NSM 2022
NORDIC STATISTICAL MEETING

Using Non-Survey Big Data to Improve the Quality of the Household Budget Survey

Marius Runningen Larsson[1], Statistics Norway, riu@ssb.no

Li-Chun Zhang, Statistics Norway

Abstract

The household budget survey (HBS) is resource heavy. Both in terms of resources used by the national statistical offices (NSO) and due to the household's response burden. The long duration and diary component make it prone to non-response errors and incorrect records. To improve the quality of the HBS we present an alternative method for collecting and processing household grocery expenditure. The method takes advantage of non-survey big data consisting of electronic grocery receipts and debit card transactions. The receipts are combined with their corresponding transactions using multiple key variables. This makes it possible to allocate receipts to households via de-identified administrative records. Using data containing more than half a billion receipts we were able to allocate approximately 70 percent of the receipts to households. The data covers 96 percent of the Norwegian grocery market for 2018. The integrated data can be used to improve the quality of the HBS. Either by replacing the food and non-alcoholic beverages diary component to reduce the response burden or as auxiliary information to improve the survey-based expenditure estimates. The method is transferable to countries where grocery transactions are mainly carried out with payment cards. High market concentration in a country's grocery market will increase the feasibility of the method.

Keywords: Non-survey big data, Household budget survey, transaction data, record linkage, model-based estimation

1. Introduction

The household budget survey (HBS) is used to measure the distribution of households' expenditure on goods and services. The results from the HBS are used by public authorities to measure distributional effects of tax changes. It serves as input for other

[1] Corresponding author

official statistics such as the consumer price index and is frequently used in research (Nygård, et al., (2019); Hansen, et al., (2008), Aasness, et al. (2003)).

The HBS is resource heavy, especially due to the response burden. The participants record their day-to-day expenditures over a two or one-week period in a diary and responds to a detailed questionnaire regarding expenditure on services and durable goods. The long duration of the survey reduces willingness to participate and induce high nonresponse rates. The manual entries in the diary are susceptible to deficient or omitted entries (Egge-Hoveid & Amdam, 2016). The HBS is also resource heavy for the national statistics office (NSO) (Holmøy & Lillegård, 2014). The manual processing of the diaries is associated with a considerable cost and can introduce challenges regarding accuracy, data quality and reproducibility (Egge-Hoveid & Amdam, 2016).

To improve the data collection process Statistics Norway (SSB) has been exploring ways to take advantage of transaction data from grocery stores (Fyrberg, et al., (2018); Holmberg, (2018); Egge-Hoveid & Amdam, (2016)). In 2020, SSB received electronic grocery receipts from three Norwegian grocery chains and debit card transactions from a provider of digital payment solutions (Linnerud & Egge-Hoveid, 2022). The purpose is to explore methods to remove the need for participants to record grocery expenditure, or to use it as auxiliary information to improve the survey-based expenditure estimates. As the receipts does not contain personal identifiable information, methods have been developed to link receipts with payment card transactions. This is the first step in linking grocery expenditure collected from non-survey big data to households via de-identified administrative records.

