALUMINIUM	CLADDING	FARM	CEILING	INDUSTRY	PELLET	MEMBRANE
REPAIR	FLOOR	AGRICULTURAL	DRAINAGE	TURBINE	GEOTHERMAL	BARRIER
RECYCLING	EXTERNAL	SLURRY	HIRE	WASTE	STOVE	RADON

we see the words cluster nicely within each topic. We also see there is a good degree of discrimination between the topics and a natural degree of overlap, with words like CONSTRUCTION, ENERGY, INSTALL and INSULATE appearing in more than one topic. Meanwhile, words such as REPAIR and RECYCLING are only associated with the "Windows and Doors" topic. This too is quite agreeable, as these words are likely less important in the construction sector than they are in other NACE sectors of EGSS, such as recycling or waste collection and disposal.

Of course, while the results displayed in Table 1 are very appealing, there is the possibility that this topic classification by vocabulary word is to some degree a fluke for this particular value of $K = 7$. Accordingly, the sensitivity of this distribution to the *a priori* set number of topics K is of interest. Ideally, if LDA is stable and the Chib procedure for selecting K is robust, then we should see a topic-word distribution similar to Table 1 for values of K close to 7. To gain some insight into the sensitivity of LDA to the number of topics, we also examined the term score ranking distribution for $K = 6$ and $K = 8$ topics; the distributions are given in the Appendix, Table A1. First, comparing the seven topic distribution in Table 1 with the six topic distribution in Appendix Table A1, we can see a fair degree of similarity. LDA has found three very similar topics; 'Windows and Doors', 'Insulation' and 'Alternative Energy'. However, in this instance LDA has not distinguished words in the area of 'Construction' or 'Water, Waste and Energy Saving' as clearly as it did with $K = 7$ above. This is pleasing as we should see a more course-grained and less relevant set of topics when K is reduced below its postulated optimum value of seven. Second, in the eight topic case LDA seems to work quite well. It has nicely split construction into predominately 'internal' and predominately 'external' construction topics. In the context of the construction sector this seems a pleasing refinement. More importantly, this straightforward sensitivity test shows that LDA is sensitive to the value chosen for K in the best possible way. A small decrease in K from seven to six yields a more course-grained set of topics. However, an increase in K from seven to eight yields a small but appealing alteration to the topics discovered, which remain meaningful from an environmental perspective. We also mention that when we set $K = 20$, we find there are about ten topics that are meaningfully discriminated by LDA, but the other ten are more of a mixed bag. This suggests the Chib procedure is an effective means for determining an appropriate setting for the number of topics K that yields a decomposition of the corpus into meaningful concepts.

6. Using LDA to Estimate EGSS Output Value

The problem we face is simply stated; can we arrive at a meaningful estimate of EGSS value for the construction sector. As remarked in the Introduction, the EGSS Practical Guide (Eurostat 2015) offers no firm method of estimation for EGSS in this sector. Accordingly, EU member states are at liberty to use any credible approach to arrive at a realistic value. With a satellite construction register an ideal solution would be to select a sample and conduct a survey of firms on this register. As noted in the introduction the BLS operated this approach for their GGSS, selecting a sample from a satellite register of about two million environmental businesses in the United States and surveying those selected. An appealing refinement of the GGSS involves using a tailored survey for different sub-sectors, such as the renewable energy or the recycling industry.

Naturally, it would be pleasing to replicate the GGSS, but smaller NSIs typically may not have the resources to operate a specific survey focussed on the environmental sector. Nevertheless, an upper bound for EGSS output in the construction sector is directly available to us. We simply take overall output from the national BR and sum it across all those firms on the satellite construction register. When we do this, we find the resulting output value is EUR 540.8m. Clearly, this is a gross over-estimate of EGSS output, because there are many construction firms where only a portion of their activity is environmental.

Remarkably, if we allow the *data-to-speak-for-itself,* LDA provides a sound approach that enables us to estimate the portion of a construction firm's activity that is genuinely environmental. We turn our focus to the vocabulary and distinguish the subset of words that for all practical purposes are *genuinely environmental* type words, from those that are *essentially construction* sector type words. Examples of *genuinely environmental* words include renewable, solar, insulation and so on, while *essentially construction* words include building, house, construction, and so on. Intriguingly, *essentially construction* words occur frequently in activity descriptions on both the satellite register and the Main BR (NACE Divisions 41-43) covering the whole construction sector. Now, by simply matching the vocabulary with the words occurring on the Main BR, the purely environmental portion of the vocabulary can be tagged and separated from the pure construction portion of the vocabulary. This gives us a vocabulary of *genuinely environmental* type words, upon which we can compute a statistically meaningful weight of evidence favouring environmental activity in each activity description. This weight reflects the semantic emphasis latent in topics that a firm places on the environmental aspects of its own activity description; we refer to it as the semantic weight sem_wt_d.

The Gibbs implementation of LDA maintains a matrix Z of dimension $D \times N$ (i.e., equal in dimension to the document X term matrix) with N being the number of terms in the longest description. This matrix records the most recent topic assignment of the Gibbs Sampler for each term in each document/description. When the Gibbs sampler reaches a steady state, the topic assignments in Z for each term in each document become fixed. Based on these assignments the posterior estimates of word-topic probabilities $\hat{\varphi}_{wj}$ and document-topic probabilities $\hat{\theta}_{jd}$ (see Equations A4 and A5 in the Appendix) are computed. Computing sem_wt_d for d^{th} description proceeds based on these posterior estimates of word-topic and document-topic probabilities. From Equation (1) we can see the probability of each term t, associated with unique vocabulary word w, in each activity description on the satellite register is $\hat{\varphi}_{wj} \times \hat{\theta}_{jd}$. Thus, in steady state, for the t^{th} term in d^{th} description we fix $z = Z_{dt} = j$ for that document-term and compute the corresponding term probability $p_{td} = \hat{\varphi}_{t=w,Z_{dt}=j} \times \hat{\theta}_{Z_{dt}=j,d}$. We define the total term weight to be the sum of these probabilities for all terms matching each vocabulary word w in that description; we label this total term weight for all terms $T(W)_d$, where W is the set of vocabulary words in the description – clearly this sum of probabilities will not in general be equal to one. Similarly, by eliminating the *essentially construction* words from this description, we can identify and retain only those specific term probabilities associated with *genuinely environmental words* in this activity description; we label the resulting total (genuinely environmental) term weight $T(W^*)_d$; by definition this quantity will always be less than or equal to $T(W)_d$ because $W^* \subseteq W$ for all those vocabulary words that appear as terms in

the d^{th} description. We define the evidence favouring environmental strength of meaning in each activity description, as the ratio of the total term weight due to environmental words to that of all words in the description (under the LDA model), this quantity is

$$sem\_wt_d = \frac{P(W^*)_d}{P(W)_d} \tag{5}$$

Multiplying the firm's overall output by this weight gives a statistically meaningful estimate of the output value the construction firm attaches to its environmental activity. Clearly, the stress a firm places on the environmental aspects within its activity description also reflects the economic importance it attaches to these functions. Accordingly, the output estimate derived from directly measuring the relative importance of those environmental aspects via sem_wt_d also has sound economic credibility. We also note the estimated term weight in a description is a linear function of the estimated probabilities p_{td}. Thus, for all words in a given description we have $T(W^*)_d + T(W^\#)_d = T(W)_d$, where $T(W^\#)_d$ is the overall term weight for essentially construction words in that description.

Of course the value of sem_wt_d for activity description depends on the probabilities assigned to *essentially construction* or *genuinely environmental* words in that description. At first sight therefore it would appear that a description with more common *essentially construction* words will have a smaller semantic score than a description with less common *essentially construction* words. This scenario implies that given the same environmental words, the description with more common construction words will be considered less environmental (having a smaller semantic score). This seems anomalous, as the existence of common construction words should not necessarily mean the activity description is less environmental. However, it will be clear from the preceding paragraph that sem_wt_d probabilities are a function of not just of the word, but also the actual topic assigned to that word in the description from the topic assignment matrix Z. Importantly, this varies for the same word across different topics and descriptions. For an essentially 'Construction' topic the scenario outlined above is likely to occur as the probability will be primarily word dependent. But in a topic like 'Alternative Energy' it is far less likely, since topic assignments in the Z matrix will be associated with the respective topic and word in that description. Interestingly, this is quite an appealing feature of LDA as it generates a statistically meaningful score that is dependent on both the type of word and topic assigned to that word at the term level within each activity description.

Implementing this (sem_wt_d) computation, we generated the genuinely environmental vocabulary by removing the set of essentially construction words. The construction words were identified by manually extracting activity descriptions on the satellite corpus that were essentially construction, for example *Carpenter and Builder* and matching these with the 712-word base vocabulary. We found that the base vocabulary was reduced from 712 words to a 249 genuinely environmental word vocabulary. However, in practice we have also found this 249 genuinely environmental word vocabulary turns out to be too restrictive. The reason for this is that certain activity descriptions, such as, "A CIVIL ENGINEERING PLANT HIRE AND DEMOLITION COMPANY" get a zero sem_wt_d value. Interestingly, we included this description on the satellite register as 'demolition' may also incorporate a latent recycling function. Accordingly, to account for this effect

we take the word 'demolition' to be a *latent synonym* for 'recycling' and include it on the genuinely environmental word vocabulary. Identifying and including all these *latent synonyms* our purely environmental word vocabulary increased its size from 249 words to 314 words. Using this extended genuinely environmental word vocabulary we are able to compute a non-zero sem_wt_d for all descriptions on our satellite construction register.

With an effective extended genuinely environmental word vocabulary in place, we computed estimates of the word-topic probabilities $\hat{\varphi}_{wj}$ and document-topic probabilities $\hat{\theta}_{jd}$. In practice, we ran 20 separate LDA simulations on the 1,077 descriptions on the satellite (environmental) register and computed the matrices $\hat{\varphi}_{wj}$ *and* $\hat{\theta}_{jd}$ on each run. Both sets of term weights $T(W)_d$ and $T(W^*)_d$ and the resulting sem_wt_d value, were then computed based on average the values of $\hat{\varphi}_{wj}$ *and* $\hat{\theta}_{jd}$ across the 20 simulation runs. A histogram (and kernel density estimate from Proc SGplot in SAS) of the resulting sem_wt_d values computed for each of the 1,077 activity descriptions is displayed in Figure 3. The plot is skewed to the left with a median value of about 0.41 and lower and upper quartiles of 0.14 and 0.66 respectively. Intriguingly, this tells us that 50% of construction firms on our satellite register, who describe themselves using explicitly environmental words, did so with a degree of environmental *semantic weight or emphasis* below 41%. Meanwhile, only 25% of the firms described themselves with an environmental *semantic weight* of 66% or higher. Recalling that stop-words have been eliminated from our descriptions, the distribution of semantic weight (sem_wt_d) values in Figure 3, suggests that companies in the sector engaged in environmental work, tend to see themselves first as construction companies and second as environmental companies. This shows the NACE coding of these companies, based on their primary activity, into the construction sector tends to be correct, which is a valuable and unforeseen quality assurance by-product of this analysis.

In Figure 4, the Estimated Environmental Output value distribution that results from multiplying each construction firm's output on the 2013 satellite construction register by sem_wt_d is displayed. Encouragingly, as one might expect for output or production value

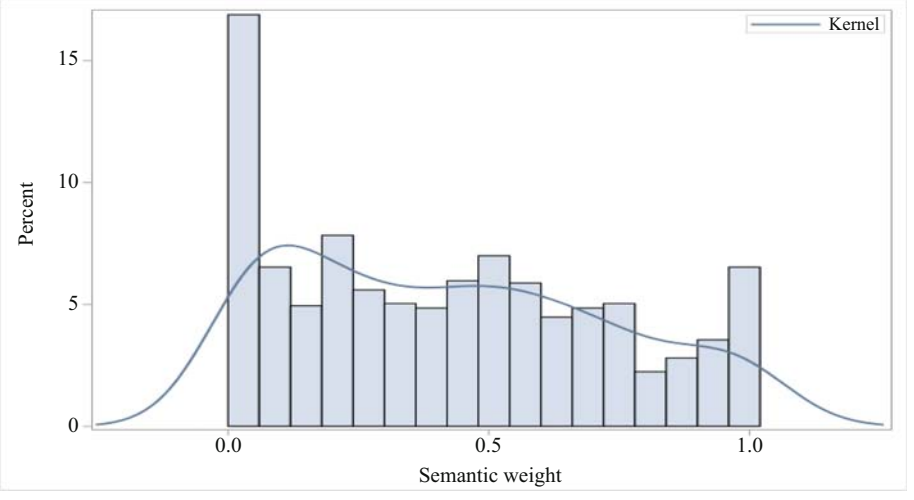

Fig. 3. Distribution of sem_wt_d for activity descriptions on the satellite construction register.

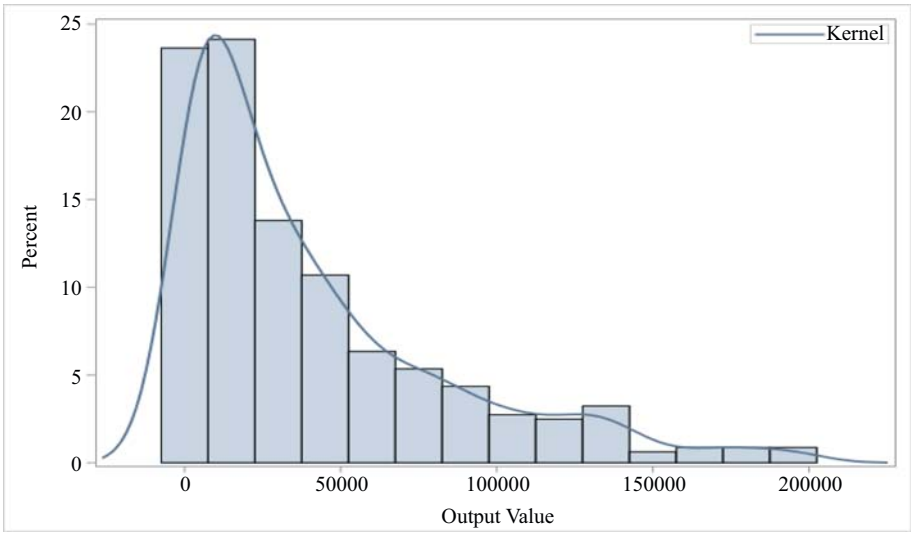

Fig. 4. Distribution of estimated environmental output on the satellite construction register.

data, the distribution is heavily skewed to the left. This is nothing other than a reflection that most construction firms tend to be small companies with a few employees and therefore tend to have a small overall output. The median of this distribution is EUR 36k with the lower and the upper quartiles being EUR 10k and EUR 107k respectively. More importantly, the overall total estimated EGSS output is EUR 229.2m. This value is 42.4% of the overall output value EUR 540.8m for all firms on the satellite construction register. This sem_wt_d based output estimate of EUR 229.2m therefore seems plausible, as many construction companies will typically have quite a mixed bag of activities, several of which will not be environmental. Meanwhile, our current best guess at overall EGSS output for 2013 in Ireland is about EUR 3.4bn, suggesting that the construction component likely accounts for about 6.7% of total EGSS output in Ireland. Interestingly, for international comparison, the 2012 UK EGSS output estimate for construction is 8.3% (ONS 2015) of total environmental output and the 2010 Estonian estimate is 10.5% (Statistics Estonia 2015). If these estimates are accurate, this indicates the sem_wt_d estimate at 6.7% may be somewhat on the low side.

Of course our output estimate is register based and therefore we are able to compute other measures such as output per employee and pay per employee, to further judge the quality of estimated output value. For our satellite register companies these values turn out to be EUR 130.4k and EUR 26.5k respectively, while the corresponding values for general construction companies on the national BR are EUR 119.1k and EUR 24.5k respectively. Considered in this light, our estimates give rise to per employee values that are consistent with the general construction sector in Ireland. Comparing internationally, the UK and Estonia estimates of output per employee are GDP 150.3k and EUR 50.6k respectively, and pleasingly our value of EUR 130.4k comes in close to the middle of these two estimates.

Thus, international evidence based on output per employee and national comparison of pay per employee shows there is a reasonable degree of consistency in our estimated

Table 2. Topic Assignments by CEPA/CReMA.

Topic	CEPA/CReMA Code	CEPA/CReMA class
Windows and doors	13B	Heat/Energy saving and management
Insulation	13B	Heat/Energy saving and management
Agriculture	9, 16	Other environment construction
Pure construction	9, 16	Other environment construction
Water and waste	2, 3	Waste Water management, Waste management
Alternative energy construction	1	Protection of ambient air and climate
Energy saving pollution	13B	Heat/Energy saving and management

output value. Of course the sem_wt_d estimate proposed here is based on emphasis or meaning in text and the degree to which this is causally linked to economic value remains uncertain. Nevertheless, in light of the comparisons made here and the sound statistical methodology (e.g., creating a satellite register and vocabulary construction) underlying the computation of sem_wt_d, it seems reasonable to assume the overall estimate of EUR 229.2m for EGSS in the construction sector in Ireland, based on sem_wt_d, is fundamentally sound.

Remarkably, we can glean more knowledge from our data than just the overall estimate of output value. Specifically, EU Regulation 691/2011 (EU 691,2011) also requires participating member states to provide a breakdown of output value by environmental protection (CEPA) and resource management (CReMA) classifications. Looking at the seven topics identified in Table 1 we can fairly readily associate these with classification headings within CEPA/CReMA as shown in Table 2.

Now, using the estimated document-topic matrix of probabilities $\hat{\theta}_{jd}$ we can allocate the output value of each firm on the satellite register, associated with the d^{th} description, according to these probabilities, giving the CEPA/CReMA value for each firm. Summing these values across all firms we arrive at the EGSS value in the construction sector broken down by CEPA/CReMA, the resulting sector totals are given in Table 3.

The figures in Table 3 show the largest sub-component of EGSS output in the construction sector in Ireland is EUR 98.5m and relates to the area of 'Heat/Energy saving and management'. This covers the provision of insulation and installation of

Table 3. Construction Value by CEPA/CReMA class.

CEPA/CReMA class	Value EUR(m)	Stand Error EUR(m)
Protection of ambient air and climate	32.7	10.7
Waste water management, Waste management	33.9	12.0
Heat/energy saving and management	98.5	20.3
Other environment construction	64.1	16.4
All	229.2	14.7

windows and doors in buildings. Separately we have estimated 'Heat/Energy saving and management' based on aggregate retro-fit insulation grant data (SEAI 2013) and new house construction data (see Department of Environment 2013) and obtained a value of between EUR 100m and EUR 115m. It is pleasing to see that the sem_wt_d estimate of EUR 98.5m computed here is close to the lower end of this interval, adding further to the credibility of our proposed approach. The standard error of the estimated value is also provided based on the 20 simulation runs. The overall standard error of the estimate across all CEPA/CReMA classes comes in at just over 6%, showing the estimated value is quite precise. Also interestingly, this breakdown comes at virtually no additional effort and therefore shows the considerable added value of using LDA to estimate output based on a set of relevant topics. We note that in practice this level of refinement would generally be possible only using a targeted survey such as the GGSS. However, here the time, cost and response burden associated with a specific survey have been avoided.

7. Closing Remarks

The key research problem addressed in this paper is whether and to what extent the semantic value provided in a construction firm's activity description text, informs us about the environmental economic value of the firm. The key assumption underlying this interconnection is, the emphasis a firm places on environmentally related terms in its descriptive text, will also reflect its economic focus and therefore the resulting productive value of the firm. Clearly, in this scenario, the output value of a firm that spends 95% of its time on pure construction work and just 5% of its time on environmental work will be overestimated, if the description comprises several environmental terms and few construction terms. However, this scenario contradicts reality as firms actually do emphasise the activities that are important to them in their activity description. Indeed, like many other NSIs, CSO in its annual BR survey specifically asks each firm to give, *as full a description as possible of its main activities,* and on this basis our underlying assumption seems valid. We also stress that our experience based on purposefully selecting 1,077 environmentally related construction activity descriptions on our satellite register tends to bear this out.

In this article, we used a relatively new applied statistical method called Latent Dirichlet Allocation (LDA) to search for meaning in activity description text strings, in the form of main topics or themes occurring on a satellite construction register. Using the activity descriptions on this register we constructed a vocabulary of 712 unique words needed as input for LDA. We also conducted a model evaluation study and established that our dataset of activity descriptions could be *softly-classified* into just seven environmentally relevant topics. With this seven topic classification we proceeded to extract the weight of evidence associated with environmental terms in each activity description. Based on LDA's estimated word-topic and document-topic probabilities, we proposed a statistically meaningful and environmentally relevant weighting factor. This is based on the ratio of the probability of *genuinely environmental words* in the activity description to the probability of all words; this ratio reflecting the semantic importance of the environmental aspects of the description conditional on the topic.

