Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Herding and exploring combinations of electronic transaction data for alternative HBS-design purposes

Anders Holmberg[1]

Abstract

When Statistics Norway cancelled its 2018 Household Budget Survey it was because of cost/quality trade-off concerns and hesitation whether an essentially traditional survey-based diary approach would be satisfactory. The decision to cancel, kicked-off intensified investigations to acquire alternative data sources for household consumption data. The national transaction data landscape can therefore now be well described, and three different sources of electronic transaction data as well as how they can be combined are being examined on experimental basis. One source is transaction data from a key payment provider in Norway (card transactions and other electronic transactions, also B2B transactions). The coverage rate is substantial and covers most card transactions done in Norway. We also investigate cash register data of from retail chains and samples of retail chain membership (loyalty card) data. All these data sources are interesting by themselves but finding ways to combine them is what really adds value, at least from a household consumption perspective. E.g. methods for record linking card transactions with cash register data to simultaneously retrieve both the demographic dimension and the detailed consumption dimension down to detailed levels of Classification of Individual Consumption by Purpose (COICOP). The paper discusses the possibilities and the methodological and technical experiences made from this work up till this date (mid-2018).

Key Words: Combining data sources; Household consumption; Electronic payment transactions.

1. Introduction

1.1 Background

Surveys where the data collection rely on the respondents keeping and reporting a diary have always had a big focus on motivating and maintaining respondent participation. (e.g. see Edgar et al. (2013)). Compared with surveys that use a conventional format of measurement instrument the response burden is usually higher (both from a volume and frequency point of view). The respondents are expected to, not only remember, record and report a variety of facts about different topics, events and activities, they are also asked to do so consistently over a longer period. In such circumstances it is natural to suppose a higher risk for nonresponse and more measurement errors. This in turn drives costs because of necessary quality assurance efforts.

A family of important societal surveys that traditionally have this characteristic and suffers the dilemma are household surveys that collect information about incomes and consumption. In some countries they are referred to as a Household Budget Survey (HBS). In other countries it is only the consumption part that is relevant for (direct) data collection, hence the name Household Expenditure Survey (HES). The latter is the focus in this paper. I will describe the explorative work done by Statistics Norway, to identify, acquire and evaluate data about households' expenditures that are tracked by electronic financial transactions and how they may be fit for use for statistical purposes.

1.2 Statistics Norway's Survey of Consumer Expenditure

In Norway, Statistics Norway (SSB) has since 1958 produced statistics about household consumption using a sample survey called FBU (short for Forbruksundersøkelsen). The main purpose of the first surveys done in 1958, 1967 and 1973 was to get detailed estimates of private consumption to update the weights of the Consumer Price Index. From

[1]Anders Holmberg, Statistics Norway, Postboks: 2633 St. Hanshaugen, Norway, 0131 Oslo

1974 to 2009 the FBU was conducted annually, and an additional aim was to monitor the consumption pattern of different household categories. By 2010 a two-year break was introduced until 2012 and today the FBU 2012 is the most recent statistics published by SSB.

Over time the basic design of the FBU have stayed relatively unchanged. The FBU2012 had as input a questionnaire, a diary (paper and electronic option) and receipts from cash-registers. The gross sample size was 7000 households. Two interviews were done, one initial and one final interview after a 14-day diary period. The interviews were conducted both by visits and telephone during a 15-month field period. The nonresponse rate was 51 percent. This reflects an increasing trend. In 1983 the nonresponse rate in the FBU was 33 percent, (Holmøy and Lillegård, 2014).

After the FBU2012 SSB initiated work to develop and modernise the survey. The target was to launch a modernised FBU in 2017 but it was later postponed to 2018. The main delivery from this modernisation work was improvements of the data collection design. A digitalized self-administrated solution for the respondents' diary reporting and increased automatic coding of receipts were suggested to provide good enough quality and keep costs down. Nevertheless, this was not convincing enough. In late 2016 SSB's board of directors decided to cancel the FBU2018 and relocate funds to other organisational projects. The next FBU is tentatively set to be done in 2022.