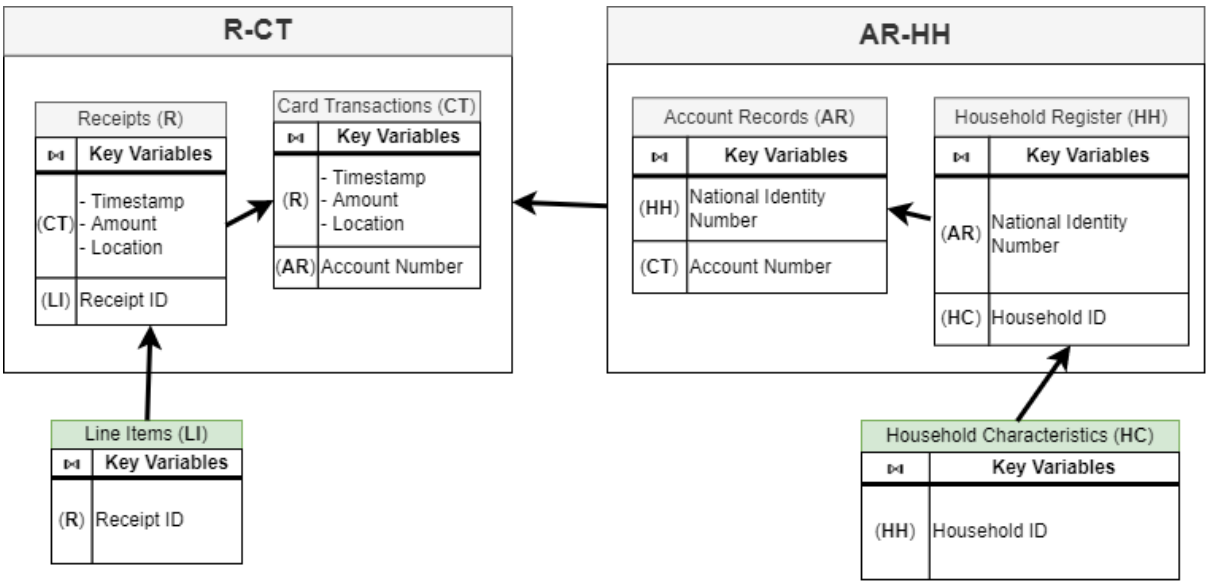
In this paper we present techniques for linking electronic grocery receipts to payment card transactions. We base our approach on proof of concept work done at SSB which identified three necessary linkage key variables: timestamp, amount and location (Fyrberg, et al., 2018). We evaluate our approach by applying it to data containing more than half a billion grocery receipts for the year 2018. In chapter 2 we start by presenting an overview over the linkages required to allocate grocery receipts to households. We continue by presenting in detail the approach and methods we use to achieve linkage between grocery receipts and payment card transactions. In chapter 3 we present the data we will use to measure the

performance of our approach. In Chapter 4 we present the linkage results. Chapter 5 contains a general discussion of the methods used and of the road ahead utilizing non-survey big data in the HBS.

2. Method

The process of linking grocery receipts to households is outlined in figure 1. We start by deriving a table of receipts (R) from the line items table (LI). LI contains variables such as item names, item prices, and EAN numbers – one row per item. In addition, it contains unique receipt ID, timestamp and total amount as repeated variables. As LI typically contains data from multiple stores belonging to a major chain it contains a variable with store name or ID. To derive R from LI we extract the *linkage key* variables $timestamp_R$, $amount_R$ and $location_R$. We're left with the table R where each row represents a single receipt. The item attributes from LI are kept separately from the linkage.

Figure 1 Linkage procedure: From grocery receipts and groceries to households



The next step is to link R with card transactions (CT) using the key variables. The CT table contains detailed information on all domestic debit card transactions and their corresponding account number. This includes $timestamp_{CT}$, $amount_{CT}$ and $location_{CT}$. Completing this linkage generates the R-CT table, which consists solely of the receipt ID and account number. Generating R-CT is the most challenging step

in the process. The methods utilized in this step will be discussed in detail later in this chapter.

In addition to R-CT we need to define a linkage between the account numbers in CT and the corresponding households. We do this by linking an account number table (AR) with the household register (HH). The AR table is a list of domestic account numbers and their owners, identified by their national identity number. The HH table is a list of all domestic households and their members, also identified by their national identity number. Linking AR and HH on the identity number we generate the AR-HH table which contains the account number and its corresponding household ID.

We can now link R-CT and AR-HH together using account number, generating the final table R-CT-AR-HH, or Y. Y contains all linked receipts, represented by their unique *receipt ID*, and their corresponding household ID's. We can link Y with auxiliary information tables. This includes variables such as household type (e.g. one-person, couple, etc.) from the household characteristics table (HC), and amount spent on items classified by Classification of Individual Consumption by Purpose (COICOP) from LI.