We applied the resulting evidence weight to the associated firm's overall output value to arrive at an estimate of the EGSS value for each construction firm. The quality of this estimate predicated on the assumption that environmental emphasis placed in the text by that firm, reflects the environmental economic value. On this basis, we arrived at an EGSS estimate for construction in Ireland in 2013 of EUR 229.2m. This accounts for about 41% of the overall output for all firms on the satellite construction register. Comparisons with two other countries, namely the UK and Estonia revealed that the value of our estimate as a proportion of total EGSS value appeared to be on the low side at 6.7%. With this caveat in mind, we viewed the estimated output value of EUR 229.2m for EGSS, arrived at here by analysis of environmental emphasis within activity descriptions, to be fundamentally sound. In addition, we are able to match the topics found by LDA with CEPA/CReMA classes leading to output classified by the latter. This is a valuable extra benefit to using LDA.

It cannot be overemphasised that we have been very purposeful in our use of LDA. Thus, as occurs in many other implementations, we have avoided the pitfall of using LDA as a *black box* to identify latent topics in a corpus of general construction descriptions that then might map to meaningful environmental concepts. Instead, we have pragmatically selected a corpus of environmental activity descriptions and prudently selected a vocabulary based on these descriptions. Thus, from the outset we have done considerable dimension reduction to our dataset before applying LDA. This has put in place the foundations to ensure a meaningful mapping between the topics LDA has discovered and real environmental concepts. Given these operational constraints, our results show that LDA is an impressive tool for identifying meaningful topics. Moreover, we feel this contributes greatly to enhancing the accuracy of our estimates of economic output derived from LDAs document-topic and word-topic probability distributions.

In the literature, LDA and its variants, such as dynamic topic models (Blei and Lafferty 2006), correlated topic models (Blei and Lafferty 2007), tagged or labelled LDA (Ramage et al. 2009) are used solely for text-based corpus analysis. These variants also extend LDA in various ways. By contrast, the analysis conducted here has been undertaken on a relatively small dataset with a small number of topics. Interestingly, we note the approach taken here, where we used a set of purely construction words from a pure construction source, is in essence a form of tagged LDA. Where it is possible to *a priori* tag certain descriptions beforehand with a tag that more precisely identifies economic value with the activity description and/or correlate descriptions, the variants mentioned may give rise to more credible estimates. For this reason and others noted earlier, we stress that the estimate of output arrived at here is not meant to replace estimates arrived at by other (e.g., demand side) means. Ideally, the estimate of EGSS output computed here should complement those others and indeed give a direct breakdown of output according to CEPA/CReMA, as required by the EGSS module in the EU Regulation.

8. Appendix

Using the notation in Section 2, we note from Equation (2) the multinomial distributions over the parameter sets for φ and θ only appear in $P(w|z)$ and $P(z)$ terms respectively. Moreover, as their respective Dirichlet priors are conjugate to these multinomial

distributions, both (matrix) parameter sets φ and θ can be integrated out to give the joint likelihood $P(w, z)$, which is proportional to the product of (see Griffiths & Steyvers 2004)

$$P(w|z) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta^W)}\right)^K \prod_{j=1}^{K} \frac{\prod_w \Gamma(n_{wj} + \beta)}{\Gamma(N_j + W\beta)} \tag{A1}$$

$$P(z) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha^K)}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_{jd} + \alpha)}{\Gamma(N_d + K\alpha)} \tag{A2}$$

where the entry n_{wj} in the word-topic count matrix (n_{wj}) give the number of times word w has been assigned to topic j in the vector of assignments z, the entry n_{jd} in document-topic count matrix (n_{jd}) gives the number of times a word from document d has been assigned to topic j and Γ is the standard gamma function. Both terms N_j and N_d are the respective topic and document totals of n_{wj} and n_{jd}, while W is the total number of words in the vocabulary. The full conditional topic distributions required for Gibbs sampling are computed from the resulting joint likelihood (see Griffiths and Steyvers 2004). More specifically, denoting the proposed topic to be assigned to term w_t in the d^{th} document by $z_t = (1 \cdots K)$, the full conditional topic distribution associated with this proposed latent assignment z_t is given by

$$P(z_t = j|z_{(-t)}, w) \propto \frac{n_{w_t,j(-t)} + \beta}{N_{j(-t)} + W\beta} \times \frac{n_{j(-t),d} + \alpha}{N_{d,(-t)} + K\alpha} \tag{A3}$$

where $(-t)$ denotes the exclusion of the proposed topic $z_t = j$ for word w_t and $N_{\cdot,(-t)}$ is the total of word-topic and document-topic counts $n_{\cdot,(-t)}$ of the current assignments $z_{(-t)}$ excluding the proposed topic $z_t = j$.

For any single sample we can estimate the word topic and topic document (matrix) parameter sets $\hat{\varphi}$ and $\hat{\theta}$ of probabilities respectively as

$$\hat{\varphi}_{wj} = \frac{n_{wj} + \beta}{N_j + W\beta} \tag{A4}$$

$$\hat{\theta}_{jd} = \frac{n_{jd} + \alpha}{N_d + K\alpha} \tag{A5}$$

The word topic Equation A4 is used to compute the term-score used in Section 4 to visualise words within their topics.

Table A1. Highest ranking term scores for vocabulary words by topic.

K = 6 Topics
Topic

Windows and doors	Insulation	Construction		Water, waste and energy saving	Alternative energy
		Agriculture	Other construction		
WINDOW	HOUSE	BUILDING	INSULATE	WATER	HEATING
DOOR	INSULATE	CLADDING	CLADDING	WIND	ENERGY
PVC	ATTIC	CONSTRUCTION	BUILDING	CONSTRUCTION	SOLAR
FITTING	WALL	TANK	CONTRACTOR	TREATMENT	PLUMBING
INSTALL	TIMBER	SLATTED	ROOFING	TURBINE	INSTALL
GLAZING	FRAME	SHED	HOME	INDUSTRY	RENEWABLE
REPAIR	CAVITY	ROOFING	ROOF	SCHEME	PELLET
UPVC	DRY	FARM	STEEL	BUILDING ENERGY RATING	WOOD
ALUMINIUM	LINING	AGRICULTURAL	SMALL	DISTRIBUTION	GEOTHERMAL
FASCIA	CEILING	SLURRY	ATTIC	RADON	BOILER

Table A1. Continued.

K = 8 Topics
Topic

Windows and doors	Insulation	Agriculture	Construction		Water and waste	Alternative energy	Energy saving and pollution
			Internal	External			
WINDOW	ATTIC	BUILDING	INSULATE	ROOFING	WATER	SOLAR	ENERGY
DOOR	INSULATE	CLADDING	CONTRACTOR	TIMBER	CONSTRUCTION	ENERGY	INSTALL
PVC	WALL	CONSTRUCTION	CLADDING	CLADDING	WIND	PLUMBING	HOME
INSTALL	HOUSE	TANK	DRY	DEMOLITION	CIVIL	RENEWABLE	ENVIRONMENT
FITTING	CAVITY	SLATTED	LINING	CONTRACTOR	TREATMENT	HEAT	AIR
GLAZING	BUILDING	SHED	CEILING	FRAME	HIRE	PUMP	BUILDING ENERGY RATING
UPVC	SMALL	ROOFING	DRYLINING	HOUSE	SALE	WOOD	ACOUSTIC
ALUMINIUM	SALE	FARM	PLASTERING	DRAINAGE	INDUSTRY	PELLET	MEMBRANE
REPAIR	INSTALL	AGRICULTURAL	SLABBING	LAND	TURBINE	GEOTHERMAL	BARRIER
FASCIA	EXTERNAL	SLURRY	DRYWALL	ROOF	WASTE	STOVE	RADON

Table A2. NACE Divisions 41–43 construction vocabulary word list.

Vocabulary words NOT selected by *tf-idf* and added by Author	Vocabulary Word selected by tf-idf scheme ordered by decending *tf-idf* rowwise (Note "S" dropped at word-end)									
EARTH	GRID	CONVERT	GEOTECHNICAL	PHONE	SERVICER	WINDOW	HIGH	COMPANIE	MAINTAIN	UNIT
STONE	DETECTION	CONVERTION	GIO	PHOTO	SHARE	WINDMILL	HOSPITAL	CONVERTING	MANAGEMENT	FITTER
SUSTAINABLE	LEAK	COOLER	GOVERNMENT	PHOTOGRAPHY	SHEEPWOOL	WOODPELLET	IMPORT	DECORATIVE	PARTITION	PRODUCT
LOG	BOREHOLE	CORE	GRAS	PLA	SIGNALLER	WRAPPING	INSTULATION	DELIVERY	PROPERTY	CIVIL
WATERMAIN	SOFTENER	CORRECTION	GROUNDSWORK	PLANTHIRE	SILO	COUNCIL	JOINER	DRILLING	SELLING	SALE
BIOMAS	PUMPING	CORRUGATE	GROWING	PLASTERBOARDIN	SKIM	WELL	JOINTING	DRIVER	SPRAY	SHED
HEATPUMP	VESSEL	COUPLED	HARVESTING	PLASTERER	SLATE	OLD	KERB	DUCT	TILING	CAVITY
GROUNDWORK	WIRING	COVERING	HAULAGE	PLASTIC	SLATTED	PIPEWORK	LED	ERECT	VENTILATION	DOMESTIC
RAIN	SHUTTERING	CRAFT	HEATER	PLU	SOIL	PIT	LIGHT	FOUNDATION	LINE	MAINTENANCE
SEWER	AGRI	CRUSHING	HERITAGE	PLUMBER	SOYA	SEWAGE	MACHINE	GARDEN	RESTORATION	HIRE
DRAIN	AGRIC	CUBICLE	HORSE	PLY	SPRAYER	AUTOMATION	MARKETING	IMPORTATION	DEVELOPER	ROOFING
FIBREGLAS	AUTHORITY	CURATIN	HORTICULTURAL	POLLUTION	SPRING	DRYWALLING	METERING	IMPORTING	ENVIRONMENT	WALL
RECLAMATION	DRYLINER	CUT	HOUSEBUILDING	POLYSTYRENE	STALLER	AIRTIGHT	MILL	LOW	LAND	
MEMBRANE	EXTENTSION	CYLINDER	HVAC	POLYURETHANE	STAT	ALTERATION	OPERATOR	MACHINERY	ACOUSTIC	
RECOVERY	FILL	DAM	IMAGING	POND	STATION	ALTERNATIVE	PARTITIONING	MAIN	CATTLE	
FILTRATION	HYDRO	DAMAGED	IMPROVEMENT	POOL	STEAM	APARTMENT	PARTNER	MECHANICAL	CONSULTANT	
UNDERGROUND	HYGIENE	DAMPROOFING	INFRASTRUCTURE	PORCH	STEELING	APPLICATION	PASSIVE	PIPING	CONTROL	
SEWERAGE	PRESERVATION	DATA	INSPECTION	POTENTIAL	STORM	AREA	PAVING	PREPARE	FIXING	
SLABBING	RUBBLE	DEALING	INST	POURING	STRAW	ASSESSMENT	PLASTERBOARD	PRINCIPAL	FLOORING	
LIGHTING	AEROBOARD	DECOMMISSION	INSTAFIBRE	POWERLINE	STYROFOAM	AUTHORITIE	PROCES	PUBLIC	KITCHEN	
SLAB	AGENT	DEMOLITIAN	INSTRUMENTATION	PRE	SU	BED	PROGRAM	SHOP	LAYING	
SOUND	AGRICULTURE	DENSITY	INTEGRATED	PRIMARY	SUBMERSIBLE	BOARD	PROP	SINGLE	MANUFACTURING	
BARRIER	AI	DESCRIPTION	INTERIOR	PRIOR	SUBSOIL	BONDED	PROPERTIE	SLATING	PAINTING	
FORESTRY	ANALYSI	DESIGNING	INTERPRETATION	PROCESSING	SUDIO	CABIN	PROTECTION	SPACE	PIPE	
BUSINESS ENERGY RATING	ANIMAL	DIG	IRRIGATION	PROD	SUN	CABLE	PROVIDER	SPECIALISE	SAVING	
DRYWALL	APPARTMENT	DISMANTLE	JCB	PROGRAMME	SUPPORT	CHP	PURCHASE	TANKING	SERVICING	
PELLET	ARCHITECTURAL	DOCK	KIT	PROMOTE	SURFACE	CLAY	RADIATOR	TESTING	SUSPENDED	
LANDFILL	ASBESTO	DOM	LAGOON	PUMPED	SURVEYING	CLEAN	RECYCLING	THATCHER	BUSINES	
GROUND	ATTENUATION	DOMESTUC	LANDLORD	PURIFICATION	SW	CLIENT	REFRIDGERATION	TRADE	DISTRIBUTION	
AGRICULTURAL	AUDIO	DRAUGHT	LANDSCAPING	QUOTE	SWEDEN	COAT	RELATED	WATERPROOFING	EXTERNAL	
GEOTHERMAL	AUDIT	DRAW	LARGE	RAFT	SWIMMING	COILLTE	REMOVAL	FILTER	FIRST	
SILAGE	AUTOMATIC	DREDGING	LASER	RAIL	SWITHGEAR	COMMISSIONING	RENTAL	GROUP	METER	
TURBINE	BALING	DRILLING	LEAKAGE	RAINWATER	SYPHONIC	CONSERVATION	REPAIR	PROOFING	PROJECT	
WASTE	BANDED	DRYLINING	LIFT	RAW	SYSTME	CONSERVATORY	RESERVOIR	DIGGER	PROVIDE	
DRYLINING	BASE	DUTIE	LINER	RD	TAPE	HYGIENIC	RETAINING	SPECIALIST	STRUCTURE	
PUMP	BASED	ECOBEAD	LIVE	RECLAIMATION	TAR	CUSTOMER	SAMPLE	SUPPLIE	SUB	
CEILING	BEDROOM	EFFICENT	LOGGING	RECLAMING	TARING	CUTTING	SECURITY	COLD	SUBCONTRACTOR	
ALUMINIUM	BEND	EFFLUENT	MAC	RECOVER	TECHNICALLY	DAIRY	SEPTIC	EXCAVATOR	SUPPLIER	
DRAINAGE	BIN	ELEMENT	MARBLE	REDUCING	TECHNICIAN	DECORATING	SERV	HOT	YARD	
WOOD	BIOFUEL	EMPTYING	MARINE	REED	TECHNOLOGY	DEVICE	SHEET	INTERNAL	CARPENTER	
LINING	BLOCKLAYING	EMULSION	MARKET	REFRIGERATED	TELEPHONE	DISPOSAL	SHIP	LABOUR	SCHEME	
DRY	BLOWN	ENGAGED	MAS	REG	TEO	DISTRIBUTOR	SKIRTING	MATERIAL	ERECTING	
TREATMENT	BODIE	ENGINE	MASTIC	REINFORCED	TEORANTA	DOE	SLOTTED	OPERATION	FARMER	
GLAZING	BOXE	ERRECTION	MATER	REMOVING	THATCHED	DRILL	SMART	PREMISE	HOUSING	
WIND	BREWERY	ESCAVATION	MEASURE	RENOVATING	THERMINAL	DRIVING	SOLID	PRODUCTION	OIL	
SITE	BRICKLAYER	EXPLORATION	METRE	REP	TIGHTNES	DUCTING	SPREADING	REFURBISHMENT	PRIVATE	
FARM	BRIDGE	EXTERIOR	MGE	RESERVOIRE	TILE	DUCTWORK	STAINLES	RETAIL	SECOND	
TIMBER	BURNER	EXTRACT	MGMT	RESLATING	TIPPER	DUMPER	STAIR	SAFETY	SLURRY	
HEATING	BUSINESSE	FACADE	MIDDLE	RESOURCE	TOOL	ECO	STORE	SHEETING	INDUSTRY	
PLUMBING	CALIBRATION	FACILITY	MILKING	RESPRAYING	TOWEE	EFFICIENCY	STRUCTURAL	SOLUTION	EQUIPMENT	
RENEWABLE	CAPPING	FACTOR	MINERAL	RETHATCHING	TRACK	ENERGIE	STRUCTURED	SPECIALISING	CLEARANCE	
FITTING	CARRIED	FARMYARD	MINI	ROADSTONE	TRAILER	ERECTOR	STUD	STORAGE	FABRICATION	
WATER	CCTV	FAST	MITIGATION	ROADWAY	TRANSFORME	ESB	SURVEY	TRACTOR	RATING	
SERVICE	CELING	FEE	MOBILE	ROADWORK	TRUCK	EXTRACTION	TACK	UNDERFLOOR	RESIDENTIAL	
CLADDING	CERTIFY	FELTING	MODULAR	ROCK	TUNNEL	FACTORIE	THATCH	WALLING	ROAD	
SYSTEM	CHAMBER	FEUL	MONITORING	ROCKWOOL	UNDER	FELT	TIMBERFRAME	ASSESSOR	STOVE	
UPVC	CLEANING	FILTERATION	MOVING	ROOD	UNDERTAKING	FIELD	TORCH	EXCAVATION	CONCRETE	
HOUSE	CLEARING	FILTERSCOOLER	NETWORK	ROOFER	UPGRADING	FILM	TRANSMISSION	RENEWAL	ROOF	
SOLAR	COATING	FIREPROOFING	ONLINE	SALVAGE	VISUAL	FIRE	UPGRADE	SLATED	BATHROOM	
CONTRACTOR	COILETE	FIREWOOD	OP	SANITARY	VOLTAIC	FIREPLACE	VALVE	GUTTER	MANUFACTURE	
CONSTRUCTION	COMBINED	FIX	OPERATE	SANITATION	WALK	FOREIGN	VOLTAGE	METAL	PLASTERING	
ENERGY	COMPLEX	FOOD	OPTION	SANITRY	WARDROBE	FOREST	WOODEN	AIR	HOME	
ATTIC	COMPLY	FORMWORK	OVERHEAD	SCANNING	WARE	FRAME	POWER	ASPECT	FLOOR	
BUILDING	COMPONENT	FREE	PARALON	SCRAP	WAREDROBE	FRIENDLY	RADON	BEAD	SMALL	
INSULATE	COMPOSITE	FUEL	PARLOUR	SCREENING	WAREHOUSE	FURNITURE	THATCHING	CRANE	ELECTRICAL	
DEMOLITION	CON	FUME	PATH	SECURING	WATERPROOF	GEO	THERMAL	DWELLING	BOILER	
DOOR	CONDITION	FUMIGATION	PERCULATION	SEI	WATERWAY	GEOLOGICAL	ACTIVITIE	EFFICIENT	FASCIA	
INSTALL	CONDITIONING	FUTURE	PERFORMANCE	SELL	WATERWORK	HANDLING	BIO	FARMING	SOFFIT	
TANK	CONSTRUCTED	GASIFICATION	PERIOD	SEPARATOR	WELDING	HANGING	CENTRAL	FOAM	STEEL	
WINDOW	CONTROLLING	GEOPHYSICAL	PHARMACEUTICAL	SERVER	WHEELIE	HEAVY	CLEAR	GARAGE	HEAT	

9. References

Blei, D., A.Y. Ng, and M. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. Available at: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed May 2016).

Blei, D. and J. Lafferty. 2006. "Dynamic Topic Models." Proceedings of the 23rd International Conference on Machine Learning, 113–120, Pittsburgh, Pennsylvania, U.S.A., June 25 −29, 2006. Doi: https://doi.org/10.1145/1143844.1143859.

Blei, D. and J. Lafferty. 2007. "A Correlated Topic Model of Science." *Annals of Applied Statistics* 1(1): 17–35. Doi: https://doi.org/10.1214/07-AOAS114.

Blei, D. and J. Lafferty. 2009. "Topic Models." Available at: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf (accessed April 2016).

BLS. 2011. "Green Goods and Services Survey." Available at: http://www.bls.gov/ggs/, BLS, USA (accessed May 2016).

Carpenter, B. 2010. "Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling." Available at: https://lingpipe.files.wordpress.com/2010/07/lda3.pdf (accessed May 2016).

Chang, J. and A. Dai. 2015. "'*Package-lda*': Collapsed Gibbs Sampling Methods for Topic Models." Available at: https://cran.r-project.org/web/packages/lda/lda.pdf (accessed May 2016).

Department of Environment. 2013. "Construction Activity Completion Statistics." Available at: http://www.housing.gov.ie/housing/statistics/house-building-and-private-rented/construction-activity-completions, Ireland (accessed April 2016).

Eurostat. 2015. "A Practical Guide for the Compilation of Environmental Goods and Services (EGSS) Accounts." Unit E2, Eurostat, Luxembourg. Doi: https://doi.org/10.2785/688181.

EU-691. 2011. "EU REGULATION (EU) No 691/2011 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 July 2011 on European environmental economic accounts." Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32011R0691 (accessed 2013).

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

Griffiths, T. and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Science*, USA, 101, 5228–5235.

Heinrich, G. 2009. "Parameter Estimation for Text Analysis." Technical Report Fraunhofer IGD, Darmstadt, Germany. Available at: http://www.arbylon.net/publications/text-est2.pdf (accessed April 2016).

Hornik, K. and B. Grün. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40(13): 1–30. Doi: https://doi.org/10.18637/jss.v040.i13.

McCallum, A. 2002. "MALLET: A Machine Learning for Language Toolkit." Available at: http://mallet.cs.umass.edu (access May 2016).

Minka, T. and J. Lafferty. 2002. "Expectation-propagation for the Generative Aspect Model." Proceedings of the Eighteenth Conference on Uncertainty in Artificial

Intelligence, 352–359, Alberta, Canada, August 1–4, 2002. Available at: https://dl.acm.org/citation.cfm?id=2073918 (accessed April 2016).