1.3 Search for Alternative Data Sources

The modernising project of the FBU also pursued alternative data sources to assess household expenditure. SSB tried to use its legislative power to acquire *loyalty membership data* from daily retail chains. This process was stalled in negotiations with the providers. After three years, not until 2017, was a set of test data delivered to SSB for statistical exploration purposes. This was too late to have any impact on the abovementioned cancellation decision. Furthermore, an evaluation and analysis of the potential using such data to get representative household expenditure statistics has not been promising. SSB has instead turned to other data sources. From the daily retail trade SSB have focused on *cash register / scanner data* from digitalised sales transactions.

The transaction data from cash register sales are very detailed for every sales event. It contains the sold commodities, prices, volume, point of sale and time etc., but it lacks information about the consumer and household. However, SSB have looked at other sources and another way where this deficiency can be overcome.

When investigating the financial transaction environment in Norway it became apparent that *electronic payment transactions* from banks, financial institutions and other enterprises engaged in payment services could be accessible for statistics. There is one central custodian of such data in Norway. Although it is not straightforward this greatly simplifies accessibility. The data involve electronic transactions made by individuals and firms and has information on transaction level about amount, type of transaction, date, and parties. While the cash register data are rich in detail and have questionable coverage, the payment transaction data do not have much detail, but they cover (in theory) all electronic payment transactions done in Norway. In the following, the idea that these sources can complement each other to get household expenditure statistics of daily retail goods is investigated.

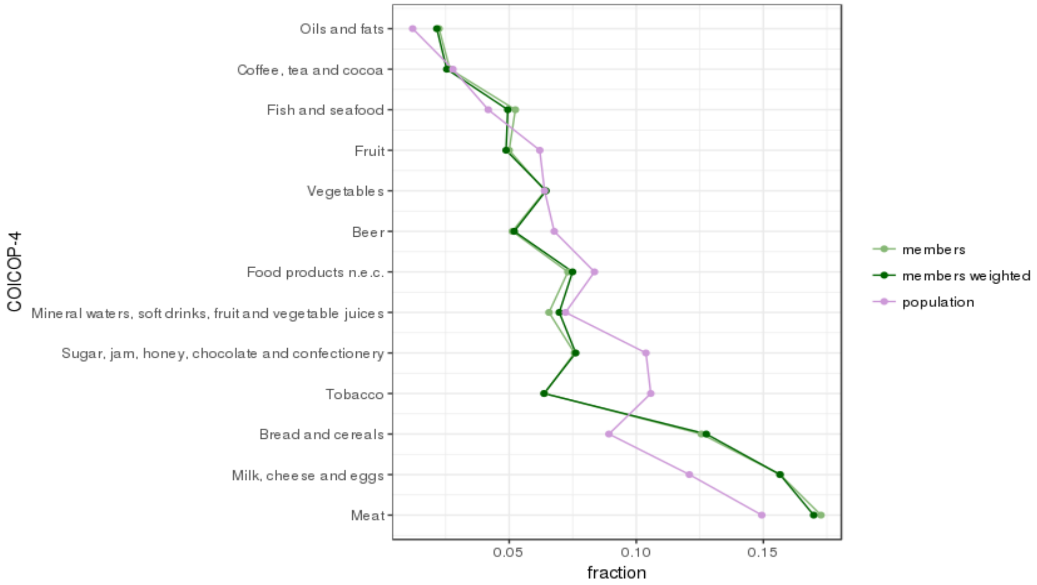
2. Transaction Data Exploration

2.1 Experiences with Loyalty Cards Data and Cash Register Data Acquisition

The Norwegian daily retail trade is dominated by three actors. SSB have invested considerable effort to establish a constructive dialogue to get access to their data. As an example, cash register data has been available on aggregated level for Consumer Price Index purposes for a long time. In recent years the retailers have built loyalty member systems with electronic cards and customer data bases that register the members' use of product offers and purchase patterns. Ideally, these databases would be a mirror of the members' and consumers' expenditure patterns. This is the background why the data could be useful household expenditure statistics.

Statistics Norway have analyzed data from one of these databases. All purchases registered by loyalty member cards during one month in 2016 was studied with respect to demographic and purchasing patterns and compared to purchases by non-members (Buelens et. al. 2018). The conclusions from the study were that there are significant issues with representativeness. The member demographics and the retail expenditure patterns revealed many significant deviations from other benchmark statistics. Attempts to correct for these deviations with auxiliary information from the Norwegian population register were fruitless, and the study concluded that differences could not be explained by age, gender and location. Moreover, certain Classification of Individual Consumption by Purpose (COICOP) products have wider sales points than just the stores of the retails chains. As an example, this is illustrated below by the low tobacco proportion in figure 2.1-1.