Record linkage

In the following we will distinguish between links and matches. Matches are two (or more) records from separate tables belonging to each other regardless of linkage. Links are established after the linkage procedure and might be a match or a nonmatch. In case the key variables in a linkage procedure form unique combinations and are free of errors, one can identify the matched records in two data files by comparing the associated key variables. In case of duplicated records in either data file, where records have the same key variables and need to be removed in advance, no comparison between the two files can resolve such duplicates. More critically, the key variables may be subject to various noises, so that two nonmatched records may appear to have the same key variables, and two matched records may appear to have different key variables. Linkage techniques that deal with the noises using an explicit statistical model are referred to as *probabilistic* (e.g. Fellegi & Sunter, (1969); Lee, et al., (2021)), whereas they are called *deterministic* if the noises are handled in a practical (if somewhat ad hoc) manner when making comparisons.

The two key variables $timestamp$ and $amount$ are essentially continuous, unlike $location$. Moreover, as we explain further below, different data sources have different noises which complicates statistical modelling. As we are not aware of any generic software for probabilistic record linkage based on continuous key variables, we develop deterministic linkage methods that are scalable to the present task.

Linking R with CT

Our goal is to use the key variables to generate R-CT by deterministic linkage, denoted as

$$R \bowtie_{\theta(\tau)} CT \tag{1}$$

where $\bowtie$ is the join operator, and θ is the conjunction rule given by

$$\theta = (amount_s = amount_z) \wedge (location_s = location_z) \wedge (timestamp_s = timestamp_z)$$

$$s \neq z \; and \; s, z \in \{R, CT\}$$

which consists solely of equality comparisons of the key variables. However, due to the noises in the key variables, the actual rule will require other operations and depend on some tuning parameters τ. In the upcoming subsections we present how we modify the comparisons and key variables in (1) to accommodate for this.

Amount

In CT, $amount_{CT}$ is represented by two variables, the transaction amount charged to the payment card and the actual purchase cost, denoted by $amount_{CT,f}$ where $f \in \{charged, cost\}$. For most cases these are identical, except for e.g., cash withdrawals where $amount_{CT,charged} > amount_{CT,cost}$. In R, $amount_{R,cost}$ is always available, either directly or by summing all line item prices for each receipt in LI when generating R. Amount charged by payment card can in some cases be present in R such that $amount_{R,f}$ depend on f in the same way as $amount_{CT,f}$.

There are advantages of using $amount_{R,charged}$over $amount_{R,cost}$. For example, if the purchase in R is paid by a combination of cash and debit card, only the card charge will be present in CT such that $amount_{R,cost} > amount_{R,charged} = amount_{CT,charged} = amount_{CT,cost}$. To link such cases using deterministic methods we preferably need $f = charge$ in both R and CT.

If only cost is available, the comparison between $amount_{CT,cost}$ and $amount_{R,cost}$ can be affected by noise. For example, when $amount_{R,cost}$ is calculated (instead of given directly), one must include all the relevant adjustments correctly, such as discount, recycling deposit and returned items. Rounding of the prices and total cost can be affected by different data types.

Moving forward, we will nevertheless use the equality comparison

$$amount_{R,f} = amount_{CT,f}$$

where f depends mainly on the data supplier of R.

Timestamp

There are generally two points in time associated with a sale: Start and end time. We denote this by $timestamp_{R,t}$ where $t \in \{start, end\}$. Start time corresponds to when the first item is scanned. End time is the moment after the transaction confirmation is received from the card terminal. Depending on the data supplier, R will contain start or end time, or both. As the sale is completed after payment confirmation, the relationship between timestamps in R and CT can be expressed as

$$timestamp_{R,start} < timestamp_{CT} < timestamp_{R,end}$$

Two examples are given in Table 1. The first sale has a duration of 30 seconds. It is initiated at 09:01:00 and concluded at 09:01:30. The timestamp in CT is 09:01:20. In the second example the duration is 81 seconds. Note that the time difference between $timestamp_{R,start}$ and $timestamp_{CT}$ vary drastically between the two examples, which depends on several factors (e.g. number of items to be scanned). The difference between $timestamp_{CT}$ and $timestamp_{R,end}$ is more stable.