OECD. 1999. "THE ENVIRONMENTAL GOODS AND SERVICES INDUSTRY Manual for Data Collection and Analysis." OECD, Paris. Available at: https://unstats.un.org/unsd/envaccounting/ceea/archive/EPEA/EnvIndustry_Manual_for_data_collection.PDF (accessed May 2016).

ONS. 2015. "UK Environmental Goods and Services Sector (EGSS): 2010–2012." Available at: http://www.ons.gov.uk/economy/environmentalaccounts/bulletins/ukenvironmentalaccounts/2015-04-15 (accessed October 2014).

Ramage, D., D. Hall, R. Nallapati, and C.D. Manning. 2009. "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248–256, Singapore, August 6–7, 2009. Available at: https://www.aclweb.org/anthology/D09-1026 (accessed May 2016).

RICS. 2016. "The Real Cost of New House Delivery, Royal Institute of Charter Surveyors." Dublin, Ireland. Available at: https://www.scsi.ie/documents/get_lob?id=885&field=file (accessed May 2016).

SEAI. 2013. "Sustainable Energy Authority of Ireland – Annual Report 2013." Available at: https://www.seai.ie/Publications/SEAI_Corporate_Publications_/Annual_Reports/SEAI-Annual-Report-2013.pdf (accessed May 2016).

Spärck Jones, K. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation* 28: 11–21. Doi: https://doi.org/10.1.1.115.8343.

Statistics Estonia. 2015. "Development of the Methodology for the Compilation of Statistics of the Environmental Goods and Services Sector (EGSS) in Estonia." Statistics Estonia.

Wallach, H.M., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. "Evaluation Methods for Topic Models." Proceedings of the 26-th International Conference on Machine Learning", 1105–1112, Montreal, Canada, June 14–18, 2009.

Wang, Y. 2008. "Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details." Available at: https://cxwangyi.files.wordpress.com/2012/01/llt.pdf (accessed May 2016).

Received July 2016
Revised September 2017
Accepted December 2017

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 653–681, http://dx.doi.org/10.2478/JOS-2019-0027

Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach

Joseph W. Sakshaug[1], *Arkadiusz Wiśniowski*[2], *Diego Andres Perez Ruiz*[2], *and Annelies G. Blom*[3]

Carefully designed probability-based sample surveys can be prohibitively expensive to conduct. As such, many survey organizations have shifted away from using expensive probability samples in favor of less expensive, but possibly less accurate, nonprobability web samples. However, their lower costs and abundant availability make them a potentially useful supplement to traditional probability-based samples. We examine this notion by proposing a method of supplementing small probability samples with nonprobability samples using Bayesian inference. We consider two semi-conjugate informative prior distributions for linear regression coefficients based on nonprobability samples, one accounting for the distance between maximum likelihood coefficients derived from parallel probability and nonprobability samples, and the second depending on the variability and size of the nonprobability sample. The method is evaluated in comparison with a reference prior through simulations and a real-data application involving multiple probability and nonprobability surveys fielded simultaneously using the same questionnaire. We show that the method reduces the variance and mean-squared error (MSE) of coefficient estimates and model-based predictions relative to probability-only samples. Using actual and assumed cost data we also show that the method can yield substantial cost savings (up to 55%) for a fixed MSE.

Key words: Bayesian inference; quota sampling; German Internet Panel; GESIS Panel; web surveys.

1. Introduction

1.1. Background

Scientific surveys based on random, probability-based samples are ubiquitously used in the social sciences to study and describe large populations. They provide a critical source of quantifiable information used by governments and policy-makers to make informed decisions. However, probability-based surveys are increasingly expensive to carry out due to declining response rates and costly intervention strategies (Tourangeau and Plewes 2013). Consequently, many survey organizations have shifted away from probability-based samples in favor of cheaper nonprobability samples usually drawn from volunteer web panels. This shift in practice has prompted significant controversy and skepticism

[1] University of Mannheim and Institute for Employment Research, Nuremberg, 90478 Germany. Email: joe.sakshaug@iab.de
[2] University of Manchester, Manchester, M13 9PL United Kingdom. Emails: a.wisniowski@manchester.ac.uk, and diego.perezruiz@manchester.ac.uk
[3] School of Social Sciences, University of Mannheim, Mannheim, 68131 Germany. Email: blom@uni-mannheim.de

over the representativeness and overall utility of nonprobability samples (Baker et al. 2013). While probability-based surveys have their own concerns regarding representativeness (Gelman et al. 2016; Wang et al. 2015), comparison studies generally show (with same exceptions: see, for example Kennedy et al. 2016) that they produce more accurate population estimates than nonprobability surveys when evaluated against benchmark data (Yeager et al. 2011; Blom et al. 2017; Malhotra and Krosnick 2007; Chang and Krosnick 2009; Dutwin and Buskirk 2017; Pennay et al. 2018; Erens et al. 2014; Callegaro et al. 2014; MacInnis et al. 2018). Hence, the field of survey research is in a situation where probability-based samples are preferred from an error perspective, while nonprobability samples are preferred from a cost perspective.

Given the advantages of both sampling schemes, it makes sense to devise a strategy to combine them in a way that is beneficial from both a cost and error perspective. In some ways, survey organizations already attempt to make use of both sample types, either by drawing a nonprobability sample whose units closely match units from a reference probability sample prior to data collection (Rivers 2007; Rivers and Bailey 2009; Ansolabehere and Rivers 2013), or by devising post-survey weights that adjust the composition of a nonprobability sample survey towards that of a reference probability survey (Lee 2006; Lee and Valliant 2009; Valliant and Dever 2011). While both approaches are cost-effective and have been shown to increase the accuracy of estimates derived from nonprobability surveys, they have some important limitations. Firstly, they assume that the matching/adjustment variables fully explain the underlying selection mechanism that leads to inclusion in the nonprobability sample – a questionable and usually untestable assumption in practice (Mercer et al. 2017). Secondly, the target variable of interest is usually not present in the reference probability survey data, and therefore, these data are usually discarded after the matching/adjustment procedure. The intended analysis is then based solely on the nonprobability survey data, which lacks important properties of randomization theory, including the ability to measure the uncertainty of sample-based estimates.

Instead of forgoing probability-based survey data collection entirely, an alternative approach is to field the same questionnaire in a parallel probability and nonprobability sample and analyze the collected data jointly. For example, Elliott and Haviland (2007) describe a methodology that supplements a traditional probability sample with a web-based convenience sample. They evaluate a composite estimator influenced by Rao (2003) that is a linear combination of a probability and convenience sample, with each sample weighted according to a bias function. The estimator, under certain conditions, yields a smaller mean-squared error (MSE) compared to the probability-only sample. In related work, Elliott (2013) proposes a method of devising pseudo-weights for a nonprobability sample based on probabilities of selection estimated using a parallel probability sample. Both samples can then be combined and analyzed with case weights as if the units were drawn from the same population frame. The method is shown to reduce bias and MSE relative to a probability-only sample.

DiSogra et al. (2012) introduce an idea referred to as "blended calibration" in which available probability sample cases are supplemented with parallel nonprobability opt-in panel cases. The two-step procedure relies on, firstly, weighting the probability sample to known population benchmarks using a raking or poststratification procedure. In the second

step, the weighted probability and unweighted opt-in cases are combined and the combined sample is calibrated to the probability-only sample on a selection of survey variables common to both samples. The method yields smaller bias and MSE compared to more traditional approaches of analyzing probability and nonprobability samples separately and jointly. Fahimi et al. (2014) extend the approach by considering a more effective range of differentiator variables to use in the calibration step.

A practical limitation of the above studies is that they require relatively large probability sample sizes. Elliott and Haviland (2007) recommend a probability sample size of at least 1,000–10,000 cases alongside a convenience sample size in the thousands, and Elliott (2013) uses a probability sample size of 50,000 in the simulation study. Blended calibration also requires a relatively large probability sample size in order to minimize the variability in the probability-based survey benchmarks.

Any data integration strategy that requires fielding a large probability sample is likely to be met with opposition, as such sample sizes are prohibitively expensive for most survey budgets. An alternative, and more budget-friendly, strategy is to draw and field a small probability sample and combine it with a parallel nonprobability sample. On the face of it, the usefulness of deliberately fielding a small probability sample is not intuitively clear. Estimates derived from small probability samples, while inferentially valid, are subject to large variability and are insufficient as a standalone source of population information. Furthermore, a small probability sample is too sparse to be used as a reference sample for sample matching and post-survey adjustment procedures. A natural question, therefore, is whether there exists any scenario in which combining a small probability sample with a nonprobability sample could be beneficial from both a cost and error perspective.

1.2. Bayesian Inference

We address this question from a Bayesian inferential viewpoint. Bayesian inference offers an attractive system of estimation that allows combining sparse scientific data, such as those from probability-based samples, with less scientific and less reliable but potentially abundant and cheap information, such as those derived from nonprobability sources (Gelman et al. 2013). There are several advantages of using Bayesian inference in the context of combining small probability samples with nonprobability samples. First, the Bayesian framework allows for estimating complex models and quantifying measures of uncertainty, which can be problematic when analyzing nonprobability data under traditional estimation frameworks. Second, unlike sample matching and post-survey adjustment procedures, the Bayesian framework allows for the analysis of probability-based sample units through the likelihood function and is principally structured to give priority to these units in the posterior estimations as the probability sample size increases. Put differently, as additional probability sample units are observed, the "prior" information brought in through the nonprobability data becomes less relevant in the estimations, and increasing weight is given to the probability units. And third, because the probability-based likelihood *borrows information* from the informative nonprobability-based prior, the resulting posterior estimates are expected to be more efficient, that is, have less uncertainty, compared to estimates derived from small probability-only samples. This result could yield potential cost savings if large reductions in uncertainty are achieved and

the marginal cost of interviewing a nonprobability sample unit is lower than that of a probability sample unit – a plausible scenario in practice.

However, a disadvantage of applying the Bayesian framework in the aforementioned context is the deliberate incorporation of (potentially) biased data into the estimation process. In contrast to sample matching and post-survey adjustment, which takes an error-prone nonprobability sample and skews it towards a presumably less error-prone probability reference sample, the Bayesian approach that we describe does the opposite. That is, the method takes a probability sample and deliberately skews it towards a nonprobability sample reflected in the prior. The posterior estimates are therefore likely to have more bias than corresponding probability-only estimates. This effect is likely to be most pronounced for small probability samples where the prior will have peak influence on the posterior estimates. On the other hand, the expected reduction in variability due to the supplementary use of nonprobability data may offset any increase in bias, resulting in an estimator that yields a smaller mean-squared error.

1.3. Research Aims

In this article, we investigate whether supplementing a probability sample with nonprobability sample priors can produce more efficient survey estimates under varying probability sample sizes. We consider three specifications of the prior distribution for a target analysis of regression coefficients and model-based predictions: (i) a reference prior that allows for the probability sample to dominate the posterior, (ii) an informative prior that decreases the weight of the nonprobability sample with increasing distance between the maximum likelihood coefficient estimates derived from the probability and nonprobability samples, thus, "protecting" against bias in the latter, and (iii) an informative prior whose weight depends on the variability and size of the nonprobability sample and is able to dominate the posterior. Further, we examine the extent to which varying levels of bias in the nonprobability sample affect the mean-squared error (MSE) of the posterior estimates. To achieve these aims, we carry out a simulation study and real-data application involving two nationally-representative, probability-based surveys and eight nonprobability web surveys fielded in parallel using the same questionnaire. Through the application, we also assess whether the method is likely to yield cost savings for a fixed MSE.

The balance of this article is organized into five sections. Section 2 describes the proposed methodology for combining probability and nonprobability samples under a Bayesian framework. Section 3 presents the simulation study examining the bias-variance tradeoff of the method for various bias and sample size parameters. Sections 4 and 5 describe the real-data application and evaluation. Lastly, Section 6 provides a general discussion of the results, their implications for survey practice, and possible research extensions.

2. Methodology

2.1. Modeling Approach

As introduced in Subsection 1.2, in Bayesian inference (for details, see Gelman et al. 2013), the likelihood distribution is multiplied by a prior distribution, and inferences are

typically summarized by random draws from this product, that is, the posterior distribution. On the one hand, Bayesian inference can utilize prior distributions that "allow data to speak for themselves," that is, to have a negligible influence on the posterior draws. These priors are known as noninformative or weakly informative. On the other hand, informative priors can be used to add information about model parameters. This may be desirable in situations where parameters cannot be identified, or due to a small number of available observations. In this section we present three models, of which *two* use informative prior distributions constructed from a *single* nonprobability sample.

Consider an $n \times 1$ response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ of observations collected from a probability-based survey. The parameter of interest is the expectation of $\mathbf{y}$, denoted by μ. We assume that y is continuous and normally distributed:

$$\mathbf{y} \sim N(\mu, \sigma^2),$$

where σ^2 is the unknown variance of $\mathbf{y}$. The simple model can be expanded to account for covariates if the researcher's substantive interest lies in interpreting their effect on the outcome variable, or in making model-based predictions of the outcome. We focus on these two scenarios. The covariates can be incorporated by using a linear regression with an $n \times p$ design matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, which leads to

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a column vector of length p and I is the $n \times n$ identity matrix. We note that this model does not explicitly reflect the survey design. In our forthcoming application, we include survey weights as a covariate in the proposed modeling approach. Adapting the proposed approach to include additional survey design features (e.g., stratification, clustering) is a topic we leave for future work.

A semi-conjugate prior distribution for a single regression coefficient, β_j, for $j = 1, \ldots, p$ is

$$\beta_j \sim N\left(\beta_{j0}, \sigma^2_{\beta_j 0}\right), \tag{1}$$

with fixed location and variance hyperparameters, β_{j0} and $\sigma^2_{\beta_j 0}$, respectively. The semi-conjugacy (or conditional conjugacy) results from the fact that the variance in (1) does not depend on σ^2 (Gelman et al. 2013, 130). We consider three specifications of these hyperparameters.

In **Model 1** we assume a weakly informative parameterization of the priors, that is;

$$\beta_{j0} = 0, \quad \sigma^2_{\beta_j 0} = 10^6.$$

This specification allows the model parameters to be estimated directly from the probability data. Therefore, we treat this model as a reference to compare the two other specifications in which we introduce information from the nonprobability samples.

In **Model 2** we introduce an informative prior by utilizing information from a single nonprobability sample. First, we define $\hat{\beta}_j^P$ and $\hat{\beta}_j^{NP}$ to be the maximum likelihood (ML) estimators of the regression coefficients using the probability (*P*) and nonprobability (*NP*) survey data, respectively. These ML estimates are equivalent to the means of the posterior distributions of these parameters under the linear regression model using

noninformative Jeffrey's priors. We implicitly assume a simple random sampling design for the nonprobability data. Next, we set the hyperparameter β_{j0} equal to the estimated regression coefficient derived from the nonprobability survey, $\hat{\beta}_j^{NP}$. For the variance hyperparameter $\sigma_{\beta_j 0}^2$ we consider the Euclidean distance between the ML regression coefficients estimated in the probability survey and in the nonprobability survey. Specifically, we consider the square of the distance as the variance hyperparameter in (1):

$$\sigma_{\beta_j 0}^2 = d^2\left(\hat{\beta}_j^P, \hat{\beta}_j^{NP}\right) = \left(\hat{\beta}_j^P - \hat{\beta}_j^{NP}\right)^2, \quad \forall j.$$

Therefore, the prior for the regression coefficient in Model 2 can be written as:

$$\beta_j \sim N\left(\hat{\beta}_j^{NP}, d^2\left(\hat{\beta}_j^P, \hat{\beta}_j^{NP}\right)\right) \tag{2}$$

This method of setting the hyperparameter for the regression coefficient implies that the standard deviation, $\sigma_{\beta_j 0}$, is equal to the difference between the probability- and nonprobability-based ML estimates and does not depend on the size of the nonprobability sample. This, on the one hand, ensures some variability around the mean while keeping the uncertainty relatively small. If the distance d is large, the prior is wider and allows the small probability sample to influence the posterior. The smaller the distance between the two ML estimates, the tighter the prior distribution and, thus, larger potential gains in reducing posterior variance. A potential limitation of this approach is that if the distance is zero, that is, the corresponding probability and nonprobability estimates are equal, then the hyperparameter will be set to zero and shrink the location parameter β_j to a fixed value being $\hat{\beta}_j^{NP}$. However, in practice, such an event has virtually zero probability.When the distance is extremely small, it may severely reduce the variance of the posterior distribution for the parameter, especially when the probability sample size is very small. The next model we consider is free from this shortcoming.

By using the probability-based estimator to construct the prior distribution, the question of using data twice arises. We address this issue by pointing out that the ML estimator from the probability sample (a measure of central tendency) is used to inform the variance, rather than the mean. Further, we use the information from the probability data only in relative comparison to the nonprobability sample. Hence, any potential shrinkage in posterior variance depends on the combination of both data sets, rather than the probability data alone.

In **Model 3** we use a bootstrap procedure instead of the squared distance to derive information about the variance hyperparameter in (1). The bootstrap method has been used in many contexts and was originally proposed by Efron (1979). The general approach is to draw random subsamples with replacement from the full sample a large number of times and estimate the statistic of interest in each subsample before combining them using a bootstrap estimator. We implement the procedure by drawing 1,000 bootstrap samples from the nonprobability survey data, estimating the regression coefficient in each of them, and then calculating the variance $\left(\hat{\sigma}_{\beta_j}^{BNP}\right)^2$ across all regression coefficients. We then set the variance hyperparameter in (1) to the estimated variance and the prior distribution for

the regression coefficient in **Model 3**:

$$\beta_j \sim N\left(\hat{\beta}_j^{NP}, \left(\hat{\sigma}_{\beta_j}^{BNP}\right)^2\right), \tag{3}$$

with mean being the ML coefficient calculated using the nonprobability sample (the mean of all bootstrapped coefficients will converge to it). This approach is an alternative to calculating the uncertainty around the nonprobability-based regression coefficient and ensures it is always positive. The method is limited in a sense that the hyperparameter relies on the bootstrapped nonprobability sample which may propagate its unrepresentativeness and homogeneity, especially when very large nonprobability sample sizes are used, again leading to a false sense of certainty about the regression coefficient. However, analogous to the distance approach, this effect is reduced the larger the size of the probability sample.

For the variance of the regression model, σ^2, we first transform it to a precision, that is, inverse variance (σ^{-2}), and set $\sigma^{-2} \sim \Gamma(r, m)$ where $\Gamma(\cdot, \cdot)$ denotes a Gamma distribution with hyperparameters r being a shape and m being a rate. In our application, we set these hyperparameters to be $r = m = 10^{-3}$. This specification for the precision parameter is approximately noninformative and gives preference to the data (Gelman et al. 2013, 128). It remains the same for Models 1 through 3, which ensures comparability of the results.

3. Simulated Data Inference

In this section, we demonstrate how the proposed methods work under various assumptions regarding bias and sample size introduced through simulated data. First, we investigate the effect of bias on the regression coefficients of the model (part A of the simulation), and second, we analyze to what extent the bias affects model-based predictions of the outcome variable (part B).

The analysis was implemented in OpenBUGS (Spiegelhalter et al. 2007) and R (R Core Team 2016) using the library *r2openbugs* (Sturtz et al. 2005). We also use *MCMCpack* to summarize the results of the simulations, *boot* package for bootstrapping, as well as *ggmcmc* and *lattice* packages for visualization. In the simulations, the posterior distributions were obtained using three MCMC chains with samples of 2,000 each and 500 burn-in samples which ensured convergence of all chains.

To generate the data, we first assume the true values of the parameters in a linear regression model with intercept $\beta_1 = 5$, two parameters $\beta_2 = 0.5$ and $\beta_3 = 1$, and standard deviation of the outcome being $\sigma_y = 5$. Predictors x_1 and x_2 have means 0 and 5, respectively, standard deviations 4 and 0.5, and are correlated with correlation $\rho = 0.1$. These assumptions yield the mean response being $\bar{y} = 10$.

To introduce bias, we multiply the true parameter $\beta_3 = 1$ by 0.5, 1 (i.e., unbiased sample), 1.5, 2, 2.5, and 3 when generating the nonprobability samples (part A of simulation). For testing the effect of bias in nonprobability samples on the predicted outcomes (part B), we generate a predictive posterior distribution for a fixed probability test sample of size 500 using coefficients generated in part A. Bias introduced in this way is quite significant. For instance, when coefficient β_3 is doubled, the expected outcome

increases to 15. These scenarios are relatively extreme to real-life applications, but aim to demonstrate the limits of the proposed methods.

In the simulation we assume three nonprobability sample sizes $NPS \in \{1000, 10000, 50000\}$ and probability sample sizes $PS \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000\}$. In each simulation, for each PS we generate 100 sets of data with each combination of bias level and NPS. In total, it yields 27,000 data sets.

3.1. Model Evaluation

First, we evaluate the performance of the three modelling approaches by calculating the bias, variance, and mean-squared error (MSE; the sum of variance and squared bias) of the posterior means of the coefficients estimated using Models 1, 2 and 3. This permits an assessment of the effect of bias in nonprobability-based informative priors on all of the model coefficients.