Figure 2.1-1
Expenditure distribution by COICOP for loyalty card members and CPI-based population expenditure.



Given the substantial selection biases observed in the loyalty member data and that considerable further investment is needed to acquire data tailored for production, SSB has halted any further pursuit of these kind of data. Instead, it was decided to concentrate on and do deeper analysis of cash register data.

From one of the main retail chains a calendar month of all registered sales were retrieved consisting of about 220 million records. The full receipts with records have a time stamp, the ID of the store, the GTIN code (Global Trade Item Number) of the sold item, price of the item, as well as total price sum. This information is used for the linking exploration described in section 2.3.

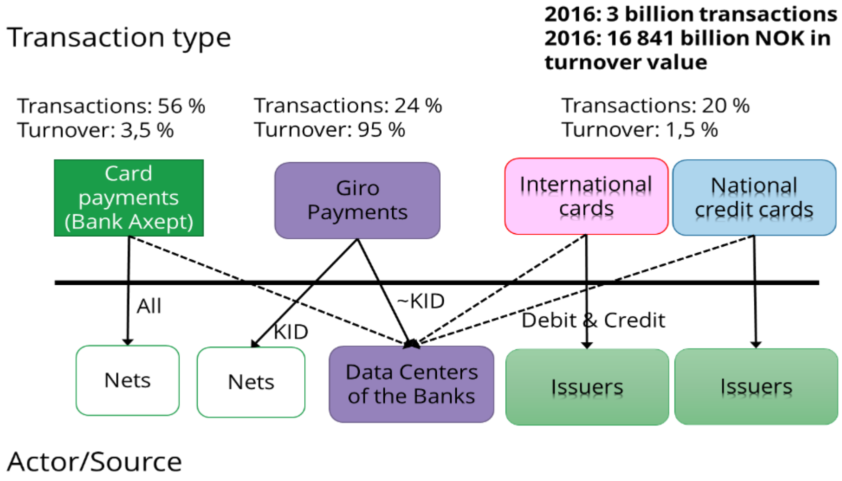
2.2 Payment Transactions

There are different types of electronic financial transactions. In this context the interest lies in those who are payments from consumers. Business to business transactions, salary transfers from a business to an employee or transfers between bank accounts without an underlying purchase are out of scope. Although they have a wider interest to make other statistics.

Figure 2.2-1 is an overview of transaction categories and actors in Norway. There are four main types, (i) debit card transactions via the national Bankaxept system, (ii) giro-transactions made between bank accounts and credit card transactions, either issued for (iii) Norwegians or as a (iv) foreign credit card. All transaction types go through and leave traces in the data centers of the banks who are active in Norway. The credit card transactions go to the credit card issuers. Giro payments can be done with or without a KID-code. The KID-code is a customer identification code that simplifies the handling of invoices by automatically tracking who has paid.

Nets is a provider for payment services that receive all giro-transactions with KID-code and every transaction that uses Bankaxept, i.e. payments done with Norwegian debit cards. Of all electronic financial transactions, Nets handles approximately 98,5 % of the monetary value and 80 % of the estimated 3 billion yearly transactions.

Figure 2.2-1
Overview of the Norwegian Financial Transaction System and the distribution of number of transaction and transaction values in 2016



SSB established contact and requested the delivery of a one-month test data from Nets. This is only possible under the statistics law. The reference period matches the loyalty card data and the cash register data retrieved from the retail chains. It is Nets' Bankaxept data that is of main interest for consumer to business payments. The transaction records of card payments contain a time stamp down to seconds, information about the owner of the Bankaxept terminal (the business) and its location, the account from which the payment is made and the transaction amount.