Table 1

$timestamp_{R,start}$	$timestamp_{CT}$	$timestamp_{R,end}$	**Duration**
09:01:00 (20 sec)	09:01:20	09:01:30 (10 sec)	30 sec
09:01:00 (70 sec)	09:02:10	09:02:21 (11 sec)	81 sec

Time difference between R and CT in parenthesis. Dates are omitted for brevity.

Based on this we construct two inequalities for timestamp. First, if end time is available we have the following inequality comparison with slack:

$$timestamp_{CT} + \tau_+ > timestamp_{R,end}$$

The choice of τ_+ concerns two types of error: a larger value introduces more duplicates and possibly false links, a smaller value risks more missing matches. Setting an initial τ and updating it iteratively is a simple but effective way in determining a suitable (but not necessarily optimal) value.

Next, if only $timestamp_{R,start}$ is available, we apply the inequality comparison

$$timestamp_{CT} - \tau_- < timestamp_{R,start}$$

Due to the uncertain and potentially long time difference between $timestsamp_{R,start}$ and $timestamp_{CT}$ we must have that $\tau_- > \tau_+$.

Location

To reduce the total number of comparisons required, some key variables are used to *block* the files, so that comparisons are only made among the records belonging to the same block. Here, *location* and *date* (extracted from *timestamp)* are natural for blocking. As we have not managed to obtain the unique business ID for all the stores in R with corresponding values in CT, $location$ is also subject to noises.

Location is typically recorded by a store name or store ID. Table 2 provides examples. In the second case, both store names are available, so that they can be compared as strings (of alphabets). However, string comparison is of less help in the third case, where the trade name is in R, but the legal name is in CT, whereas it is infeasible in the first case, where we only have numeric ID in R but name in CT.

Table 2

$location_R$ – store name	$location_R$– store ID	$location_{CT}$– store name
-	100	Your Local Food Store
The Neighborhood Store	200	the neighbor store
The Cheapest Supermarket	300	Kari Nordmann AS

To solve this problem, we generate a store catalogue like table 2 for all unique store identifiers in R. To achieve this, we apply *distribution comparison* as follows. First, apply (1) with only $amount$ and $timestamp$ (with slack) as the key variables. For each store r in R, there will be many records in CT joined to it. Let n_r be the number of receipts in store r and let $n_{r,ct}$ be the number of card transactions to store ct (in CT) which are joined to *r* in this way. Let

$$location_r = location_{ct^*} \qquad where \qquad ct^* = \arg\max_{ct} \frac{n_{r,ct}}{n_r}$$

which is the $location$ (in CT) of store ct^* that has the most joined records. The idea is similar to deciphering a text written in permuted alphabets.

Table 3

	$n_{r,ct}/\mathrm{n}_r$ (for store ct in CT)				
Store r (in R)	Your Local Food Store	The Blue Flower Store	the neighbor store	The Neighborhood Store AS	Kari Nordmann AS
100	0.74	0.05			
150	0.68				
200			0.56	0.52	
300					0.85

Empty cells represent a link rate of 0

Table 3 illustrates the idea which shows a submatrix of the scores $n_{r,ct}/\mathrm{n}_r$. The primary situation of distribution comparison can be seen for store 100, for which the highest rate is 0.74 with *Your Local Food Store.* The joining rates with the other stores are considered to have occurred by pure chance (0.05 with *The Blue Flower Store)*. Notice that *Your Local Food Store* has a high rate both for stores 100 and 150, which can occur when a store changes its ID during the period in which the records are included for distribution comparison. Similarly, store 200 (in R) has a high rate with two stores in CT, which can occur if the latter changes its name during the

same period. This illustrates the need for using $date$ as a blocking variable jointly with $location$, by which means such complications can be eliminated.