Second, to evaluate model-based predictions, we split the probability survey data randomly into two parts: a training set (denoted by $\mathbf{y}$) and a test set ($\tilde{\mathbf{y}}$). We then use the training set to fit the models specified in Subsection 2.1. Next, we predict the outcome variable in the test set $\tilde{\mathbf{y}}$. We do so by applying posterior distributions of model parameters estimated using $\mathbf{y}$ to the covariates in the test set. The resulting distributions are called posterior predictive distributions, that is, posteriors for each data point.

Next, to evaluate the error properties of the predictions for the three models, we calculate the bias, variance, and MSE of the means, denoted by $\bar{\tilde{\mathbf{y}}}$, of the posterior predictive distributions for $\tilde{\mathbf{y}}$. In the simulation, we define the MSE as:

$$MSE(\bar{\tilde{\mathbf{y}}}) = \mathbb{E}\left[(\bar{\tilde{\mathbf{y}}} - \tilde{\mathbf{y}})^2\right],$$

which can be decomposed into variance and bias $MSE(\bar{\tilde{\mathbf{y}}}) = Bias^2(\bar{\tilde{\mathbf{y}}}) + Var(\bar{\tilde{\mathbf{y}}})$.

We compute the bias as the difference between the mean of the posterior means, $\bar{\tilde{\mathbf{y}}}$, and the mean of the test sample outcome $\tilde{\mathbf{y}}$, i.e., $Bias(\bar{\tilde{\mathbf{y}}}) = \frac{1}{n}\sum \bar{\tilde{\mathbf{y}}} - \frac{1}{n}\Sigma\tilde{\mathbf{y}}$ whereas $Var(\bar{\tilde{\mathbf{y}}})$ is the unbiased estimator of the variance of $\bar{\tilde{\mathbf{y}}}$.

We calculate the bias, variance, and MSE of the posterior predictive means for the three models described in Subsection 2.1 under different probability sample size scenarios. To accomplish this, we run the models on training sets ranging in size from 50 to 600 cases with intervals of 50, and from 600 to 1,000 with intervals of 100. The samples are constructed cumulatively so that the same cases used in the smaller samples are also included in the larger samples.

3.2. Results

Having generated the artificial probability and nonprobability samples for each size and level of bias as described in the previous section, we applied the three modelling approaches (Model 1, 2, and 3) as described in Subsection 2.1 to produce posterior distributions of model parameters and predictive distributions for the test sample in simulation part B. We then compare the effect of bias introduced in the nonprobability sample on bias, variance, and MSE of the coefficients and means of the posterior

predictive distributions as defined in Subsection 3.1. The bias, variance, and MSE are averaged over 100 simulated data sets.

3.2.1. Part A: Regression Coefficients

Figure 1 presents the bias, variance, and MSE for the three coefficients, where β_3 has been generated with bias in the nonprobability (*NP*) sample. First, we observe that Model 2 does not lead to bias in the coefficients and performs similarly to Model 1, which relies on weakly-informative priors without information from the *NP* samples. It also leads to improvements in variance (middle panel of Figure 1) and MSE (lower panel). For Model 3, we observe larger improvements in variance compared to Models 1 and 2. However, in the presence of bias, the MSE tends to be dominated by it. This results from the fact that the

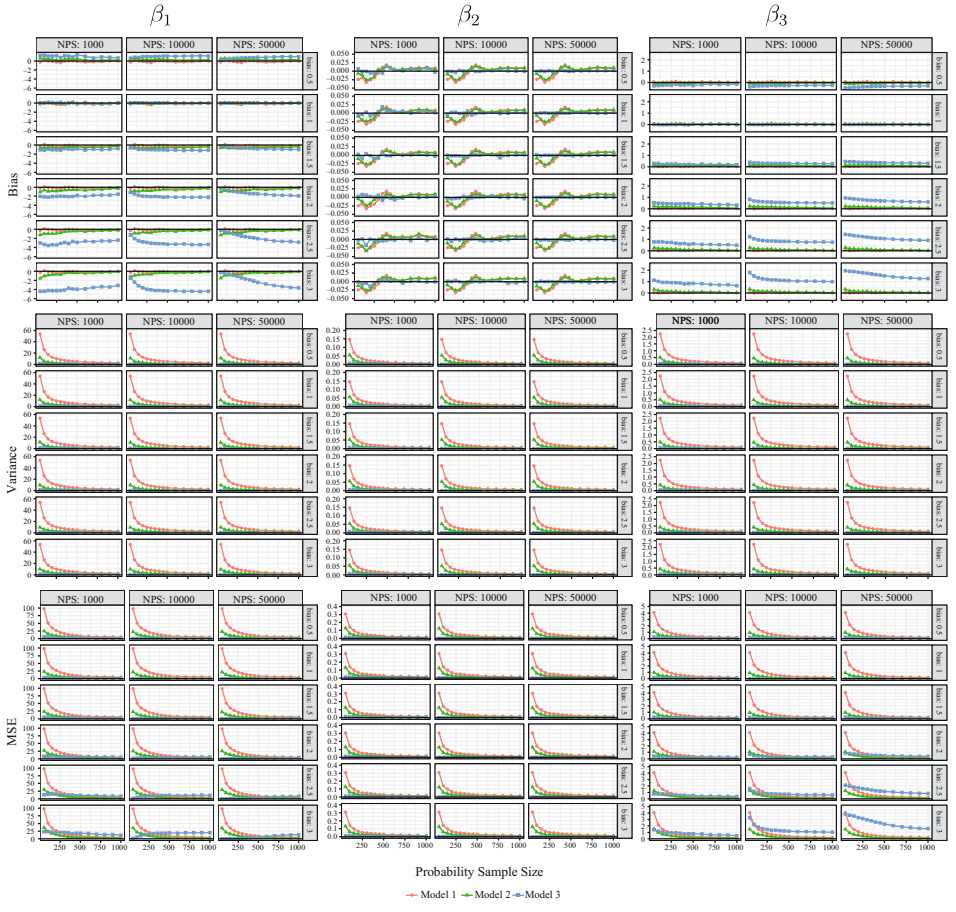

Fig. 1. Effect of bias in nonprobability samples on regression coefficient. Note: Regression parameters are in columns; measures in rows, where top row is bias (difference between the posterior mean from the model and the true coefficient), middle row is variance, and bottom row is mean-squared error (MSE), averaged over 100 simulations. Each panel shows the combination of three nonprobability sample sizes (NPS) and six levels of bias introduced to β_3 in nonprobability sample (bias:1 denotes unbiased coefficient, i.e., $\beta_3 \times 1$), with probability sample size on the x-axis.

prior in Model 3 relies on the size and variability of the *NP* sample and does not protect against bias present in it.

More precisely, in Model 3 the positive bias in the posterior mean of β_3 (top right panel) is increasing with the introduced bias (difference between posterior mean of the coefficient and the true coefficient) and it is more persistent with larger nonprobability sample sizes (*NPS*). This is offset by the negative bias in the intercept β_1 as the regression equation needs to be consistent with the expectation of the outcome in the probability sample, $E[y] = 10$. However, for large *NPS* (10,000 or 50,000), the prior for β_1 is relatively tight and it dominates the posterior of β_1 for small probability sample sizes (*PS*), which subsequently leads to bias in the predictions of the outcome (see Figure 2 and the following Subsection 3.2.2). With an increase in *PS*, the posterior becomes more and more dominated by the unbiased probability sample, which first increases the bias in the posterior of β_1 and decreases in β_3 (e.g., $NPS = 10{,}000$ and $bias = 3$ in top left and right panel of Figure 1) to gradually decrease bias in both coefficients (e.g., $NPS = 1{,}000$ and $bias = 2.5$) and output predictions (left panel in Figure 2). Coefficient β_2 remains unaffected by bias.

3.2.2. Part B: Model-Based Predictions

Figure 2 shows the effect of bias introduced in the nonprobability samples on the predictive ability of the models when priors are based on those samples. We average over means of posterior predictive distributions (referred to as predictions for brevity) for 500 generated outcome data points. In all comparisons, we utilize the true generated outcome.

In Figure 2 we observe that Model 2, compared with the weakly informative Model 1 without input from nonprobability samples, yields mostly unbiased predictions. For Model 3, as indicated in the previous section, the bias in predictions changes with the size of bias in β_3. A large bias in the coefficient yields larger prediction bias, larger variance, and larger MSE. Also, for larger nonprobability sample sizes (*NPS*), the bias persists for larger probability sample sizes (*PS*). However, for a moderate bias (β_3 multiplied by 0.5 to 1.5), Model 2 and Model 3 show a reduction in the prediction variance and MSE (presented on log scale) compared with Model 1 and for nonprobability sample sizes of 1,000 and 10,000.

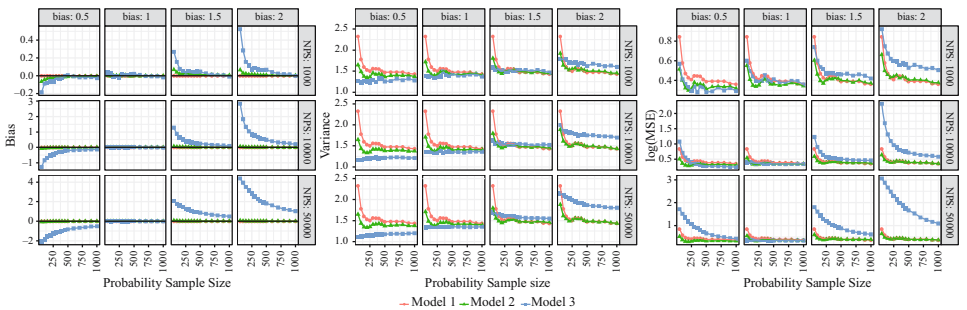

Fig. 2. Effect of bias in nonprobabiliy samples on predicted outcome. Note: Left panel shows bias (difference between the average predicted outcome and the average true generated outcome), middle panel shows variance, and right panel shows the logarithm of mean-squared error (MSE), averaged over 100 simulations. Each panel shows a combination of four levels of bias in β_3 (Beta:1 denotes no bias, i.e., $\beta_3 \times 1$) and three nonprobability sample sizes (NPS), with probability sample size on the x-axis.

For $NPS = 50{,}000$ and larger amounts of bias, reductions in variance and MSE are observed only for Model 2 and they are relatively small compared with Model 1 predictions.

4. Real-Data Application

To demonstrate the proposed methods on actual survey data, we make use of two probability-based surveys: the German Internet Panel (GIP) and the GESIS Panel, and eight nonprobability surveys. Each survey implemented the same questionnaire to respondents during overlapping field periods. Relevant details of the surveys are provided below.

We demonstrate the proposed Bayesian method on two continuous outcome measures: an additive index of a subset of Big Five (BIG-5; Digman 1990; Goldberg 1993) personality items and an additive index of a subset of Need for Cognition (NFC; Cacioppo and Petty 1982) scale items. The Big Five index included four items related to "trust of people", "artistic interests", "finding fault in others", and having an "active imagination" with each item using a 5-point response scale from "strongly disagree" to "strongly agree." The distribution of additive values approximately followed a normal distribution. The NFC index included four items about "knowing answers without understanding their rationale", "being confronted with tricky tasks to solve", "preferring to solve complex to simple problems", and "thinking only because one has to." Each item used a 7-point response scale from "strongly disagree" to "strongly agree." A square-root transformation was applied to the index to achieve approximate normality.

4.1. *German Internet Panel*

The GIP is an ongoing individual-level longitudinal online survey, which is designed to be representative of the population aged 16–75 in Germany. It is the central data collection project of the Collaborative Research Center 884 "Political Economy of Reforms" funded by the German Research Foundation (DFG). In 2012 and 2014, the GIP recruited sample members by means of a 3-stage stratified probability area sample and face-to-face recruitment interviews. At the first sampling stage, a random sample of areas was drawn from a database of 52,947 areas in Germany, each containing approximately equal numbers of households. Within each PSU, listers recorded every household along a predefined random route. Subsequently, a random sample of households to be interviewed drawn. All age-eligible members of sampled households were invited to become online panelists (Blom et al. 2015). The GIP covers individuals without computer and/or internet access by equipping them with the necessary devices (Blom et al. 2016a; Herzing and Blom 2019). The first recruitment process, which took place in 2012, yielded a recruitment rate of 18.5% (also based on Response Rate 2; AAPOR 2016) and in the second recruitment process in 2014 a recruitment rate of 20.5% (also based on AAPOR Response Rate 2) was achieved. Every two months, all panel members are invited to take part in an online survey of about 20–25 minutes on various social, economic, and political topics. The questionnaire module used in the present study was implemented 1–31 March 2015. Out of 4,989 original panel members, 3,426 completed the survey for a completion rate of 68.7%. Despite the low recruitment rate, the representativeness of the GIP compares well to other probability-based surveys in Germany (Blom et al. 2017).

4.2. GESIS Panel

Like the GIP, the GESIS Panel is an ongoing individual-level probability-based longitudinal survey. It is designed to be representative of the German-speaking population aged between 18 and 70 years, permanently residing in Germany. The sample was drawn from municipal population registers using a stratified multistage sampling procedure. All sample members who were interviewed with face-to-face recruitment interviews were asked to participate in the panel. The recruitment process, which took place in 2013/14, yielded a panel registration rate of 28.4% (based on Response Rate 1; AAPOR, 2016). Subsequent interviews are conducted on a bi-monthly basis using a mix of mail and web data collection. Mail questionnaires are sent to participants who are unable or unwilling to complete the interview online. Interviews are divided into two parts: a 15-minute interview on modules submitted by external researchers and a five-minute interview devoted to longitudinal core study topics developed by GESIS. The core study covers a range of topics, including values, political behavior, well-being, and usage of information technology. The questionnaire module we use was approved by the GESIS Panel team and fielded 18 February–14 April 2015. Out of 6,210 original panel members, 3,822 completed the interview (61.5%). More details of the GESIS Panel methodology can be found in Bosnjak et al. (2017), where they show the representativeness of the panel to be similar to other probability-based surveys in Germany (see also Blom et al. 2016b).

4.3. Nonprobability Surveys

The eight nonprobability web surveys were conducted by different commercial vendors. The vendors were recruited through a call for tender published in November 2014. The tender call sought to implement a ten-minute questionnaire on a sample of approximately 1,000 respondents in three waves of data collection. Initial data collection was to take place in March 2015 with two follow-up surveys in September 2015 and March 2016. The primary stipulation was that the sample should be representative of the general population aged 18–70 years living in Germany. Exactly how representativeness was to be achieved (e.g., quota sampling) was left to the discretion of each vendor. Out of 17 bids, seven commercial vendors were commissioned based on technical requirements and budgetary considerations. An eighth commercial vendor, upon learning about the study goals of the project, voluntarily offered to participate without compensation. Further details of each nonprobability survey, including cost information, is provided in Table 1. To maintain confidentiality, we do not identify the commercial vendors by name and simply refer to the nonprobability surveys by number, that is, Survey 1, Survey 2, and so on. The actual cost of the commercial surveys (excluding the gratis survey) ranged from EUR 5,392.97 to EUR 10,676.44. The average cost per respondent therefore ranged from EUR 5.40 to EUR 10.29. We do not have cost information for the GIP and GESIS Panel surveys.

4.4. Comparison of Outcome Variables Between Surveys

Here, we examine the extent to which the outcome variables differ within and between the probability and nonprobability surveys. Figure 3 displays estimated means and 95%

Table 1. List of probability and nonprobability surveys.

Survey	No. respondents	Quota variables	Fieldwork period	Total cost (in Euros)	Average cost per respondent (in Euros)
GIP	3,426	N/A	1st–31st March 2015	Unavailable	Unavailable
GESIS	3,822	N/A	18th February–14th April 2015	Unavailable	Unavailable
1	1,012	Age, gender, region, education	1st–31st March 2015	0 (pro bono)	N/A
2	1,000	Age, gender, region	5th–18th March 2015	5,392.97	5.40
3	999	Age, gender, region	2nd–11th March 2015	5,618.57	5.63
4	1,000	Age, region	1st–18th March 2015	7,061.11	7.07
5	994	Age, gender, region	2nd–16th March 2015	7,411.00	7.46
6	1,002	Age, gender, region, education	25th March–1st April 2015	7,636.22	7.62
7	1,000	Age, gender, region	3rd–9th March 2015	8,380.46	8.39
8	1,038	Age, gender, region	5th–11th March 2015	10,676.44	10.29

confidence intervals (CIs) for the BIG-5 (left panel) and NFC (right panel) outcome variables in the GIP and GESIS Panel surveys.

The figures show very little difference between the GIP and GESIS Panel estimates of BIG-5 and NFC. Both probability surveys yield mean estimates that overlap by their respective confidence intervals. Larger differences are apparent between the probability and nonprobability surveys. For the BIG-5 variable, all nonprobability surveys yield mean estimates that fall outside of the GIP and GESIS Panel confidence intervals. All but one of the nonprobability-based means is lower than the GIP and GESIS Panel means. Differences between the nonprobability surveys are less pronounced, as most of the estimates are relatively homogeneous and lie within a close range. For the NFC variable, the nonprobability mean estimates are larger than the corresponding GIP and GESIS Panel estimates. All but two of the nonprobability surveys yield mean estimates that lie outside of the GIP and GESIS Panel CIs. Analogous to the BIG-5 estimates, most of the nonprobability NFC estimates are similar to each other. In summary, it is apparent that differences in the means exist between the probability and nonprobability surveys, but differences are less apparent between the nonprobability surveys.

4.5. Comparison of Regression Coefficients Between Surveys

Next, we compare the ML estimates of regression coefficients of BIG-5 and NFC obtained from the probability and nonprobability surveys. Control variables include age (four categories), sex (binary), marital status (three categories), occupation (four categories), secondary education certificate (three categories), region of residence (binary), internet access (binary), and housing tenure (binary). We also include a survey weight variable,

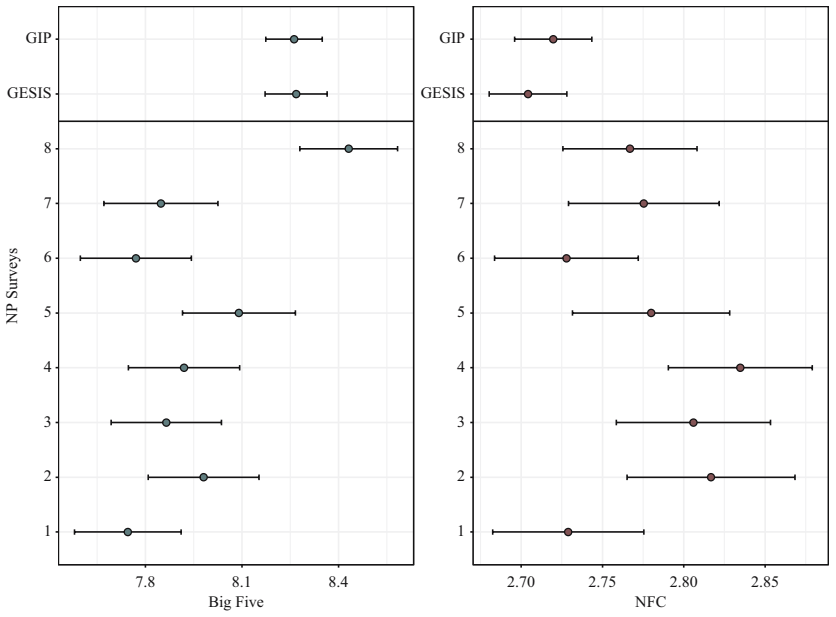

Fig. 3. Means and 95% confidence intervals for BIG-5 (left panel and Need for Cognition (NFC) (right panel) on the probability (GIP and GESIS Panel) and eight nonprobability (NP) surveys.

which was produced to reduce bias through a raking adjustment to population benchmarks (Blom et al. 2017), as a covariate in the regression. For the regression analysis of the GESIS Panel and the nonprobability surveys, we use the same independent variables, minus region and the survey weight, which were both unavailable.

Figure 4 shows the regression coefficients and 95% CIs from the BIG-5 model estimated from the GIP Panel with corresponding coefficients estimated from the nonprobability surveys. The conclusions for the GESIS Panel (not shown) are virtually the same. Overall, there is a close correspondence between the probability and nonprobability coefficients across the models. Very few of the nonprobability estimates lie outside of the CI ranges of the probability estimates. The results contrast with the results presented in Subsection 4.4, where differences in the outcome variable between the probability and nonprobability surveys were more pronounced. Our finding that regression coefficients are less affected by bias than univariate estimates in nonprobability samples is consistent with other work (Ansolabehere and Schaffner 2014; Pasek 2016).