2.3 Exploration of Linking the Transaction sources and household demographics

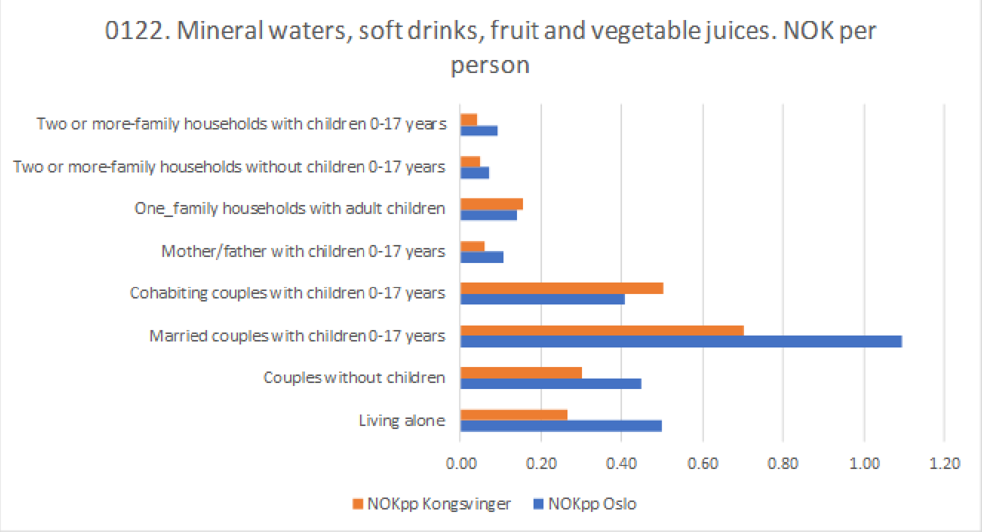
Fyrberg et. al. (2018) describe a proof of concept which was done to see whether the cash register data and the payment transactions combined could be used to make household expenditure statistics. To get there a sequence of linking operations and confidentiality preserving routines must be performed.

The unit in the payment transaction and the cash register data is the monetary transaction event. The sender and a receiver information are attributes. In household expenditure statistics the unit of interest is the household. Therefore, a connection between the households and the transaction data must be established. SSB can do so because of its statistical use of the population register, business register and a register of bank accounts. The bank account register is a listing of legal account owners. By utilizing this information for linking, pseudonymizing personal and household information, the Bankaxept payment transaction data can be linked to the account owner making the payments and labeled with a household category. Any detailed error analysis of this procedure has not been done. Neither for the linking or the household categorization. However, the linking is done with personal ID numbers and it is the same procedure to categorize households that is used in the Norwegian register-based census. Hence, it has been considered reliable with respect to household statistics before. The extent of whether people lend their debit cards to non-household members or buy goods for other households are unknown and a matter for further studies.

The abovementioned procedure connects household categories to payment transactions as expenditure events. The next step is to add detailed information about the expenditure. In the proof of concept, the cash register data from the daily retail trade was then linked. This was done by record linking the transaction data from one day. To create the linked records *time*, *place of sale* and *total sale sum*, was used. This method resulted in that 66 percent of all cash register transactions could be matched to an account transaction and labeled with a household category. Previous studies have shown that about 70 percent of all purchases in Norway are made with debit cards using Bankaxept, so that matching rate is expected. The other 34 percent of non-matched cash register transactions are payments done with credit cards, cash and cards issued by the retail chain. (Possible differences between the set of matched and the set of un-matched transactions are being studied.)

The matched dataset for this single trading day and single retail chain contained circa 775 000 unique purchases with an average of just above 6 items per purchase and a total value of 155 million NOK. Every good has an EAN code enabling linking to the COICOP classification. Figure 2.3-1 illustrates the level of detail that can be given when processing the resulting dataset.

Figure 2.3-1
Estimate of Daily Expenditure of COICOP 0122 per Person in Oslo and Kongsvinger by Household Type



Since it was out of scope, the proof of concept never investigated the quality of statistics such as the one illustrated above. That would also have been very difficult because the level of detail simply has not been accessible before. Note the small domain aspect with statistics from a small commune such as Kongsvinger by household category, the sample size and the old design of the FBU is not even close to allow decent accuracy for such estimates.

3. Summary and Further work

3.1 Discussion

The cancellation of the FBU sparked an effort to search for and herd alternative data into Statistics Norway for investigating different ways to do the survey. So far, the experiences are promising. New and previously unknown data sources have been discovered and explored. The idea of using data from loyalty member cards have been put on hold in favor of pursuing a general idea of combining different transaction data sources with each other, and with available administrative data. A proof of concept study on transaction data has encouraged SSB to carry on. Especially since the uncovered data sources have potential use in other statistics than just for household expenditure. Potential for new health and nutrient statistics, for price statistics and for looking into business to business transactions for structural business statistics, are just some of the spin-offs that have emerged.