We use $location_R^{DC}$ to denote the $location$ obtained for R by distribution comparison, which is now directly comparable to $location_{CT}$ in CT and can be used for blocking.

Final linkage

The deterministic linkage rule for (1) is now given as

$$\theta(f, \tau_+, \tau_-) = (amount_{R,f} = amount_{CT,f}) \wedge$$

$$\left(timestamp_{R,end} < timestamp_{CT} + \tau_+ \,\middle|\, timestamp_{R,start} > timestamp_{CT} - \tau_-\right) \wedge \quad (2)$$

$$(location_R^{DC} = location_{CT})$$

with tuning parameters (f, τ_+, τ_-), where | denotes disjunction.

When applying (2) multiple joins can occur when two or more receipts have identical amount in the same store inside the same time slack $\tau_+ \; or \; \tau_-$. Regardless how many receipts and card transactions this involves, one can either assign the links among them randomly, or treat it with a pseudo-Hungarian algorithm whereby the total time differences are minimised given the chosen links. Notice that such multiple joins are most likely to occur with the receipts consisting of few items, which are usually identical in terms of the item's COICOP classification. If this is the case the associated linkage errors do not affect the expenditure statistics and the receipts can be assigned at random.

Both the distribution comparison and the final linkage in (2) is implemented using PySpark, a Python API for Apache Spark (Zaharia, et al., 2016). The linkages generated for the distribution comparison is done on data containing a minimal number of days, but all unique store identifiers in R is present at least once (but with all their receipts that day) to reduce number of comparisons required.

3. Data

The receipt data is supplied by the three largest grocery chains in Norway covering approximately 96% of the total grocery market for the calendar year of 2018. The three chains supplied LI data in different formats and with different types of variables. When referring to the three different chains we will refer to them by the status of their

data rather than by their name. We convert each LI to three separate R^ks by extracting the relevant variables.

One supplier delivered data in a rawest form, which includes variables such as $amount_{R,charged}$ and $timestamp_{R,end}$. We will denote this source as raw. Data for January and most of February is missing from this supplier. The second supplier delivered a somewhat more processed LI table. It contains purchase cost (not charged) and end time. We will denote this source $medium$. The data from the final chain requires the most processing. Amount is not directly given, and the timestamp is start time. We will denote this one as $processed$.

All suppliers delivered at least one variable to uniquely identify their stores, which are handled by distribution comparison. See Table 4 for a summary of the different variables for each $R^k, k \in \{processed, medium, raw\}$.

Table 4

Table / Key	$R^{processed}$	R^{medium}	R^{raw}	CT
Time	$timestamp_{R,start}$	$timestamp_{R,end}$	$timestamp_{R,end}$ $timestamp_{R,start}$	$timestasmp_{CT}$
Amount	$amount_{R,cost}$	$amount_{R,cost}$	$amount_{R,charged}$ $amount_{R,cost}$	$amount_{CT,charged}$ $amount_{CT,cost}$
Location	$location_R^{DC}$	$location_R^{DC}$	$location_R^{DC}$	$location_{CT}$

The total number of unique stores across all R^k is approximately 3 500. The total number of unique receipts is well above half a billion which puts the number of line items in LI in the billions. We are unable to handle a small number of stores by distribution comparison, mainly due to missing unique identifier in CT (for unknow reasons). We leave out all the stores for which we do not obtain $location_R^{DC}$. These account for 0.15% of all receipts.

The payment card transaction data CT is supplied by a single supplier covering all domestic debit card transactions for 2018 (not only grocery stores). The number of transactions for the entire year surpasses 1.5 billion. After generating the store

catalogue, we remove all transactions from CT not pertaining to grocery receipts. This leaves CT approximately 75% the size of R in terms of number of transactions /receipts (averaged over all R^k). In other words, the maximum share of receipts we can link is approximately 75%. The remaining 25% can be attributed to other types of payment methods.