5. Application Results

5.1. Evaluation and Efficiency

In this section, we evaluate the performance of the three modelling approaches on the GIP and GESIS Panel data by using the model-based predictions as described in Subsection 3.1. Splitting the probability survey data into training and test sets in the applicaton is done for evaluation purposes only and takes advantage of the abundant number of probability

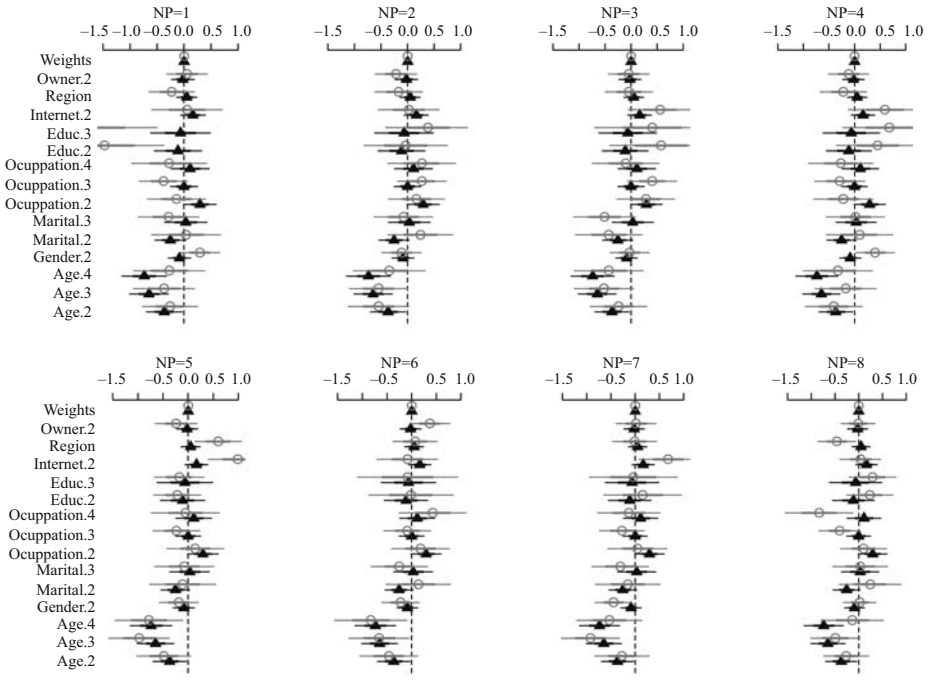

Fig. 4. Comparison of OLS regression coefficients and 95% confidence intervals for BIG-5 in the German Internet Panel (triangles) and eight nonprobability surveys (circles).

cases we have at our disposal. In practice, we envision the practitioner would only have access to a small probability sample and therefore this evaluation step would not be feasible. We then use the training set to fit the models specified in Subsection 2.1.

After excluding cases with missing data and assigning 1,000 cases from the probability survey to the training set, the remaining cases are assigned to the test set. For the BIG-5 outcome, the test set includes 1,924 and 2,150 cases for the GIP and GESIS Panel surveys, respectively. For the NFC outcome, the respective sample sizes are 1,891 and 2,088 cases. The nonprobability sample sizes are not altered.

In the application, we use $MSE(\bar{\tilde{\mathbf{y}}}) = \mathbb{E}\left[\left(\bar{\tilde{\mathbf{y}}} - \bar{\tilde{\mathbf{y}}}_{adj}^{IS}\right)^2\right]$, where $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$ are the model-adjusted, in-sample (superscript *IS*) predictions in the test set of the probability survey. These predictions are adjusted by (i) applying the regression model with the same covariates as in Models 1, 2, and 3 exclusively to the test set, with noninformative Jeffrey's priors, and then (ii) computing posterior predictive means and using them as $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$. By using the adjusted predictions rather than the original observations, we account for the fact that our model may be unrealistic and explain only a small part of data variability. An important distinction between $\bar{\tilde{\mathbf{y}}}$ and $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$ is that the former are out-of-sample predictions made by using one of the three specifications of models described in Subsection 2.1 on the training set, whereas the latter are in-sample predictions informed exclusively by the withheld test set. Analogously, the bias here is the difference between the mean of the posterior means, $\bar{\tilde{\mathbf{y}}}$, and the mean of the model-adjusted predictions $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$, that is, $Bias(\bar{\tilde{\mathbf{y}}}) = \frac{1}{n}\sum \bar{\tilde{\mathbf{y}}} - \frac{1}{n}\sum \bar{\tilde{\mathbf{y}}}_{adj}^{IS}$ (cf. Subsection 3.1).

Finally, to assess the efficiency of the two models informed by the nonprobability data (Models 2 and 3) relative to the reference model (Model 1), which is based on only weakly-informative priors, we examine the ratio of the variances of the posterior predictive means:

$$\epsilon\left(Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right), Var\left(\bar{\tilde{\mathbf{y}}}_{Model2}\right)\right) = \frac{Var\left(\bar{\tilde{\mathbf{y}}}_{Model2}\right)}{Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right)},$$

$$\epsilon\left(Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right), Var\left(\bar{\tilde{\mathbf{y}}}_{Model3}\right)\right) = \frac{Var\left(\bar{\tilde{\mathbf{y}}}_{Model3}\right)}{Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right)}.$$

Analogously, we examine the ratio of the MSEs of the posterior predictive means. If the value of any ratio is less than 1, then the informative model is more efficient than the reference model. Conversely, if the ratio is equal to or greater than 1 then the informative models do not produce efficiency gains over the reference model.

5.2. *Variance, Bias, and MSE*

This section presents the results of the three modeling approaches (Model 1, 2, and 3) implemented on the GIP and GESIS Panel surveys. The variance, bias, and MSE as defined above are computed for the posterior predictive means (hereinafter referred to simply as the mean estimates) of the two outcome variables produced under each model. The entire procedure of splitting the probability data into training and test sets was conducted 100 times to produce 100 estimates of variance, bias, and MSE for each probability sample size. The forthcoming results report the averages of these 100 repetitions. Each of the models was fitted using the independent variables described in Subsection 4.4.

The posterior characteristics were computed, as in Section 3, using three MCMC chains with samples of 1,000 and a 100 iteration burn-in sample. This ensured convergence of all chains used for creating the posterior distributions. We investigated the convergence using a larger number of iterations and found that the results were robust with respect to the number of iterations used.

Results for the BIG-5 and NFC means are shown for both GIP and GESIS Panel data in Figure 5. For brevity, we show the results using only one nonprobability survey, NP = 5, the middle-priced of the seven paid-for nonprobability surveys. Similar results (not shown) were found when the other nonprobability surveys were used.

Models 2 and 3 yield very similar variance estimates and are virtually indistinguishable in the figures. For the smallest probability sample sizes, both models yield substantially smaller variance estimates compared to the reference model (Model 1). Maximum variance efficiency is achieved with a probability sample size of 50, while efficiency gains tend to diminish as the sample size increases. All three models converge to variance equivalency at about n = 500. What is most striking is that the variance estimates produced under Models 2 and 3 for the smallest sample sizes are approximately equivalent to the variance estimates produced under the reference model for the largest probability sample size of 1,000. In other words, a probability sample size of only 50 cases with a supplement of 1,000 nonprobability cases achieves roughly the same variance as a much larger probability sample size of 1,000 does on its own.

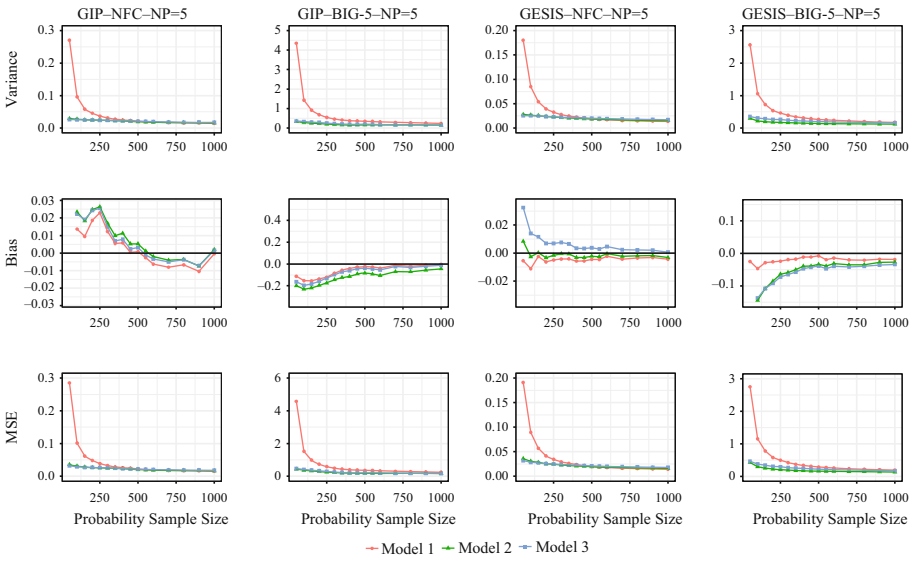

Fig. 5. Variance, bias, and mean-squared error (MSE) for estimates of BIG-5 and need for Cognition (NFC) in the GESIS Panel and GIP. Note: Results shown for one nonprobability survey (NP = 5) only. Similar results were found when other nonprobability surveys were used.

Concerning bias, as expected, the majority of plots show a slightly larger bias in Models 2 and 3 relative to the reference model for the smallest probability sample sizes, where the nonprobability-based priors have their strongest influence on the posterior estimations. In general, however, the magnitude of the bias is quite small, which is consistent with the results of the comparison of regression coefficients in Subsection 4.5.

In terms of MSE, the figures reveal that for small probability sample sizes Models 2 and 3 yield MSE values that are substantially smaller than those of the reference model. These MSE reductions persist at a decreasing rate until the probability sample size reaches about 500, at which point the values from all three models converge. The results clearly indicate that any increase in bias due to using the nonprobability-based priors is offset by the reduction in variance. Analogous to the variance results, the MSE values under Models 2 and 3 remain similarly small across the sample size spectrum. The practical implication is that the same MSE achieved through a large probability sample can be roughly achieved by supplementing a very small probability sample (e.g., 50–100 cases) with a larger nonprobability sample.

5.3. Model Efficiency and Cost Implications

In the final analysis, we summarize the MSE/variance efficiencies achieved through Models 2 and 3 and examine whether they would have likely resulted in a cost saving compared to Model 1 for a given MSE. Figure 6 presents efficiency ratios of MSE and variance for mean estimates of BIG-5 (upper panel) and NFC (lower panel) for the GIP and GESIS Panel surveys. The ratios are averaged across all eight nonprobability surveys (with equal weight given) to provide an overall summary measure of model efficiencies.

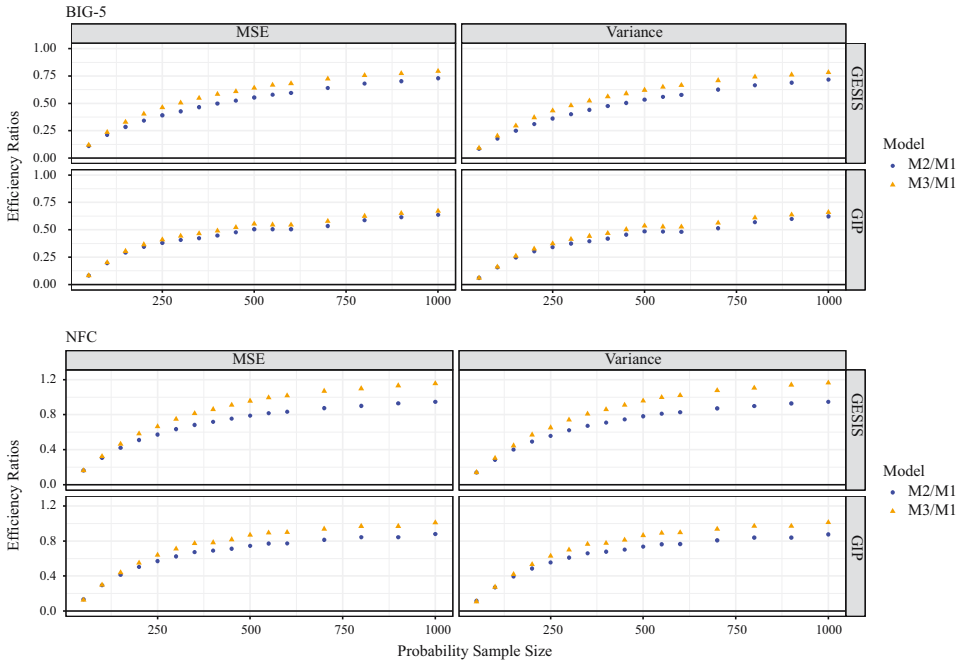

Fig. 6. Efficiency ratios of mean-squared error (MSE) and variance (columns) for mean estimates of BIG-5 (upper panel) and Need for Cognition (NFC, lower panel) in the GESIS Panel and GIP (rows). Note: Ratios are averaged across all eight nonprobability surveys.

Four observations can be made from Figure 6: (i) as observed in the previous analyses, MSE/variance efficiency gains are largest for the smallest probability sample sizes. For example, Models 2 and 3 reduce MSE and variance by more than 80%, on average, compared to Model 1 for the smallest sample size of 50. Even when the sample size is doubled to 100 cases, MSE/variance reductions of at least 70% are observed; (ii) the BIG-5 variable experiences larger efficiency gains than the NFC variable, and both variables yield slightly larger efficiency gains in the GIP than in the GESIS Panel; (iii) gains in variance efficiency are only slightly larger than gains in MSE efficiency, which indicates that the bias due to utilizing nonprobability-based priors is marginal compared to the corresponding variance reduction; and (iv) Models 2 and 3 yield very similar gains in MSE and variance efficiency with slightly larger gains achieved under Model 2.

To demonstrate the cost implications (and potential cost savings) of the different models, we utilize actual cost data for the nonprobability surveys (see Table 1) and hypothetical cost data for the probability-based GIP survey. For the GIP survey, we assume a cost per respondent of 22 euros, which is roughly 2 and 4 times larger than the most and least expensive nonprobability surveys (excluding the gratis survey), respectively. Using these data, we perform a crude estimation of the expected cost of performing a probability-only survey (under Model 1) that would achieve the same MSE that was actually achieved under Model 3 – the more conservative of the two models utilizing nonprobability-based priors. We then compare the estimated

Table 2. Estimated cost differences (and percent cost savings) between model 1 and model 3 for fixed mean-squared error (MSE) values achieved under model 3 for mean estimates of BIG-5 in the German Internet Panel.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 50							
MSE (Model 3)	0.288	0.300	0.521	0.462	0.423	0.406	0.222
Cost (Model 3)	6,492.97	6,718.57	8,161.11	8,511.00	8,736.22	9,480.46	11,776.44
Est. cost (Model 1; same MSE)	14,267.37	13,904.09	8,811.64	9,917.81	10,739.26	11,122.18	16,473.01
Cost difference (Est. M1 – M3)	7,774.40	7,185.52	650.53	1,406.81	2,003.04	1,641.72	4,696.57
Est. cost savings (%)	54.49	51.68	7.38	14.18	18.65	14.76	28.51
GIP sample size = 100							
MSE (Model 3)	0.249	0.288	0.433	0.408	0.358	0.370	0.212
Cost (Model 3)	7,592.97	7,818.57	9,261.11	9,611.00	9,836.22	10,580.46	12,876.44
Est. cost (Model 1; same MSE)	15,526.23	14,267.37	10,521.24	11,076.31	12,292.75	11,987.14	16,840.42
Cost difference (Est. M1 – M3)	7,933.26	6,448.80	1,260.13	1,465.31	2,456.53	1,406.68	3,963.98
Est. cost savings (%)	51.10	45.20	11.98	13.23	19.98	11.73	23.54
GIP sample size = 150							
MSE (Model 3)	0.227	0.262	0.397	0.373	0.346	0.345	0.203
Cost (Model 3)	8,692.97	8,918.57	10,361.11	10,711.00	10,936.22	11,680.46	13,976.44
Est. cost (Model 1; same MSE)	16,292.76	15,092.86	11,331.37	11,912.13	12,607.50	12,634.15	17,179.17
Cost difference (Est. M1 – M3)	7,599.79	6,174.29	970.26	1,201.13	1,671.28	953.69	3,202.73
Est. cost savings (%)	46.65	40.91	8.56	10.08	13.26	7.55	18.64

Table 2. Continued.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 200							
MSE (Model 3)	0.213	0.235	0.346	0.336	0.315	0.321	0.194
Cost (Model 3)	9,792.97	10,018.57	11,461.11	11,811.00	12,036.22	12,780.46	15,076.44
Est. cost (Model 1; same MSE)	16,803.26	16,009.08	1,2607.50	12,876.97	13,464.97	13,293.84	17,525.78
Cost difference (Est. M1 – M3)	7,010.29	5,990.51	1,146.39	1,065.97	1,428.75	513.38	2,449.34
Est. cost savings (%)	41.72	37.42	9.09	8.28	10.61	3.86	13.98
GIP sample size = 250							
MSE (Model 3)	0.194	0.209	0.317	0.294	0.277	0.283	0.194
Cost (Model 3)	10,892.97	11,118.57	12,561.11	12,911.00	13,136.22	13,880.46	16,176.44
Est. cost (Model 1; same MSE)	17,525.78	16,952.48	13,407.64	14,084.37	14,610.07	14,421.98	17,525.78
Cost difference (Est. M1 – M3)	6,632.81	5,833.91	846.53	1,173.37	1,473.85	541.52	1,349.34
Est. cost savings (%)	37.85	34.41	6.31	8.33	10.09	3.75	7.70

Table 3. Estimated cost differences (and percent cost savings) between model 1 and model 3 for fixed mean-squared error (MSE) values achieved under model 3 for mean estimates of need for cognition (NFC) in the German Internet Panel.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 50							
MSE (Model 3)	0.055	0.039	0.033	0.032	0.025	0.034	0.021
Cost (Model 3)	6,492.97	6,718.57	8,161.11	8,511.00	8,736.22	9,480.46	11,776.44
Est. cost (Model 1; same MSE)	4,470.38	7,104.11	8,560.89	8,837.23	11,098.38	8,294.81	12,695.68
Cost difference (Est. M1 – M3)	−2,022.59	385.54	399.78	326.23	2,362.16	−1,185.65	919.24
Est. cost savings (%)	0	5.43	4.67	3.69	21.28	0	7.24
GIP sample size = 100							
MSE (Model 3)	0.048	0.033	0.029	0.029	0.022	0.029	0.018
Cost (Model 3)	7,592.97	7,818.57	9,261.11	9,611.00	9,836.22	10,580.46	12,876.44
Est. cost (Model 1; same MSE)	5,441.20	8,560.89	9,732.26	9,732.26	12,272.42	9,732.26	14,071.49
Cost difference (Est. M1 – M3)	−2,151.77	742.32	471.15	121.26	2,436.20	−848.20	1,195.05
Est. cost savings (%)	0	8.67	4.84	1.25	19.85	0	8.49
GIP sample size = 150							
MSE (Model 3)	0.042	0.029	0.027	0.027	0.020	0.026	0.017
Cost (Model 3)	8,692.97	8,918.57	10,361.11	10,711.00	10,936.22	11,680.46	13,976.44
Est. cost (Model 1; same MSE)	6,488.51	9,732.26	10,388.86	10,388.86	13,136.09	10,736.71	14,568.12
Cost difference (Est. M1 – M3)	−2,204.46	813.69	27.75	−322.14	2,199.87	−943.75	591.68
Est. cost savings (%)	0	8.36	0.27	0	16.75	0	4.06

Table 3. Continued.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 200							
MSE (Model 3)	0.039	0.028	0.025	0.027	0.019	0.026	0.017
Cost (Model 3)	9,792.97	10,018.57	11,461.11	11,811.00	12,036.22	12,780.46	15,076.44
Est. cost (Model 1; same MSE)	7,104.11	10,054.22	11,098.38	10,388.86	13,594.43	10,736.71	14,568.12
Cost difference (Est. M1 – M3)	−2,688.86	35.65	−362.73	−1,422.14	1,558.21	−2,043.75	−508.32
Est. cost savings (%)	0	0.35	0	0	11.46	0	0
GIP sample size = 250							
MSE (Model 3)	0.037	0.026	0.024	0.026	0.018	0.025	0.018
Cost (Model 3)	10,892.97	11,118.57	12,561.11	12,911.00	13,136.22	13,880.46	16,176.44
Est. cost (Model 1; same MSE)	7,553.97	10,736.71	11,474.45	10,736.71	14,071.49	11,098.38	14,071.49
Cost difference (Est. M1 – M3)	−3,339.00	−381.86	−1,086.66	−2,174.29	935.27	−2,782.08	−2,104.95
Est. cost savings (%)	0	0	0	0	6.65	0	0

Model 1 costs with the actual and estimated costs of Model 3 for the fixed MSE. The analysis is conducted in two steps. First, a linear regression model of GIP costs (log-transformed) on MSE (and MSE squared) is fitted using the Model 1 MSE results. Next, we plug-in the MSE values achieved under Model 3 into the fitted model to estimate the (back-transformed) cost of collecting a probability-only sample. Lastly, we calculate differences between the estimated Model 1 costs and the actual/estimated Model 3 costs for each realized MSE and compute the expected cost savings (in percentages) under Model 3.

Tables 2 and 3 show the estimated cost differences between Model 1 and Model 3 for the BIG-5 and NFC outcomes, respectively. The cost differences are shown for the five smallest probability sample sizes (50, 100, 150, 200, and 250). Regarding the BIG-5 outcome, cost savings are evident for each sample size. In general, the largest cost savings occur for the smallest sample size of 50, followed by 100, and so on, which is consistent with the MSE reductions observed in the previous analyses. However, there is large variation in the amount of cost savings across the seven (paid-for) nonprobability surveys. For example, when the two least expensive nonprobability surveys (surveys 2 and 3) are used to construct the priors then estimated cost savings of about 55% and 52% are achieved, respectively, for the BIG-5 outcome with a probability sample size of 50. The other, more expensive, nonprobability surveys yield cost savings ranging from about 7% to 29% for the same sample size. For larger probability sample sizes of 100 and 150, the range of cost savings for the BIG-5 outcome is slightly reduced to between 12% and 51%, and 8% to 47%, respectively, across all nonprobability surveys. Beyond 150 probability cases, the two least expensive nonprobability surveys continue to achieve significant cost savings (greater than 30%), but as for the more expensive nonprobability surveys, the cost savings are more modest (less than 15%).