In terms of household expenditure statistics more development work remains. Not just to attain more knowledge from the new data sources, but also to decide on a general design to make statistics. Thus far one conclusion is that by adding more data from the other main actors in daily retail trade and adding more days; it is likely that household expenditure statistics about daily retail products would be better than that of the old FBU. As far as errors are concerned, the nonresponse and measurement errors in a diary-based FBU, would be replaced by linking errors, unit errors (see Zhang 2012) and possibly issues with representativeness if payment habits change, or with new big actors on the market. The accuracy, actuality and analysis potential would be improved immensely with the transaction data.

However, compared to the previous purpose of the FBU the coverage of different expenditure types is poor. As an example, the daily retail trade actors do not cover goods such as alcohol and electronics very well. For each such poorly covered group of goods, there may be a big hurdle to climb to access relevant transaction data. It is doubtful if such investments are worthwhile just for household expenditure statistics. Every new data source must be negotiated, processed and combined. Technical solutions and a secure data delivery must be put in place and maintained and if there is a lack of standardization, which it often is, it could be costly to set up separate lines of input production systems for several data providers. Although Norwegian transaction data also may be able to pick up signals from other larger expenditures such as cost for housing, it appears farfetched and unwise that a household expenditure survey should be based on electronic transactions alone. At least if the content demands are similar those as in the old FBU.

Instead, it is appealing to reflect on a survey design that is a hybrid between a traditional survey and the electronic transaction data. The transaction data can capture a tremendous accuracy and detail for certain types of expenditures. E.g. the daily retail goods and other regular expenditures that is identifiable and can be assigned to households. Then a complementary sample survey can be designed to assemble information about other necessary expenditure types that are not covered. To keep costs down the level of detail needed for those expenditure groups must be lower than those covered by the transaction data. Given what we know about measurement errors when respondents are keeping diaries or answering memory recall type of questions; and given that one of the cost driving factors is the data collection. It seems little point of having a diary in such a complementary survey if the detailed daily goods are determined by transactions. The respondent contact should focus on acquiring information that cannot be obtained from the transaction data and to check whether other necessary assumptions about expenditure distributions among households are reasonable. This could for example be the expenditure proportions for leisure activities, travel etc.

SSB will continue to look at the transaction data to make household expenditure statistics. It is too early to say if the next published statistics will be based on transaction data only, a hybrid between transaction data and a traditional survey or on a modernized FBU survey. Personally, I believe that if the survey is decided the diary approach must be taken out or seriously reformed. In the light of what we know we can get from alternative sources, there should be no or a different role for the diary than as previously improving/assuring detailed and timely reporting.

On the other hand, if the electronic transactions become the main data source, then it is also necessary to ascertain and control that coverage is sufficient, representativeness in terms of household units linked to transactions, and in terms of expenditure types. If not, the complexity and cost driving processes to supplement transactions with other data may very well be a worse solution than a traditional survey. In Europe there is a plan to legally regulate household budget statistics within the European Statistical System. Depending on how the regulation is written it will have impact on the way SSB will act. The higher the requirements are on more details about other expenditures than daily retail goods, the less likely it is that SSB will set up a statistical design with many combined sources.

References

Buelens, B., S. Amdam, and H. Holgersen (2018), “The use of loyalty card transactions data in Household Budget Statistics: an exploration”, paper presented at the European Conference on Quality in Official Statistics, 26-29 June 2018, Krakow, Poland.

Edgar, J., D. Nelson, L. Paszkiewicz, and A. Safir (2013), “The Gemini Project to Redesign the Consumer Expenditure Survey: Redesign Proposal”, Project report 26 June, Bureau of Labour Statistics.

Fyrberg, J., J. Zhiyang, J. Åmberg, A. Vestfossen, H. Grini, and A. Frøberg (2018), “Proof of concept – linking payment transaction data with purchase transaction data”, unpublished report, Oslo, Norway: Statistics Norway.

Holmøy, A., and M. Lillegård (1988), “Forbruksundersøkelsen 2012: Dokumentasjonsrapport”, *Documents 2014/17*, Statistics Norway.

Zhang, L.-C. (2012), “Topics of statistical theory for register-based statistics and data integration”, *Statistica Neerlandica*, 66(1), pp. 41-63.