4. Results

As quality measures of the linkage we consider two types of linkage errors. *False linkage* is the case if a linked pair of records are not a correct match, whereas *missing match* is the case if a matched pair of records are not linked. Since we can be quite confident at identifying the transactions admitted in CT, which only involve grocery purchases and each transaction is assumed to have a corresponding receipt, an operation strategy is to aim at linking as many as possible of the records in CT while keeping the false linkage error as low as possible.

Let N_{CT} be the total number of grocery card transactions in CT. It is assumed to be the number of matches between R and CT. Let $N_{CT,1}$be the total number of card transactions to which exactly one linked receipt is obtained directly by a given $\theta(f,\tau)$ --- these are all considered to be correct links. Let $N_{\mathrm{CT},>1}$be the total number of card transactions to which a linked receipt is obtained from multiple joins (i.e. possible links) --- these are the total number of possible false links. An upper bound of *missing match rate (MMR)* is

$$MMR = 1 - \frac{N_{CT,1}}{N_{CT}}$$

The closer MMR is to zero, the closer we are to full linkage. MMR will have a negative relationship with τ if an increase in τ leads to $\Delta N_{CT,1} > \Delta N_{CT,>1}$.

The upper bound of *false linkage rate (FLR)* is by definition

$$FLR = \frac{N_{CT,>1}}{N_{CT,1} + N_{CT,>1}}$$

To make it easier to compare the magnitude of the two types of error, we instead use an upper bound of the *proportion of false links (PFL)*, which is given by

$$PFL = \frac{N_{CT,>1}}{N_{CT}}$$

The relationship between τ and PFL is positive if $\Delta N_{CT,>1} > 0$. As with MMR, we want PFL to be as low as possible.

The results of the linkage can be seen in Tables 5 for each of the data suppliers. All numbers are for the whole year. The results vary by $\tau = (\tau_+, \tau_-)$ in (2), where τ_+ is used for R^{medium} and R^{raw}, and τ_- for $R^{processed}$.

For all R^k we have a positive relationship between τ and PFL. This is expected as an increased interval around $timestamp$ will increase the probability of a payment transaction being linked with two or more receipts. The relationship between τ and MMR is negative for all but R^{medium}. This indicates that the increased number of linkages for R^{medium} gained by an increase in τ is offset by the increase in false links. Out of the three τ values tested, linkage for R^{medium} has the best performance at $\tau = 15$.

Table 5

	(τ_+, τ_-)	$R^{processed}$	R^{medium}	R^{raw}
MMR	$(15, 75)$	0.1324	0.0205	0.6795
	$(30,150)$	0.0461	0.0206	0.0048
	$(60, 300)$	0.0306	0.0210	0.0038
PFL	$(15, 75)$	0.00315	0.00034	0.00006
	$(30,150)$	0.00462	0.00143	0.00069
	$(60, 300)$	0.00694	0.00220	0.0017

The highest MMR values can be found at $(\tau_+, \tau_-) = (15, 75)$ for $R^{processed}$ and R^{raw}. Increasing τ to 30 for R^{raw} greatly reduce MMR. A further increase to 60 only see a marginal MMR change, but PFL more than doubling. In figure 2 we can see the distribution of time differences ($timestamp_R - timestamp_{CT}$) for linked transactions in

$N_{CT,1}$. It corresponds well with what we observe in table 5 and explains the sharp decrease in MMR for R^{raw} when we increase τ_+ from 15 to 30. For $R^{processed}$ we have a steady decrease in MMR for increased τ. Despite this, MMR for $R^{processed}$ at the highest τ level is still greater than for R^{medium} and R^{raw}.