Cost savings for the NFC outcome are much less pronounced. Only nonprobability survey 6 yields a modest cost savings (about 21%) for a probability sample size of 50. The remaining nonprobability surveys produce cost savings of less than 8% for the same sample size, and some surveys achieve no cost savings at all. With a probability sample size greater than 150 cases, the majority of nonprobability surveys yield no cost savings. Thus, the cost-effectiveness of Model 3 appears to be sensitive to the probability sample size, differences in per respondent costs between the probability and nonprobability surveys, and the outcome variable of interest.

6. Discussion

This study demonstrated a novel method of using Bayesian inference to supplement small- and modest-sized probability samples with nonprobability samples in a way that can improve the cost and error properties of survey estimates. Specifically, we proposed two ways of constructing informative nonprobability-based priors. We then showed that using these priors to inform estimates derived from small probability samples yields substantially lower mean-squared errors (MSEs) compared to estimates derived from probability-only samples. Moreover, applying these informative priors to small probability samples (e.g., 50 or 100 cases) through a real-data application yielded estimates that were approximately as efficient as estimates based on much larger

probability-only samples (e.g., 1,000 cases). Reductions in MSE were primarily driven by large reductions in variability which completely offset any increases in bias. By using simulated data, we also demonstrated general applicability of the method and its mechanism for various sample sizes and levels of bias in the nonprobability samples.

Using actual cost data for several nonprobability surveys and a plausible assumed cost for a probability survey, we showed that the method can lead to large expected cost savings (up to 55% in our application) compared to a probability-only sample for a given MSE. However, the extent of cost savings depended on the outcome variable of interest and the nonprobability sample costs which varied across the survey vendors used. The largest cost savings tended to occur when the per-respondent costs were about four times greater in the probability survey than in the nonprobability survey.

At a time when many survey researchers are shifting away (or abandoning altogether) probability samples and embracing less-expensive nonprobability samples despite their known caveats, our results suggest that it is possible to retain the benefits of both sampling approaches in a way that is beneficial from both a cost and error perspective. The proposed method is ideally suited for tight survey budgets in which only a small probability sample (e.g., 50–100 cases) can be afforded alongside a larger nonprobability sample. The finding that the method can yield estimates that are just as efficient as estimates derived from very large probability samples is a particularly attractive feature for survey practice.

However, there are potential issues with the Bayesian method that should be considered. First, it is possible that some nonprobability samples may contain large biases that, when utilized as prior distributions, could negate reductions in variability and yield larger MSEs compared to probability-only samples. We did not face this issue in our application, as the estimated regression coefficients used in our models were not substantially different between the probability and nonprobability surveys. When using simulated data, we found that if the interest is in the size of the effect (regression coefficient), the combination of probability and nonprobability samples yields reductions in variance and MSE of that effect with minimal amount of bias. However, using nonprobability-based priors for model-based predictions or imputation of a missing outcome variable may not produce desired improvements if bias in the nonprobability sample is substantial (in our simulation study a bias of around 50% of the outcome variable). Thus, it would be prudent for the researcher to adjust the nonprobability sample data in advance of constructing priors to minimize bias at the outset, especially if prediction is the ultimate objective.

A further issue with the Bayesian approach is the vast number of modeling specifications and prior configurations that one could employ. We deliberately kept the modeling and prior specification as basic as possible. This sometimes required choosing simplicity over complexity in order to facilitate implementation and minimize computation time. Further refinements of the modeling approach could be developed to account for more complex data structures, such as categorical outcome variables. In addition, adapting the modeling approach to incorporate complex sample design features (e.g., stratum, cluster indicators) is an area for future work.

In conclusion, we find that augmenting a probability sample with a nonprobability sample under a Bayesian framework can produce survey estimates with smaller MSE and

potentially large cost savings relative to probability-only samples. The proposed method, which turns the usual approach of treating a probability sample as an unbiased prior for a nonprobability sample "on its head" as one reviewer put it, could be a useful import to survey practice where cost-saving measures and error-reduction tools are in high demand. However, despite the advantages of the method, survey organizations using nonprobability samples may still be skeptical to the idea of fielding a small probability sample survey in parallel when the nonprobability sample will likely dominate the inference. Here, we would contend that adopting a system of estimation that accounts for both sampling streams, yet incentivizes probability-based observations and allows for the direct quantification of uncertainty in survey estimates is a more defensible strategy than one that renounces probability sampling entirely along with all of its attractive theoretical properties. Moreover, the idea of enhancing a small, but carefully designed, probability sample with abundant but potentially error-prone data is not a new idea and is a widely accepted strategy in small area applications where sparse probability samples are routinely supplemented with alternative data sources to improve the cost and error properties of population estimates (Marchetti et al. 2016; Porter et al. 2014; Briggs et al. 2007; Schmertmann et al. 2013).

7. References

AAPOR. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed July 2019).

Ansolabehere, S. and D. Rivers. 2013. "Cooperative Survey Research." *Annual Review of Political Science* 16: 307–329. Doi: https://doi.org/10.1146/annurev-polisci-022811-160625.

Ansolabehere, S. and B.F. Schaffner. 2014. "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison." *Political Analysis* 22(3): 285–303. Doi: https://doi.org/10.1093/pan/mpt025.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. *Report of the AAPOR Task Force on Non-Probability Sampling*. American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf (accessed July 2019).

Blom, A.G., D. Ackermann-Piek, S.C. Helmschrott, C. Cornesse, and J.W. Sakshaug. 2017. "The Representativeness of Online Panels: Coverage, Sampling and Weighting." *Paper Presented at the General Online Research Conference*.

Blom, A.G., C. Gathmann, and U. Krieger. 2015. "Setting Up an Online Panel Representative of the General Population: The German Internet Panel." *Field Methods* 27(4): 391–408. Doi: https://doi.org/10.1177/1525822X15574494.

Blom, A.G., J.M.E. Herzing, C. Cornesse, J.W. Sakshaug, U. Krieger, and D. Bossert. 2016a. "Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence from the German

Internet Panel." *Social Science Computer Review* 35(4): 498–520. Doi: https://doi.org/10.1177/0894439316651584.

Blom, A.G., M. Bosnjak, A. Cornilleau, A.-S. Cousteaux, M. Das, S. Douhou and U. Krieger. 2016b. "A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe." *Social Science Computer Review* 35(1): 8–25. Doi: https://doi.org/10.1177/0894439315574825.

Bosnjak, M., T. Dannwolf, T. Enderle, I. Schaurer, B. Struminskaya, A. Tanner, and K.W. Weyandt. 2017. "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel." *Social Science Computer Review* 36(1): 103–115. Doi: https://doi.org/10.1177/0894439317697949.

Briggs, D., D. Fecht, and K. De Hoogh. 2007. "Census Data Issues for Epidemiology and Health Risk Assessment: Experiences from the Small Area Health Statistics Unit." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170(2): 355–378. Doi: https://doi.org/10.1111/j.1467-985X.2006.00467.x.

Cacioppo, J.T. and R.E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42(1): 116. Doi: https://doi.org/10.1037/0022-3514.42.1.116.

Callegaro, M., A. Villar, J. Krosnick, and D. Yeager. 2014. "A Critical Review of Studies Investigating the Quality of Data Obtained with Online Panels." In *Online Panel Research. A Data Quality Perspective*, edited by M. Callegaro, R.P. Baker, J. Bethlehem, A.S. Goeritz, J.A. Krosnick, and P.J. Lavrakas, 23–53. Chichester, UK: John Wiley & Sons. Doi: https://doi.org/10.1002/9781118763520.ch2.

Chang, L. and J.A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4): 641–678. Doi: https://doi.org/10.1093/poq/nfp075.

Digman, J.M. 1990. "Personality Structure: Emergence of the Five-factor Model." *Annual Review of Psychology* 41(1): 417–440. Doi: https://doi.org/10.1146/annurev.ps.41.020190.002221.

DiSogra, C., C. Cobb, E. Chan, and J. Dennis. 2012. "Using Probability-Based Online Samples to Calibrate Non-Probability Opt-in Samples." *Presentation at: 67th Annual Conference of the American Association for Public Opinion Research (AAPOR)*. Available at: http://www.websm.org/uploadi/editor/1361444163DiSogra_et_al_2012_Using_Probability_Based_Online_Samples.ppt (accessed July 2019).

Dutwin, D. and T.D. Buskirk. 2017. "Apples to Oranges or Gala Versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples." *Public Opinion Quarterly* 81(S1): 213–239. Doi: https://doi.org/10.1093/poq/nfw061.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, 1–26. Doi: https://doi.org/10.1007/978-1-4612-4380-9_41.

Elliott, M.N. and A. Haviland. 2007. "Use of a Web-based Convenience Sample to Supplement a Probability Sample." *Survey Methodology* 33(2): 211–215. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10498-eng.pdf?st=A8NHMZ2v (accessed July 2019).

Elliott, M.R. 2013. "Combining Data from Probability and Non-probability Samples Using Pseudo-weights." *Survey Practice* 2(6). Doi: https://doi.org/10.29115/SP-2009-0025.

Erens, B., S. Burkill, M.P. Couper, F. Conrad, S. Clifton, C. Tanton, A. Phelps, J. Datta, C.H. Mercer, P. Sonnenberg, et al. 2014. "Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison with a Probability Sample Interview Survey." *Journal of Medical Internet Research* 16(12). Doi: https://doi.org/10.2196/jmir.3382.

Fahimi, M., F.M. Barlas, W. Gross, and R.K. Thomas. 2014. "Towards a New Math for Nonprobability Sampling Alternatives." *Presented at the 69th Annual Conference of the American Association for Public Opinion Research (AAPOR)*.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2013. *Bayesian Data Analysis*, Third Edition. Boca Raton, FL, USA: Chapman & Hall/CRC. ISBN: 9781439840955.

Gelman, A., S. Goel, D. Rothschild, and W. Wang. 2016. "High-frequency Polling with Non-representative Data." In Political Communication in Real Time: Theoretical and Applied Research Approaches (eds. D. Schill, R. Kirk, and A.E. Jasperson). Routledge, 117–133.

Goldberg, L.R. 1993. "The Structure of Phenotypic Personality Traits." *American Psychologist* 48(1): 26. Doi: https://doi.org/10.1037/0003-066X.48.1.26.

Herzing, J.M.E. and A.G. Blom. 2019. "The Influence of a Person's IT Literacy on Unit Nonresponse and Attrition in an Online Panel." *Social Science Computer Review* 37(3): 404–424. Doi: https://doi.org/10.1177/0894439318774758.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. 2016. Evaluating Online Nonprobability Surveys. Vendor Choice Matters; Widespread Errors Found for Estimates Based on Blacks and Hispanics, Pew Research Center. Available at: http://www.pewresearch.org/2016/05/02/evaluatingonline-nonprobability-surveys/ (accessed July 2019).

Lee, S. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22(2): 329. Available at: https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/propensity-score-adjustment-as-a-weighting-scheme-for-volunteer-panel-web-surveys.pdf (accessed July 2019).

Lee, S. and R. Valliant. 2009. "Estimation for Volunteer Panel Web Surveys using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research* 37(3): 319–343. Doi: https://doi.org/10.1177/0049124108329643.

MacInnis, G., J.A. Krosnick, S. Ho, and M.J. Cho. 2018. "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension." *Public Opinion Quarterly*. Volume 82, Issue 4, 707–744. Doi: https://doi.org/10.1093/poq/nfy038.

Malhotra, N. and J.A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis*, 286–323. Doi: https://doi.org/10.1093/pan/mpm003.

Marchetti, S., C. Giusti, and M. Pratesi. 2016. "The Use of Twitter Data to Improve Small Area Estimates of Households' Share of Food Consumption Expenditure in Italy." *AStA Wirtschafts-und Sozialstatistisches Archiv* 10(2–3): 79–93. Doi: https://doi.org/10.1007/s11943-016-0190-4.

Mercer, A.W., F. Kreuter, S. Keeter, and E.A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference." *Public Opinion Quarterly* 81(S1): 250–271. Doi: https://doi.org/10.1093/poq/nfw060.

Pasek, J. 2016. "When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence." *International Journal of Public Opinion Research* 28(2): 269–291. Doi: https://doi.org/10.1093/ijpor/edv016.

Pennay, D.W., D. Neiger, P.J. Lavrakas, K.A. Borg, S. Mission, and N. Honey. 2018. "The Online Panels Benchmarking Study: a Total Survey Error Comparison of Findings from Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia." *Australian National University, Centre for Social Research and Methods Paper NO. 2/2018*. Available at: http://csrm.cass.anu.edu.au/sites/default/files/docs/2018/12/CSRM_MP2_2018_ONLINE_PANELS.pdf (accessed July 2019).

Porter, A.T., S.H. Holan, C.K. Wikle, and N. Cressie. 2014. "Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates." *Spatial Statistics* 10: 27–42. Doi: https://doi.org/10.1016/j.spasta.2014.07.001.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/ (accessed July 2019).

Rao, J.N. 2003. *Small-Area Estimation*. Wiley Online Library. Doi: https://doi.org/10.1002/0471722189.

Rivers, D. 2007. "Sampling for Web Surveys." Presented at the *Joint Statistical Meetings*. Available at: https://pdfs.semanticscholar.org/fffa/a7e52c5d163a0944974a68160ee6e0a6b481.pdf (accessed July 2019).

Rivers, D. and D. Bailey. 2009. "Inference from Matched Samples in the 2008 US National Elections." In *Proceedings of the Joint Statistical Meetings*, Volume 1, 627–639. Palo Alto, CA: YouGov/Polimetrix. Available at: https://pdfs.semanticscholar.org/e566/fb48f88ae34640b729387cbd4006249f8c45.pdf (accessed July 2019).

Schmertmann, C.P., S.M. Cavenaghi, R.M. Assunção, and J.E. Potter. 2013. "Bayes Plus Brass: Estimating Total Fertility for Many Small Areas from Sparse Census Data." *Population Studies* 67(3): 255–273. Doi: https://doi.org/10.1080/00324728.2013.795602.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2007. OpenBUGS user manual, version 3.0.2. *MRC Biostatistics Unit, Cambridge*.

Sturtz, S., U. Ligges, A. Gelman, et al. 2005. "R2WinBUGS: A Package for Running WinBUGS from R." *Journal of Statistical Software* 12(3): 1–16. Doi: https://doi.org/10.18637/jss.v012.i03.

Tourangeau, R. and T. Plewes. 2013. *Nonresponse in Social Science Surveys: A Research Agenda*. National Academies Press. Doi: https://doi.org/10.17226/18293.

Valliant, R. and J.A. Dever. 2011. "Estimating Propensity Adjustments for Volunteer Web Surveys." *Sociological Methods & Research* 40(1): 105–137. Doi: https://doi.org/10.1177/0049124110392533.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. "Forecasting Elections with Non-representative Polls." *International Journal of Forecasting* 31(3): 980–991. Doi: https://doi.org/10.1016/j.ijforecast.2014.06.001.

Yeager, D.S., J.A. Krosnick, L. Chang, H.S. Javitz, M.S. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-probability Samples." *Public Opinion Quarterly* 75(1): 709–747. Doi: https://doi.org/10.1093/poq/nfr020.

Received November 2018
Revised February 2019
Accepted April 2019

sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 683–697, http://dx.doi.org/10.2478/JOS-2019-0028

Tests for Price Indices in a Dynamic Item Universe

Li-Chun Zhang[1], *Ingvild Johansen*[2], *and Ragnhild Nygaard*[2]

There is generally a need to deal with quality change and new goods in the consumer price index due to the underlying dynamic item universe. Traditionally axiomatic tests are defined for a fixed universe. We propose five tests explicitly formulated for a dynamic item universe, and motivate them both from the perspectives of a cost-of-goods index and a cost-of-living index. None of the indices that are currently available for making use of scanner data satisfies all the tests at the same time. The set of tests provides a rigorous diagnostic for whether an index is completely appropriate in a dynamic item universe, as well as pointing towards the directions of possible remedies. We thus outline a large index family that potentially can satisfy all the tests.

Key words: Axiomatic test; cost of goods; cost of living; quality change; new goods.

1. Introduction

The failure to take full account of quality change and new goods is one of the important sources of the potential bias of the consumer price index (CPI). See for example CPI Manual. By convention the term 'quality change' pertains to situations where new products, models, services and so on are compared to the old items they are deemed to replace, whereas new goods are thought of as wholly new types of items, such as when microwave ovens were first introduced in the market. The cause of the problems is thus the same, namely, the item universe of the CPI is dynamic such that one needs to compare the prices of different sets of items over time.

In recent years, greater uses are made of scanner data, where one has access to the average transaction price and quantity of each item (identified by the Global Trade Item Number, GTIN) over the period of data collection. A key research question is how to make use of the quantity data that have become available below the level of elementary aggregation. Explicit replacement of the old items by the new ones for matched price comparisons becomes exceedingly resource demanding, if applied to all the available items. The hedonic methods are often infeasible due to the lack of characteristics data. The research is active at the moment regarding index formulae that are only based on the price and quantity data. See for example Chessa et al. (2017), Dalén (2017), Diewert and Fox (2017), Zhang et al. (2017). However, there is currently a lack of consensus on how to evaluate them. While the choice of index formula may not seem to have a big impact on certain consumption segments, including food (e.g., Chessa et al. 2017; Zhang et al. 2017),

[1] University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom. Email: L.Zhang@soton.ac.uk
[2] Statistics Norway, P.O. Box 2633 St. Hanshaugen NO-0131 Oslo, Norway. Emails: ingvild.johansen@ssb.no and ragnhild.nygaard@ssb.no

the choice does matter in many other situations, such as when there is a high item churn rate, and/or where the prices of goods undergo strong decline during their respective life spans. See for example Chessa et al. (2017) for some relevant empirical evidence on men's t-shirts and television.

The test approach provides a valuable theoretical framework for index numbers, in addition to the economic and stochastic approaches. However, the traditional axiomatic tests (Fisher 1922; CPI Manual 2004, chap. 16) are all defined for a fixed item universe. Tests for multilateral index have been considered for international comparisons; see, for example Diewert (1999) and Balk (1996, 2001). These tests have been adapted in terms of price and quantity comparisons over time (ABS 2016), though without addressing the issue that the universe is actually dynamic over time. None of the major multilateral methods considered in ABS (2016) passes all the tests.

In this article, we propose five tests explicitly formulated for a dynamic item universe. The tests will be motivated both from the perspectives of a *cost-of-goods index (COGI)* and a *cost-of-living index (COLI)*. As mentioned in paragraph 16.2 of the CPI manual, "different price statisticians may have different ideas about which tests are important [...]" Moreover, there is generally the possibility that a test (or axiom), which otherwise seems attractive, may be in conflict with the principles of another theoretical approach to index numbers. Therefore, the plausibility of a test is clearly strengthened if it can be motivated from both the COGI and COLI perspectives.

Notice that definition of a COGI in a dynamic item universe is not evident. According to Schultze and Mackie (2002), COGI "measures the change in expenditures required by a household to purchase a fixed-weight basket of goods and services when prices change between some initial reference period and a subsequent comparison period", whereas COLI "measures the change in expenditures a household would have to make in order to maintain a given standard of living". Accordingly, we shall maintain that a COGI is calculated based on a fixed-weight basket of items despite the fact that the item universe is dynamic over time, regardless of how one settles the choice of the items to be included in such a fixed basket.

The proposed tests will be applied to the indices for scanner data, which have so far received the most attention. These include the Geary-Khamis (GK) index (Geary 1958; Chessa 2016), the generalised unit value (GUV) index family (Dalén 2001; De Haan 2001; Auer 2014), the weighted geometric means (WGM) index family including the time product dummy (TPD) index (De Haan and Krsinich 2014), and the Gini-Eltetö-Koves-Szulc (GEKS) index (Ivancic et al. 2011). It is shown that none of them satisfy all five tests at the same time. We will outline a large index family, which includes the GUV index family as a subclass in general and the Fisher index as a special case of bilateral index in a fixed item universe. As will be explained, this family of indices can potentially satisfy all the tests, thus providing impetus for future research.

Overall, extending the test approach to a dynamic universe has at least three advantages. (i) Formulating explicit tests strengthens the rigour of investigation, because a test result is dichotomous by construction whereas an empirical comparison may be inclusive on its own. (ii) It can help to clarify certain conceptual ambiguities. For instance, transitivity can obviously prevent chain drifting in a fixed universe. However, as we shall discuss in Subsection 2.3, a rigorous definition of transitivity is impossible in

a dynamic universe. The proposed tests can help to articulate the intuition against chain drifting and to avoid an index method that suffers from chain drifting. (iii) Where an index fails a test, it points to the direction in which the index can possibly be improved. For instance, it will be shown that the bilateral GK index fails the responsiveness test T5 in general.

