Towards an integrated Consumer Expenditure Survey -- Combining Multi-mode Data Collection and Big Data Extracts

Gustav Haraldsen (gha@ssb.no), Sverre Amdam¹, Li-Chun Zhang²

**Keywords**: Multi-mode, Transaction data, Big data, Response burden, Consumer Expenditure

1. **INTRODUCTION**

The Consumer Expenditure/Household Budget Survey is traditionally based on meticulous diary reports over a 14 days period, combined with an initial and concluding interview. The survey has suffered a high response burden and low response rate; typically about 50%. Interview costs and coding based on hand written specification or enclosed receipts have also made the survey expensive. Several redesign options have been suggested for this kind of surveys (e.g. Cantor, Mathiowetz et al. 2013). Our principal approach in the 2017-survey will be to combine digital transaction data with different kinds of questionnaires. In addition we will try to extract consumer patterns from databases kept by grocery retailers and link these results to the survey sample, e.g. to counteract bias because of nonresponse or to replace the most burdensome part of the diary.

2. **METHODS**

2.1. **Multi-mode Survey Design**

Figure 1 depicts a multi-mode design where data from different inputs are collected into a common self-completion web questionnaire and, when completed, data are transferred to the statistical office for statistical processing. The core element in this design is a common metadata-base that both define the response categories in the questionnaire and the deliveries from different inputs. The most noticeable input source in this design is e-receipts from debit and credit card transactions that are picked up from the shops payment systems and electronically transferred into the respondents’ diary of daily expenditures (1). The technology for this is developed and will soon be tested in a pilot study. In a similar way we will try to collect e-bank statements covering invoice payments (2).

---

¹ Statistics Norway
² University of Southampton/Statistics Norway
We also envisage that some diary data can be entered by a tablet questionnaire (3) or scanned from smartphones (4). In addition to these input sources, some data may be collected by interviewers (5) or from paper diaries and questionnaires that are scanned by delivery (6). For respondents who are unable or unwilling to conduct the survey using digital inputs 1-4, input 5-6 (telephone and paper diary) will replace these.

2.2. Big Data Extracts

The multi-mode design will reduce the response burden for those who allow us to automatically collect transactions. These records will also be more accurate. Self-completion and digitalized data will reduce costs. But for those respondents who cannot or do not want to give us access to electronic transactions, the consumer expenditure surveys will still be a heavy burden. And there will still be unit and item nonresponse. We will try to meet these challenges with an alternative approach to sample-based statistics. Instead of generalizing from a small number of observations, as we do in surveys, we will try to extract expenditure consumption patterns from large data files and link these results to the same groups as those generated from the survey data.

The most burdensome part of the diary completion is to report daily grocery shopping. In Norway most of the grocery market is split between three major grocery retailer chains (GRC). Two of these, covering approximately 75% of the grocery market, offer loyalty programs which record what members buy in their shops. When accumulated these files form a rich source of data. By linking social characteristics to those present in the retailer’s database, the statistical office can produce statistics analogue to that produced from survey samples. The data can be accumulated over time to avoid seasonal variation associated with sample diary report. The large amount of available data also means that sampling uncertainty is a lesser error in this source. The key quality issues in this kind of data are rather selectivity and potential coverage errors. The loyal GRC customers is unlikely a simple random sample of the target population of the Expenditure Survey. Not all the household members of the loyal GRC customers may be found in the same database. Retail at GRC may not cover the whole range of retail expenditure. Methodological issues to be explored are alternative possibilities of detecting and adjusting for such errors. The principal alternatives can be distinguished by whether they require linkage between the Expenditure Sample and the GRC databases or not.
Both methods have strengths and weaknesses which have to be dealt with, but potentially survey data and big data extracts supplement each other.

3. FUNDAMENTALS

Receipts electronically captured from shops are already offered to customers of a digital post box in Norway, and Statistics Norway is able to use the same technology. One important constraint, however, is the number of shops which have taken the system in use. In particular, the coverage of GRC’s is still low.

Only just above 4% of the expenditures of Norwegian households are in cash (Norges Bank 2014). To leave this out of a consumer expenditure survey seems to be a minor problem. We are, however, dependent on an informed consent from the respondents in order to download their transaction data. The number of respondents who are willing to give us access to digital transaction data, opposed to those that want to conduct the survey in the original diary/interview form have a lot to say for both the business case of multi-mode design, as well as the quality of the data.

We have surveyed individual’s willingness to let Statistics Norway access debit/credit card data, as well as data from bank accounts in a limited reporting period. The purpose of the study was to get a better understanding of people’s attitudes towards disclosing this kind of personal financial information for statistical purposes, as well as understanding what affects the decision to participate. The survey was included as a part of Statistic Norway’s omnibus survey in July/August 2014, which had a net sample of 1082 respondents.

The results showed that 40% of those asked said they were willing to give SSB access to bank transactions. Only age had a significant impact on the willingness, where younger age groups were more willing to disclose debit/credit card information, as well as bank account transactions. Surprisingly people’s internet experience did not seem to have a significant impact on willingness in the survey. An Integrated Expenditure Survey (IES), which also makes use of GRC database extracts, will have important differences from a modernised mixed-mode expenditure survey (MES) across the production processes, including sampling design, questionnaire design, editing and estimation. For instance, loyal GRC customers may be over-represented in the sample. The questionnaire for these sample households will focus more on supplementary retail data, and other information related to potential coverage errors. The editing and estimation may no longer be based on sample weighting alone, and statistical model-based methods will be necessary.

4. SUMMARY

The technology and the widespread use of credit/debit card in the population form a basis for a multi-mode consumer expenditure survey primarily based on digital transactions. The main challenges are to sell in the technology to shops, particularly GRCs, as well as sell in the transaction option to those sampled. As an alternative we are therefore looking into ways to extract consumptions patterns from data files established by customers’ loyalty programs.

REFERENCES