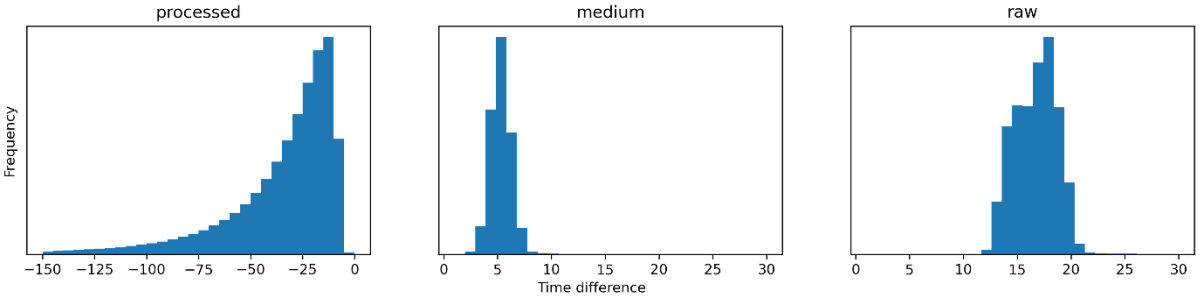


Figure 2 Time difference between $timestamp_R$ and $timestamp_{CT}$ for linked transactions in $N_{CT,1}$. Single month (τ_+, τ_-) = (30, 150)

Contrary to what one could be led to believe without running the linkage empirically, it is clear that τ cannot be too small. Furthermore, τ must be set separately for the different R^ks, even those containing $timestamp_{R,end}$. The different stores in each R^k could also have different delays between payment time and end time for various reasons. This could explain why we are able to link more than 30% (i.e., $1 - MMR - PFL$) of the transactions with R^{raw} despite setting $\tau = 15$. Since it is impossible for us to obtain and investigate all the details that could explain the distributional differences across the sources, the linkage method needs to be configured statistically based on the data obtained.

The linkage between R-CT and AC-HH (to obtain Y) can be directly based on the account numbers, although perfect linkage is still not possible due to missing account numbers for some of the household members (due to multiple reasons, e.g. debit card registered to businesses). Initial tests indicate that we manage to link approximately 90%-95% of all transactions to their respective households. This is not investigated further in this paper.

5. Discussion

In this paper we have presented an approach for linking grocery receipts to payment card transactions with the purpose of allocating grocery expenditure to households in the HBS. For most intents and purposes, we achieve acceptable missing match rates

indicating that the approach is well suited for this type of application. There are perhaps several tweaks one could implement to reduce the noises even further to improve the results. However, the potential gains to the match rate by increasing the timestamp slack might very well be offset by an increase in false linkages. The trade-off between reduced MMR and increased PFL must be evaluated in light of the statistics the NSO wish to produce. If most transactions included in PFL have duplicate linkages for receipts with identical COICOP categories, trivial deduplication methods can be utilized to generate suitable data for the HBS. By comparing the linkage between the different supplied datasets, we show that data quality and relevant variables are most important for the linkage using our methods – The linkage of R^{raw} is close to saturated. We suspect the linkage of R^{raw} outperforms the other two because it is the only one containing $amount_{R,charged}$. It is doubtful that more complicated probabilistic methods could have improved the results.

To our knowledge, the approach outlined in this paper has not previously been presented in the literature making it difficult to compare with other results. Proof of concept (POC) work conducted at SSB have shown that a similar approach can generate acceptable linkage on a single day of data for one chain (Fyrberg, et al., 2018). The linkage result of the POC is similar to the results from $R^{processed}$ presented here. However, the key variables in the POC are prepared using more ad hoc methods which might not generalize, since their data quality are considerably worse than the data in our $R^{processed}$. We believe the methods presented in this paper, all based on the key variables identified in the POC, are more general and easier to implement. As an example, solely using distribution comparison to generate match between location compared to string comparison is more general and allows for the usage of de-identified store ID's to increase confidentiality.

Since we only have access to debit card transactions administered through the supplier of the CT table, the approach does not cover the receipts paid for by customer accounts, cash, and credit cards. Furthermore, online grocery shopping is expected to increase in popularity. This might lead to a more fragmented grocery market in the future. Collecting receipts from independent and foreign stores is infeasible. We do not know the extent to which people purchase groceries for other households than their own.