The rest of the article is organised as follows. In Section 2, the necessary notations and concepts are introduced, following which the tests are defined and explained, including a discussion of why some other seemingly obvious tests are not included in the proposal. In Section 3 the tests are applied to the indices mentioned above. Finally, an index family is outlined in Section 4 which potentially can satisfy all the tests, and a few concluding remarks are given in Section 5.

2. Tests for a Dynamic Item Universe

2.1. Notations

Let us formally introduce the notations and terms that are necessary for a dynamic item universe. Consider a given *item universe* in period t, denoted by $U_t = \{1, 2, \ldots, N_t\}$, which constitutes a subset (or sector) of the entire CPI item universe. For instance, U_t may refer to all food and non-alcoholic beverages sold at supermarkets, or all personal computers sold at electro warehouses, and so on. Notice that generally we allow the items to be specific for an outlet or chain. For example, the same mobile phone model sold at two different outlets can be treated as two different items. Assume that one has available the unit value price p_i^t and the transaction total v_i^t, for each $i \in U_t$, where $v_i^t = p_i^t q_i^t$ and q_i^t denotes the transaction quantity of the item. Note that the unit value price refers to the average transaction price of an item in the period t, where the actual transaction price may change over the period. Denote the set of quantities by $q(U_t) = \left\{q_i^t; i \in U_t\right\}$ and the set of prices by $p(U_t) = \left\{p_i^t; i \in U_t\right\}$. Denote all the data from period t by

$$D_t = D(U_t) = \{p(U_t), q(U_t); U_t\}.$$

Denote by (U_0, U_t) the *comparison universe*, for which we seek a price index from 0 to t, denoted by $P^{0,t}$. We refer to 0 as the *base* period and t the *statistical* (or *current*) period. Denote by $U_{R(0,t)} = \{U_r; r \in R\}$ the *reference universe*, where R is the set of reference periods involved, and the notation $U_{R(0,t)}$ emphasises the dependence of R on the comparison universe – we may suppress $(0, t)$ whenever such an emphasis is unnecessary. We consider price indices that are functions of the data $D_R = D(U_R) = \{D_t; t \in R\}$, which can be given as

$$P^{0,t} = f(D_R). \tag{1}$$

Two choices of the reference universe U_R are most immediate, that is

$$R_B = \{0, t\} \quad \text{and} \quad U_{R_B} = U_0 \cup U_t,$$
$$R_M = \{0, 1, \ldots, t\} \quad \text{and} \quad U_{R_M} = U_0 \cup U_1 \cup \cdots \cup U_t,$$

where R_B implies direct comparison between 0 and t, for example December 2014 (0) and July 2015 (t), and R_M implies all the periods from 0 to t, for example December 2014 to July 2015. An index based on the reference universe U_{R_B} is referred to as a *bilateral* index; it is *multilateral* otherwise.

We notice that, when 0 is fixed as t moves forward, the choice of R_M above corresponds to using a fixed base expanding window. It is, of course, possible to use another reference set, such as a rolling window of fixed length counting backwards from t, which includes period 0. Moreover, the definition of reference universe formally covers the use of chained index, for example when t moves beyond a year-length window into the next one. We do not discuss chained index explicitly in the sequels, because a chained index would fail a test if any part of the chain fails that test. Finally, unless otherwise noted, one may keep in mind the reference universe R_M above when reading this article, because the conclusions then also hold for other choices of R_M.

2.2. The Tests

In this subsection we formulate five tests for a dynamic item universe, and motivate them from both the COGI and COLI perspectives. Note that in the case of fixed item universe and $R = R_B$, both the tests T1 and T2 reduce to well-known tests for a fixed item universe. However, more generally, $U_0 = U_t$ represents the special case where a dynamic universe may sometimes return to the same state of affairs, even when $U_r \neq U_0 = U_t$ for all $r \neq 0, t$. By the tests T1 and T2 we require that a dynamic-universe index should not then have counter-intuitive properties.

Let $U_{0t} = U_0 \cap U_t$ be the *persistent* item universe at 0 and t, and $U_{t\backslash 0} = U_t \backslash U_0$ the set of *birth* items and $U_{0\backslash t} = U_0 \backslash U_t$ that of the *death* items. The item universe is (always) *fixed* if $U_t = U_0 = U_{0t}$ for any $t \neq 0$, in which case $U_{t\backslash 0} = U_{0\backslash t} = \emptyset$; it is *dynamic* otherwise.

Identity test (T1) If $U_0 = U_t$ and $p_i^0 \equiv p_i^t$ for all $i \in U_0$, then $P^{0,t} = 1$.

Since the item universe is the same at 0 and t, so must be the items eligible for a COGI, when the comparison universe is (U_0, U_t). Thus, despite the changes of item universe that take place between the two time points, where $t > 1$, the identity test can be motivated for a COGI. Now that all the prices are the same at 0 and t, a COGI index must necessarily be 1, regardless of how the reference quantities of the goods are calculated. Therefore, a cost-of-goods index should satisfy the identity test. Next, consider a COLI. Let $V^{0,t} = \sum_{i\in U_t} q_i^t p_i^t / \sum_{i\in U_0} q_i^0 p_i^0$ be the ratio of total expenditures. Under the stipulated setting, it is obviously possible to maintain the same utility without changing the total expenditure. Thus, insofar as $V^{0,t} \neq 1$, all the change in expenditure must be attributed to the change in utility but *not* prices, under the assumption of rational consumer behaviour. A COLI should therefore be equal to 1.

Despite the item universe varies at other points than 0 and t, an index that satisfies the identity test is said *not to drift* in this situation, which can be used to examine for example, a multilateral index. We notice that chain drift is often contrasted with transitivity. However, as will be discussed in Subsection 2.3, we find it difficult to define a transitivity

test for a dynamic item universe. For the moment the identity test T1 is the only test we have with respect to chain drift.

Fixed basket test (T2) If $U_0 = U_t$ and $q_i^0 \equiv q_i^t$ for all $i \in U_0$, then $P^{0,t} = V^{0,t}$.

The test is obvious for a COGI. It can easily be satisfied by any bilateral COGI. Otherwise, in a dynamic universe, one may have $q_i^r \neq q_i^0$, for $0 < r < t$, so that the item reference quantity can differ from $q_i^0 = q_i^t$, and a fixed basket-of-goods index may not be equal to $V^{0,t}$. It follows that, in order for a COGI to satisfy T2, one may need to avoid the use of multilateral indices. The test T2 is readily motivated for a COLI. Given the quantities $q(U_t)$ are actually the same as $q(U_0)$, no utility adjustment of $V^{0,t}$ is needed, and a COLI at the observed utility is equal to $V^{0,t}$.

Upper bound test (T3) If $U_0 \subseteq U_t$, and $p_i^t \leq p_i^0$ for all $i \in U_0$, then $P^{0,t} \leq 1$.

That is, the item universe may be constant if $U_0 = U_t$ or strictly expanding if $U_0 \subset U_t$, and the price of each persistent item is the same or reduced, that is $p_i^t \leq p_i^0$ for all $i \in U_{0t} = U_0$. The proof is rooted in the economic theory of Preferences. Basically, the premises of test T3 set up a situation which *guarantees* that a consumer can maintain their utility without increasing the total expenditure. Given this is the case, the intuition is that any rational change of consumption can only occur because the consumer has found a better utility-cost ratio, as the choices increase with the expanding item universe. The arguments are given below in details.

Firstly, suppose substitution does not occur, in which case $q_i^t = q_i^0$ for all $i \in U_0$ and $q_i^t = 0$ for all $i \in U_{t\backslash 0}$, even if $U_{t\backslash 0}$ is nonempty. The actual comparison universe reduces then to U_{0t}, so that the test T2 applies, yielding $P^{0,t} = V^{0,t} \leq 1$ under the stipulated setting. Next, suppose substitution occurs only among the persistent items, i.e., $q_i^t = 0$ for $i \in U_{t\backslash 0}$ and $q_i^t \neq q_i^0$ for some $i \in U_{0t}$. Substitution can be accounted for from the perspective of COLI. Given the actual $\{q_i^t; i \in U_{0t}\}$ and the corresponding utility at t, it cannot cost less for the same $\{q_i^t; i \in U_{0t}\}$ at 0 since $p_i^t \leq p_i^0$ for all $i \in U_{0t}$. It follows that a COLI must be less than or equal to 1. Finally, suppose substitution also involves the items in $U_{t\backslash 0}$. Let $\{\tilde{q}_i^t; i \in U_{0t}\}$ be a hypothetical set of persistent items that would have yielded the same utility as the actual $\{q_i^t; i \in U_t\}$. Owing to rational behaviour, the expenditure of $\{\tilde{q}_i^t; i \in U_{0t}\}$ at t cannot be less than the actual expenditure of $\{q_i^t; i \in U_t\}$; whereas the expenditure of $\{\tilde{q}_i^t; i \in U_{0t}\}$ at 0 cannot be less than that at t. It follows again that a COLI must be less than or equal to 1.

Lower bound test (T4) If $U_t \subseteq U_0$, and $p_i^t \geq p_i^0$ for all $i \in U_t$, then $P^{0,t} \geq 1$.

That is, the item universe may be constant or strictly shrinking, and the price of each persistent item is the same or increased. Firstly, suppose substitution does not occur, in which case $q_i^t = q_i^0$ for all $i \in U_{0t}$ and $q_i^t = 0$ for all $i \in U_{0\backslash t}$. Then, the comparison universe reduces to U_{0t}, and the test T2 applies, yielding $P^{0,t} = V^{0,t} \geq 1$ under the stipulated setting. Next, suppose substitution occurs only among the persistent items. Given any actual $\{q_i^t; i \in U_{0t}\}$ and the corresponding utility at t, it cannot cost more for the same $\{q_i^t; i \in U_{0t}\}$ at 0 since $p_i^0 \leq p_i^t$ for all $i \in U_{0t}$. It follows that a COLI must be greater than or equal to 1. Finally, suppose substitution also involves the items in $U_{0\backslash t}$. Let $\{\tilde{q}_i^0; i \in U_{0t}\}$ be a hypothetical set of persistent units that would have yielded the same

utility as the actual $\{q_i^0; i \in U_0\}$. The expenditure of $\{\tilde{q}_i^0; i \in U_{0t}\}$ at 0 cannot be less than the actual expenditure of $\{q_i^0; i \in U_0\}$; whereas the expenditure of $\{\tilde{q}_i^0; i \in U_{0t}\}$ at t cannot be less than that at 0. It follows again that a COLI must be greater than or equal to 1.

Under the setting of test T3, there exists a clear downwards trend of the prices of persistent items. We should have $P^{0,t} \leq 1$ even if this leads to an increase of expenditure, that is $V^{0,t} > 1$. Under the setting of test T4, there exists a clear upwards trend of the prices of persistent items. We should have $P^{0,t} \geq 1$ even if the price increase causes the expenditure to drop, that is $V^{0,t} < 1$.

It is possible to formulate two somewhat sharper versions of the tests T3 and T4, respectively, according to which $P^{0,t}$ can possibly deviate from 1 in a particular direction depending on whether the item universe is expanding or shrinking, when all the prices of persistent items remain the same. These are thus clearly the implications of the fact that the item universe is dynamic.

Test t3 If $U_0 \subset U_t$ and $p_i^0 = p_i^t$ for all $i \in U_0$, then $P^{0,t} \leq 1$.

Test t4 If $U_t \subset U_0$ and $p_i^0 = p_i^t$ for all $i \in U_0$, then $P^{0,t} \geq 1$.

Responsiveness test (T5) For $U_0 \neq U_t$, $P^{0,t}$ should not always reduce to $f(D_{0t})$, where $D_{0t} = D(U_{0t})$.

That is, one should not always be able to reduce a COGI to a fixed-basket index, where the basket items only consist of the persistent items. This is necessary for any COGI that, in principle, can be applied to a dynamic item universe; whereas one should not always be able to reduce a COLI to $f(D_{0t})$, since it must allow for substitution that involves the birth and death items.

One can formulate a sharper version of the test T5, where $p_i^0 = p_i^t$ for all $i \in U_{0t}$. According to T1, we have then $P^{0,t}(D_{0t}) = 1$, which is the price index of the persistent item universe U_{0t}. Any $P^{0,t}$ that is always equal to 1, regardless of $D(U_{t\backslash 0})$ or $D(U_{0\backslash t})$, is not responsive.

Test t5 For $U_0 \neq U_t$, if $p_i^0 = p_i^t$ for all $i \in U_{0t}$, then $P^{0,t}$ cannot always be equal to 1, regardless of $D(U_{t\backslash 0})$ and $D(U_{0\backslash t})$.

2.3. Discussion

The proposed tests are certainly not the only ones possible. However, we have not included any other tests here for several reasons. Firstly, it would have made little difference to include a test that is easily satisfied, an example of which is the time reversal test. Next, some tests seem no longer relevant given the birth and death items. The quantity reversal and price reversal tests are two such examples. Moreover, we have excluded some familiar tests and only retained a sharper version of them. An example is the proportionality test. Since the proportionality test implies the identity test T1, the latter is sharper than the former, in the same sense that the test t3 is sharper than T3. Finally, there may be additional concerns that make a test difficult to formulate. An example is the transitivity test, as we discuss below.

Conceptually, an index is transitive if $P^{0,t} = P^{0,r}P^{r,t}$ for any $r \neq 0, t$, provided all the three indices in the form of (1) are calculated in the *same* way, which generally involves

three different reference universes $U_{R(0,t)}$, $U_{R(0,r)}$ and $U_{R(r,t)}$ when the item universe is dynamic. Now, a motivation for transitivity is to prevent chain drifting, when chaining is used to alleviate the difficulty one would encounter in making direct price comparisons between U_0 and U_t, where $U_{t\backslash 0}$ and $U_{0\backslash t}$ may be non-negligible in size compared to U_{0t}. However, in order to verify whether or not chain drifting is the case, one must compare the chained index to the direct index that could have been calculated between 0 and t. Thus, one cannot avoid running into the same difficulty that has motivated the chaining in the first place. To push the difficulty to the logical extreme, suppose $U_0 \cap U_t = \emptyset$, i.e., the item universe is completely renewed. What are the conditions of non-drifting, or transitivity, in this case?

From a more pragmatic point of view, the GEKS index has been adapted for temporal price comparisons (Ivancic et al. 2011), as a means to achieve transitivity, provided $P^{0,r}$ are well-defined and time reversible for any $0 < r \leq t$. However, international comparisons have a fixed reference set of countries (or regions), and are symmetric in the sense that any two countries are eligible for comparison. The temporal extension has a direction and is ever-changing. Regarding the direction of time, it seems counter-intuitive to require $P^{0,r}P^{r,t} = P^{0,t}$, for an arbitrarily chosen period r, where $r < 0$ or $r > t$. Regarding the changing reference set, the GEKS index $P^{0,t}$ calculated at t will generally differ to $P^{0,t}$ calculated at t', for $t' > t$. It follows from both accounts that in reality the disseminated GEKS index is nevertheless intransitive – see Subsection 3.4 for details.

3. Application

The test results are summarised in Table 1. We show that the constant-value adjustment of the GK index, which is necessary in the context of international comparisons involving different currencies, results in unresponsiveness if $R = R_B$. Dropping the constant-value adjustment yields the modified GK (MGK) index, which is a member of the GUV index family. We consider also the WGM index family, which includes the TPD index as a special case in a dynamic universe and the Törnqvist index as a special case in a fixed universe. Finally, we discuss the GEKS index, which can use any bilateral time-reversible index as its component.

Table 1. Summary of test results.

	Identity	Fixed-basket	Upper-bound	Lower-bound	Responsiveness
GK	Yes if R_B No if R_M	Yes	Yes	Yes	No if R_B
GUV (MGK)	Yes if R_B No if R_M	Yes	Yes	Yes	No if R_B and in Setting of t3 (or t4)
WGM (TPD)	Yes if R_B No if R_M	No	Yes	Yes	No if R_B and in Setting of t3 (or t4)
GEKS	No	No	No	No	Yes

3.1. The GK Index

Deflating a constant-value reference-price quantity index yields the GK index:

$$P_{GK}^{0,t} = V^{0,t}/Q^{0,t} \quad \text{and} \quad Q^{0,t} = \sum_{i\in U_t} p_i q_i^t / \sum_{i\in U_0} p_i q_i^0, \tag{2}$$

$$p_i = \sum_{r\in R_i} \frac{p_i^r}{P^{0,r}} q_i^r / \sum_{r\in R_i} q_i^r, \tag{3}$$

where the observed price p_i^r is adjusted to a constant-value price by $P^{0,r}$, and R_i contains the periods at which the item i is in the market.

The GK index given by (2) and (3) does not satisfy the identity test T1 except when $R = R_B$, in which case $p_i = p_i^0 = p_i^t$. Next, it satisfies the fixed-basket test T2 since $Q^{0,t} = 1$. Thirdly, it satisfies the upper bound tests T3 and t3, provided $p_i^t \le p_i \le p_i^0$ by (3), such that

$$Q^{0,t} = \frac{\sum_{j\in U_0} p_j q_j^t + \sum_{j\in U_{t\setminus 0}} p_j^t q_j^t}{\sum_{j\in U_0} p_j q_j^0} \ge \frac{\sum_{j\in U_0} p_j^t q_j^t + \sum_{j\in U_{t\setminus 0}} p_j^t q_j^t}{\sum_{j\in U_0} p_j^0 q_j^0} = V^{0,t}.$$

Moreover, it satisfies the lower bound tests T4 and t4, since we then have $Q^{0,t} \le V^{0,t}$.

When it comes to the test T5, below we give a proof that the bilateral GK index is generally unresponsive. Previously, Geary (1958) pointed out that for spatial comparison with a matched commodity universe between two countries, the GK index reduces to a fixed weights price index, where the weight of a commodity is given by the harmonic means of its quantities in both countries. As we now demonstrate, the same result holds even when the two universes are not completely matched. In the present notation, the bilateral GK index can be given as E_0/E_t, which together with the reference prices satisfy the following linear system

$$p_i = \begin{cases} E_0 \dfrac{p_i^0 q_i^0}{q_i^0 + q_i^t} + E_t \dfrac{p_i^t q_i^t}{q_i^0 + q_i^t} & \text{if } i \in U_{0t} \\ E_0 p_i^0 & \text{if } i \in U_{0\setminus t} = U_0 \setminus U_{0t} \\ E_t p_i^t & \text{if } i \in U_{t\setminus 0} = U_t \setminus U_{0t} \end{cases}$$

$$E_r = \sum_{i\in U_r} p_i q_i^r / \sum_{i\in U_r} p_i^r q_i^r \qquad \text{for } r = 0, t.$$

That the GK index reduces to a matched-universe index can be seen in the following:

$$\left(\sum_{i\in U_t} p_i^t q_i^t\right) E_t = E_0 \left(\sum_{i\in U_{0t}} \frac{p_i^0 q_i^0}{q_i^0 + q_i^t} q_i^t\right) + E_t \left(\sum_{i\in U_{0t}} \frac{p_i^t q_i^t}{q_i^0 + q_i^t} q_i^t + \sum_{U_{t\setminus 0}} p_i^t q_i^t\right)$$

$$\Rightarrow E_t \left(\sum_{i\in U_{0t}} \frac{q_i^0 q_i^t}{q_i^0 + q_i^t} p_i^t\right) = E_0 \left(\sum_{i\in U_{0t}} \frac{q_i^0 q_i^t}{q_i^0 + q_i^t} p_i^0\right).$$

3.2. The GUV Index Family

Removing the constant-value adjustment via $P^{0,r}$ in (3), one may let

$$p_i = \sum_{r \in R_i} p_i^r q_i^r / \sum_{r \in R_i} q_i^r. \tag{4}$$

By (2) this yields the Lehr index (Lehr, 1885, 39) as a bilateral index in a fixed universe, and a *modified* GK (MGK) index in a dynamic universe. More generally, replacing (4) by any suitable p_i yields a family of GUV indices:

$$P_{GUV}^{0,t} = \frac{\sum_{i \in U_t} p_i^t q_i^t / \sum_{i \in U_t} p_i q_i^t}{\sum_{i \in U_0} p_i^0 q_i^0 / \sum_{i \in U_0} p_i q_i^0} = V^{0,t} / Q_{RP}, \tag{5}$$

where $Q_{RP} = \sum_{i \in U_t} p_i q_i^t / \sum_{i \in U_0} p_i q_i^0$ can be formulated as a reference-price (RP) quantity index. Auer (2014) emphasises the interpretation of p_i as an adjustment factor which transforms the transaction quantities $\left(q_i^0, q_i^t\right)$ into the "intrinsic-worth units" $\left(\tilde{q}_i^0, \tilde{q}_i^t\right)$, where $\tilde{q}_i^r = p_i q_i^r$ for $r = 0, t$. What then matters to the resulting index is only the relevant ratios p_i/p_j for any $i \neq j$. As pointed out by Dalén (2001), if hedonic regression is used, the factor p_i could be determined based on the difference in characteristics between the item and the numeraire (or chosen reference item). In this way the GUV index family can incorporate the hedonic approach.