Model-based estimation methods will be necessary to replace the current diary scheme by transaction data for grocery expenditure. Collecting both diary data and transaction data for the HBS 2022 allows us to use the diary sample as an audit sample to assess the accuracy of such model-based estimates (Zhang, 2021)(a). The linked dataset can be used as auxiliary information for purely survey-based estimations improving the expenditure estimations (Zhang, 2021)(b).

The methods we present should be of interest to other NSOs, particularly in the Nordic countries. In general, the Nordic countries have grocery markets with relatively high market concentration and share of bank card payments. This makes it more feasible to gather data compared to countries with a more fragmented market. A more thorough discussion surrounding the practical and legal difficulties for an NSO to retrieve these types of data can be found in Linnerud & Egge-Hoveid (2022).

Finally, a brief discussion on confidentiality is in order. Record linkage allows us to derive detailed information on households' purchases. It is critical for the NSO and the data providers that both commercial interest and data confidentiality are protected. The data, both in its raw and linked status, should be treated as highly sensitive. At SSB, Multiple steps are being taken to safeguard confidentiality. Several additional (and in an NSO perspective perhaps radical) measures are currently being evaluated to further increase data security (Zhang & Haraldsen, 2022).

References

Berg, N. & Seferi, G., 2022. *What is the effect og digitalisation in household surveys in official statistics? - A descriptive study and a preliminary assessment of HBS 2022, Q1.* Reykjavik, Iceland, Hagstofa Íslands Nordic Statistical Meeting.

Egge-Hoveid, K. & Amdam, S., 2016. *Redeveloping the Norwegian Household Budget Survey.* Madrid, s.n.

Fellegi, I. & Sunter, A., 1969. A theory of record linkage. *Journal of the American Statistical Association*, Issue 328, pp. 1183-1210.

Fyrberg, J. et al., 2018. *Proof of concept - transaction records as a new source for Statistics,* Oslo: Statistics Norway Internal Notes.

Hansen, K., Lian, B., Nesbakken, R. & Thoresen, T. O., 2008. *LOTTE-Skatt-en mikrosimuleringsmodell for beregning av direkte skatter for personer,* s.l.: Statistisk sentralbyrå.

Holmberg, A., 2018. *Herding and exploring combinations of electronic transaction data for alternative HBS-design purposes.* s.l., s.n.

Holmøy, A. & Lillegård, M., 2014. *Forbruksundersøkelsen 2012. Dokumentasjonsrapport. Notater 2014/17.,* s.l.: Statistics Norway.

Horvitz, D. G. & Thompson, D. J., 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, Issue 260, pp. 663-685.

Jentoft, S., Toth, B. & Muller, D., 2022. *From manual to machine: Challenges in machine learning for COICOP coding.* Reykjavik, Iceland, Hagstofa Íslands Nordic Statistical Meeting.

Kuhn, H. W., 1955. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, p. 83–97.

Lee, D., Zhang, L.-C. & Kim, J. K., 2021. Maximum entropy classification for record linkage. *Survey Methodology*, pp. 1-23.

Linnerud, K. & Egge-Hoveid, K., 2022. *Big data for HBS - Gains and lessons learned.* Reykjavik, Iceland, Hagstofa Íslands Nordic Statistical Meeting.

Nygård, O. E., Slemrod, J. & Thoresen, T. O., 2019. Distributional implications of joint tax evasion. *The Economic Journal*, Issue 620.

Zaharia, M. et al., 2016. Apache spark: a unified engine for big data processing. *Communications of the ACM*, pp. pp.56-65.

Zhang, L., 2021a. Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pp. pp.571-588.

Zhang, L. C., 2021b. *Estimation and inference for expenditure statistics,* Oslo: Statistics Norway Internal Notes.

Zhang, L.-C. & Haraldsen, G., 2022. Secure big data collection and processing: framework, means and opportunities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* .

Aasness, J., Biørn, E. & Skjerpen, T., 2003. Distribution of preferences and measurement errors in a disaggregated expenditure. *The Econometrics Journal*, Issue 2, pp. 374-400.