The test results for the GUV index are similar to the GK index except for the responsiveness. It does not satisfy the identity test T1 except when $R(0,t) = R_B$. It obviously satisfies the fixed-basket test T2. It satisfies the upper bound tests T3 and t3, provided $p_i^t \leq p_i \leq p_i^0$ for $i \in U_R$. Similarly, it satisfies the lower bound tests T4 and t4 provided $p_i^0 \leq p_i \leq p_i^t$ for $i \in U_R$. However, the bilateral GUV index fails the responsiveness test t5 only in the settings of tests t3, where $p_i = p_i^0 = p_i^t$ for $i \in U_0$ and $p_i = p_i^t$ for $i \in U_{t \setminus 0}$ due to $R = R_B$, such that

$$Q^{0,t} = \frac{\sum_{i \in U_t} q_i^t p_i}{\sum_{i \in U_0} q_i^0 p_i} = \frac{\sum_{i \in U_t} q_i^t p_i^t}{\sum_{i \in U_0} q_i^0 p_i^0} = V^{0,t} \quad \text{and} \quad P_{GUV}^{0,t} = 1,$$

regardless of $D(U_{t \setminus 0})$. Similarly in the setting of test t4.

3.3. The WGM Index Family

A WGM index does not have a direct connection to the expenditure ratio $V^{0,t}$ in general. Like the GUV index, it employs a reference price p_j, for $j \in U_R$, and is given by

$$P_{WGM}^{0,t} = \frac{\prod_{i \in U_t} \left(p_i^t / p_i\right)^{w_i^t}}{\prod_{i \in U_0} \left(p_i^0 / p_i\right)^{w_i^0}} = \left(\frac{\prod_{i \in U_t} \left(p_i^t\right)^{w_i^t}}{\prod_{i \in U_0} \left(p_i^0\right)^{w_i^0}} \right) / \left(\frac{\prod_{j \in U_t} p_j^{w_j^t}}{\prod_{j \in U_0} p_j^{w_j^0}} \right), \tag{6}$$

with the weights $\sum_{i \in U_t} w_i^t = 1$ and $\sum_{i \in U_0} w_i^0 = 1$. When $R = \{0, t\}$ and $U = U_0 = U_t$, setting $w_j^0 = w_j^t = \frac{1}{2}\left(q_j^0 p_j^0 / \sum_{i \in U} q_i^0 p_i^0 + q_j^t p_j^t / \sum_{i \in U} q_i^t p_i^t\right)$ reduces (6) to the Törnqvist

index. The TPD index (De Haan and Krsinich 2014) is given by

$$p_j = \prod_{r \in R_j} \left(\frac{p_j^r}{P^{0,r}} \right)^{\frac{w_j^r}{\sum_{b \in R_j} w_j^b}} \quad \text{and} \quad w_j^r = \frac{q_j^r p_j^r}{\sum_{i \in U_r} q_i^r p_i^r}.$$

Fattore (2010) considers the axiomatic properties of the geo-logarithmic family (GLF) of indices. The GLF index is a special case of the WGM index, with fixed universe $U_0 = U_t$ and the same weights at both periods 0 and t, i.e., $w_i = w_i^0 = w_i^t$, where w_i can depend on data at both 0 and t.

The WGM index (6) does not satisfy the identity test T1 except when $R(0,t) = R_B$, since otherwise one can not ensure $p_j = p_j^0 = p_j^t$ in a dynamic universe. It generally does not satisfy the fixed-basket test T2 due to the lack of direct connection to $V^{0,t}$. It satisfies the upper bound test T3, provided $p_i^t \le p_i \le p_i^0$, such that

$$P_{WGM}^{0,t} = \left(\prod_{i \in U_t} \left(p_i^t / p_i \right)^{w_i^t} \right) \left(\prod_{i \in U_0} \left(p_i / p_i^0 \right)^{w_i^0} \right) \le 1.$$

Similarly, it satisfies the lower bound test T3, provided $p_i^0 \le p_i \le p_i^t$. Under the settings of tests t3 and t4, we have $P_{WGM}^{0,t} = P_{WGM}^{0,t}(D_{0t}) = 1$, provided $p_j = p_j^0 = p_j^t$, such that it satisfies these tests while failing the responsiveness test t5 at the same time.

3.4. The GEKS Index

Provided $R(0,t) = R_M$ and $t \ge 2$, the GEKS index from 0 to r, for $0 < r \le t$, is given by

$$P_{GEKS}^{0,r} = \left(\prod_{s=0}^{t} P^{0,s} P^{s,r} \right)^{\frac{1}{t+1}} = \left((P^{0,r})^2 \prod_{s \ne 0,r} P^{0,s} P^{s,r} \right)^{\frac{1}{t+1}}. \tag{7}$$

For any $r < t$, it involves indirect comparisons via the periods outside $\{0, \ldots, r\}$. For example, if $t = 2$ and $r = 1$, we have $P^{0,1} = ((P^{0,1})^2 P^{0,2} P^{2,1})^{\frac{1}{3}}$ where both $P^{0,2}$ and $P^{2,1}$ are only available in period 2 but not period 1. Therefore, in practice, the disseminated GEKS index, denoted by $\hat{P}_{GEKS}^{0,r}$ is always the one with $r = t$ in (7). It is intransitive since, for any $0 < r < t$,

$$\hat{P}_{GEKS}^{0,t} = \left((P^{0,t})^2 \prod_{0<s<t} P^{0,s} P^{s,t} \right)^{\frac{1}{t+1}} = \left((P^{0,t})^2 P^{0,r} P^{r,t} \prod_{0<s<t; s \ne r} P^{0,s} P^{s,t} \right)^{\frac{1}{t+1}}$$

$$\ne \hat{P}_{GEKS}^{0,r} \hat{P}_{GEKS}^{r,t} = \left((P^{0,r})^2 \prod_{0<s<r} P^{0,s} P^{s,r} \right)^{\frac{1}{r+1}} \left((P^{r,t})^2 \prod_{r<s<t} P^{r,s} P^{s,t} \right)^{\frac{1}{t-r+1}},$$

where the set of reference periods is $R(0,t) = \{0, 1, \ldots, t\}$ for $\hat{P}_{GEKS}^{0,t}$, it is $R(0,r) = \{0, 1, \ldots, r\}$ for $\hat{P}_{GEKS}^{0,r}$ and $R(r,t) = \{r, r+1, \ldots, t\}$ for $\hat{P}_{GEKS}^{r,t}$, according to the default

choice of R_M. For example, for $t = 2$, the three GEKS indices are intransitive, where

$$\hat{P}^{0,2}_{GEKS} = ((P^{0,2})^2 P^{0,1} P^{1,2})^{\frac{1}{3}}, \quad \hat{P}^{0,1}_{GEKS} = ((P^{0,1})^2)^{\frac{1}{2}} = P^{0,1},$$

$$\hat{P}^{1,2}_{GEKS} = ((P^{1,2})^2)^{\frac{1}{2}} = P^{1,2}.$$

It is, of course, possible to use a different R_M, such as a 13-month window that moves with t. But the GEKS remains intransitive, because $P^{0,r}$ calculated with $R_M = \{r - 12, r - 11, \ldots, r\}$ is different from $P^{0,r}$ calculated with $R_M = \{t - 12, t - 11, \ldots, t\}$, provided $t - 12 \leq 0 < r < t$.

The components in (7) can be any bilateral time-reversible index. Now that the reference universe of the GEKS index by definition cannot be U_{R_B} for $t > 1$, it generally does not pass any other tests than the responsiveness test T5.

4. A Reference-Quantity-Price Index Family

None of the indices considered in Section 3 satisfies all the five tests proposed in this article. Two observations seem worth noting. First, a multilateral index generally does not satisfy the identity test T1 nor the fixed-basket test T2. However, we do not therefore conclude that a bilateral index is preferable to a multilateral index in practice, since none of them are perfect and it is possible to compensate for a small shortcoming in one respect with better properties in others. Second, there is a tension between the bound tests t3 and t4 on the one hand, and the responsiveness test t5 on the other hand. As a potential means to a resolution, we outline below a large index family, which includes the GUV index family as a subclass. Let

$$P^{0,t}_{RQP} = \left(P^{0,t}_{RQ}\right)^{1-\alpha} \left(P^{0,t}_{GUV}\right)^{\alpha}, \tag{8}$$

where α is a constant of choice, for $0 \leq \alpha \leq 1$, and $P^{0,t}_{GUV}$ is given by (5), and the *reference-quantity* index $P^{0,t}_{RQ}$ is given by

$$P^{0,t}_{RQ} = \sum_{i \in U_{0 \cup t}} q_i p^t_i / \sum_{i \in U_{0 \cup t}} q_i p^0_i, \tag{9}$$

where $U_{0 \cup t} = U_0 \cup U_t$, and q_i is a reference-quantity for $i \in U_{0 \cup t}$. We shall refer to (8) as the *reference-quantity-price (RQP)* index. It reduces to a GUV index if $\alpha = 1$.

Provided $0 < \alpha < 1$, an RQP index makes use of both reference quantities q_i and reference prices p_i. The expression (8) shows it as a weighted geometric mean of two price indices. It can equally be expressed as deflating the expenditure ratio $V^{0,t}$ by a weighted geometric mean of two quantity indices $Q^{0,t}_{RP}$ and $V^{0,t}/P^{0,t}_{RQ}$, i.e.,

$$P^{0,t}_{RQP} = V^{0,t} / \left[\left(Q^{0,t}_{RP}\right)^{\alpha} \left(V^{0,t}/P^{0,t}_{RQ}\right)^{1-\alpha} \right]$$

In particular, at $\alpha = 0.5$, the RQP index can be considered to generalise the Fisher index, defined in the special case of $U = U_0 = U_t = U_{0\cup t}$ and $R(0,t) = R_B$. That is,

$$P_L^{0,t} P_P^{0,t} = P_L^{0,t} V^{0,t} / Q_L^{0,t} = P_P^{0,t} V^{0,t} / Q_P^{0,t},$$

where $P_L^{0,t}$ is the Laspeyres price index given as $P_{RQ}^{0,t}$ with $q_i = q_i^0$, and $P_L^{0,t}$ is the Paasche price index given as $P_{RQ}^{0,t}$ with $q_i = q_i^t$, and $Q_L^{0,t}$ is the Laspeyres quantity index given as $Q_{RP}^{0,t}$ with $p_i = p_i^0$, and $Q_P^{0,t}$ is the Paasche quantity index given as $Q_{RP}^{0,t}$ with $p_i = p_i^t$.

There are many possible choices for the reference price in $P_{GUV}^{0,t}$ and the reference quantity in $P_{RQ}^{0,t}$. In the existing though limited studies and practices of $P_{GUV}^{0,t}$, the reference price p_i is usually set to be the unit-value price of item i over the chosen reference universe, that is calculated over the periods in which the item is available. However, in certain situations, one may instead consider using the introductory price or another representative price. When it comes to the reference quantity q_i in $P_{RQ}^{0,t}$, one can obviously extend the various arithmetic and geometric means defined for the fixed universe. Or, one may set q_i to be the ratio between the average expenditure of item i and the reference price p_i calculated for the GUV-counterpart. In particular, we believe it will be necessary to study these questions together with the formation of homogeneous products, which are defined at a level that is between the items identified by (GTIN, outlet) and the elementary aggregate. However, it is beyond the scope of this paper to address these issues.

The RQ index $P_{RQ}^{0,t}$ satisfies obviously the identity test T1. It satisfies the fixed-basket test T2 provided $R(0,t) = R_B$. It satisfies the responsiveness test T5, as long as $q_i > 0$ for $i \in U_{0\cup t} \backslash U_{0t}$. Moreover, it provides a means to resolve the tension between the bound tests t3 and t4 and the responsiveness test t5. To satisfy the test t5 in the setting of the test t3, where $U_{0\cup t} = U_0 \cup U_{t\backslash 0}$, $p_i^0 = p_i^t$ for $i \in U_0$ and $q_i^t > 0$ for $i \in U_{t\backslash 0}$, we require $P_{RQ}^{0,t} < 1$. Since $\sum_{U_0} q_i p_i^t = \sum_{U_0} q_i p_i^0$ regardless of the choice of q_i for $i \in U_0$, we need $\sum_{i\in U_{t\backslash 0}} q_i p_i^0 > \sum_{i\in U_{t\backslash 0}} q_i p_i^t$, given any choice of q_i for $i \in U_{t\backslash 0}$. This can be achieved by imputing a price $\hat{p}_i^0$, where $\hat{p}_i^0 > p_i^t$ for $i \in U_{t\backslash 0}$. Provided such $\hat{p}_i^0$, the imputed reference-quantity expenditure in period 0 would be higher than the reference-quantity expenditure in period t, that is

$$\sum_{U_0} q_i p_i^0 + \sum_{U_{t\backslash 0}} q_i \hat{p}_i^0 > \sum_{U_0} q_i p_i^0 + \sum_{U_{t\backslash 0}} q_i p_i^t = \sum_{U_0} q_i p_i^t + \sum_{U_{t\backslash 0}} q_i p_i^t.$$

It follows that the imputed $P_{RQ}^{0,t}$ is less than 1, which satisfies the upper bound test $t3$ and the responsiveness test t5 at the same time. Similarly, the imputed RQ index $P_{RQ}^{0,t}$ satisfies jointly the tests t4 and t5, provided $\hat{p}_i^t > p_i^0$ for $i \in U_{0\backslash t}$.

Imputation seems a natural remedy for the RQ index because, unlike the MGK, GUV or WGM index, it lacks otherwise a mechanism that accounts for the differing sizes of the item universes U_0 and U_t. The inclusion of $\sum_{U_{t\backslash 0}} q_i \hat{p}_i^0$ or $\sum_{U_{0\backslash t}} q_i \hat{p}_i^t$ can be considered as a means to incorporate a dynamic basket from the COGI perspective, or to align the utility over time from the COLI perspective. In the setting of test T3, where $p_i^t < p_i^0$ for at least some $i \in U_0$, we have $\sum_{U_0} q_i p_i^t < \sum_{U_0} q_i p_i^0$ regardless of the choice of q_i for $i \in U_0$. It follows that the imputed $P_{RQ}^{0,t}$ satisfies the test T3, provided any $\hat{p}_i^0 \geq p_i^t$ for $i \in U_{t\backslash 0}$,

including the choice of $\hat{p}_i^0 = p_i^t$. Similarly, it satisfies the lower bound test T4, provided $\hat{p}_i^t \geq p_i^0$ for $i \in U_{0\setminus t}$, including the choice of $\hat{p}_i^t = p_i^0$.

The test results of the RQP index can be deduced from those of $P_{GUV}^{0,t}$ and $P_{RQ}^{0,t}$. Thus, given a judicious choice of the imputed $P_{RQ}^{0,t}$, it can potentially satisfy all the five tests.

5. Concluding Remarks

The proposed set of tests provide a rigorous diagnostic for whether an index can be considered completely appropriate in a dynamic item universe, as well as pointing towards the directions of possible remedies. The RQP index family can potentially satisfy all the tests. It extends the GUV index family that has received much attention in the recent years. But more research is needed regarding the imputation method and the mixing weight α.

We reiterate that failing one or more tests does not in itself make an index unacceptable in practice, because not exactly satisfying a test does not mean that it is not satisfied approximately, and it is possible for an index to compensate for a small shortcoming in one respect with better properties in others. Moreover, the test approach does not directly provide the solutions to the many other choices one necessarily has to make in practice. These include the use of fixed base period versus moving base and indirect measurement of the short-term price index, the aggregation structure of the CPI including the formation of homogeneous products, the balance between automatic item-matching and manual intervention, the decision between bilateral and multilateral indices in a given CPI sub-universe, and so on. For these reasons, we believe it is important, in future research, to develop sensible *empirical criteria*, regarding when an index based on the unit value price data can be considered acceptable for practical purposes.

6. References

ABS. 2016. *Making Greater Use of Transactions Data to Compile the Consumer Price Index, Australia*. The Australian Bureau of Statistics (Catalogue No. 6401.0.60.003). Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0.60.003 (accessed June 2019).

Auer, von L. 2014. "The Generalized Unit Value Index Family." *Review of Income and Wealth* 60: 843–861. Doi: https://doi.org/10.1111/roiw.12042.

Balk, B.M. 1996. "A Comparison of Ten Methods for Multilateral International Price and Volume Comparison." *Journal of Official Statistics* 12: 199–222. Available at: https://search.proquest.com/docview/1266834049?pq-origsite=gscholar (accessed June 2019).

Balk, B.M. 2001. *Aggregation Methods in International Comparisons: What Have We Learned?* ERIM Report, Erasmus Research Institute of Management, Erasmus University Rotterdam. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=370897 (accessed June 2019).

Chessa, A.G. 2016. "A New Methodology for Processing Scanner Data in the Dutch CPI." *Eurostat review of National Accounts and Macroeconomic Indicators* 1: 49–69. Available at: https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2016-art2.pdf (accessed June 2019).

Chessa, A.G., J. Verburg, and L. Willenborg. 2017. *A Comparison of Price Index Methods for Scanner Data*. Paper presented at the fifteenth Ottawa Group meeting, 10–12 May 2017, Eltville, Germany. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/A%20comparison%20of%20price%20index%20methods%20for%20scanner%20data%20-Antonio%20Chessa,%20Johan%20Verburg,%20Leon%20Willenborg%20-Paper.pdf (accessed June 2019).

CPI Manual. 2004. *Consumer Price Index Manual: Theory and Practice*. International Labour Organization. Available at: http://www.ilo.org/public/english/bureau/stat/guides/cpi/index.htm (accessed June 2019).

Dalén, J. 2001. *Statistical Targets for Price Indexes in Dynamic Universes*. Paper presented at the sixth Ottawa Group meeting, 2–6 April 2001, Canberra, Australia. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+6/$file/2001%206th%20Meeting%20-%20Dal%C3%A9n%20J%C3%B6rgen%20-%20Statistical%20targets%20for%20price%20indexes%20in%20dynamic%20universes.pdf.

Dalén, J. 2017. *Unit Values in Scanner Data Some Operational Issues*. Paper presented at the fifteenth Ottawa Group meeting, 10–12 May 2017, Eltville, Germany. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Unit%20values%20in%20scanner%20data%20%E2%80%93%20some%20operational%20issues%20-%20J%C3%B6rgen%20Dal%C3%A9n%20-Paper.pdf (accessed June 2019).

De Haan, J. and F. Krsinich. 2014. "Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes." *Journal of Business & Economic Statistics* 32: 341–358. Doi: https://doi.org/10.1080/07350015.2014.880059.

De Haan, J. 2001. *Generalized Fisher Price Indexes and the Use of Scanner Data in the CPI*. Paper presented at the sixth Ottawa Group meeting, 2–6 April 2001, Canberra, Australia. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+6/$file/2001%206th%20Meeting%20-%20de%20Haan%20Jan%20-%20Generalised%20Fisher%20Price%20Indexes%20and%20the%20Use%20of%20Scanner%20Data%20in%20the%20CPI.pdf (accessed June 2019).

Diewert, E.W. 1999. "Axiomatic and Economic Approaches to International Comparisons." In *International and Interarea Comparisons of Income, Output, and Prices*, edited by A. Heston and R.E. Lipsey, 13–87. Studies in Income and Wealth, Vol. 61. Chicago: University of Chicago Press.

Diewert, E.W. and K.J. Fox. 2017. *Substitution Bias in Multilateral Methods for CPI Construction Using Scanner Data*. Paper presented at the fifteenth Ottawa Group meeting, 10–12 May 2017, Eltville, Germany. Available at: http://irs.princeton.edu/sites/irs/files/Diewert%20and%20Fox%20Substitution%20Bias%20and%20MultilateralMethodsForCPI_DP17-02_March23.pdf (accessed June 2019).

Fattore, M. 2010. "Axiomatic Properties of Geo-logarithmic Price Indices." *Journal of Econometrics* 156: 344–353. Doi: https://doi.org/10.1016/j.jeconom.2009.11.004.

Fisher, I. 1922. *The Making of Index Numbers*. Boston: Houghton-Mifflin.

Geary, R.C. 1958. "A Note on Comparisons of Exchange Rates and Purchasing Power Between Countries." *Journal of the Royal Statistical Society, Series A* 121: 97–99. Available at: https://www.jstor.org/stable/pdf/2342991.pdf (accessed June 2019).

Ivancic, L., K.J. Fox, and E.W. Diewert. 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.

Lehr, J. 1885. *Beitrüge zur Statistik der Preise insbesondere des Geldes und des Holzes*, F.D. Sauerländer Verlag, Frankfurt a. M.

Schultze, C.L. and C. Mackie. 2002. *At What Price?: Conceptualizing and Measuring Cost-of-Living and Price Indexes*. The National Academies Press. Doi: http://dx.doi.org/10.17226/10131.

Zhang, L.C., I. Johansen, and R. Nygaard. 2017. *Testing Unit Value Data Price Indices*. Paper presented at the fifteenth Ottawa Group meeting, 10–12 May 2017, Eltville, Germany. Available at: http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Testing%20unit%20value%20data%20price%20indices%20-%20Li-Chun%20Zhang,%20Ingvild%20Johansen,%20Ragnhild%20Nygaard%20-%20Paper.pdf (accessed June 2019).

Received November 2018
Revised March 2019
Accepted May 2019