Use of new data sources – the case of Statistics Norway

Hans Viggo SÆBØ and Xeni Kristine DIMAKOS
Statistics Norway

Integration of privately held data on third parties in the production of official statistics is one of the main challenges for modernisation and quality improvements of such statistics. Several initiatives in the international statistical community have been taken to facilitate access and use of such data, also commonly denoted big data. Proper statistical legislation is one of the preconditions for such access. In Norway, a new Statistics Act was adopted in 2019. The act emphasises quality requirements to official statistics, the coordination role of Statistics Norway for all such statistics which are described in a multiannual programme, and the access to privately held data for use in official statistics. For data access, use of receipts from grocery stores linked to bank transaction data represents a test of the new statistics act and the GDPR. Such data are very relevant for the household budget survey, a survey which places a heavy response-burden on the participants, leading to quality challenges. The paper summarises some of the Statistics Norway's experiences with accessing privately held data, exemplified by the work to acquire data to develop a new household budget survey. Key aspects in addition to the legal ones are cost-benefit analyses, collaboration, privacy protection and methodological developments. The result of Statistics Norway's work to collect and use receipts and bank transactions is still unclear. However, experiences on the process so far should anyway be useful beyond the national level.

1. Introduction

Access to data to produce official statistics is one of the pillars in the work of a national statistical institute (NSI). Such access has been granted by statistical legislation.

In the former Norwegian Statistics Act from 1989 it was stated that Statistics Norway shall have the right to use administrative data-processing systems in the state administration and in nationwide municipal organisations as a basis for official statistics. In Norway a population register with a national identity number was established in 1964. Archive statistics based on this, and other public registers were gradually developed, also based on a register of properties formally established in 1980 and a register of legal entities established in 1995. Today, to a large extent, Norwegian official statistics are based on administrative data systems or registers. Statistics Norway uses more than 100 such registers from about 30 public institutions as a basis for its production of statistics. Statistics

Norway has agreements of cooperation with these institutions, and structured quality reports exist for all registers used for its production of statistics.

However, the data revolution characterised by the abundance of data, data providers, new technology and private holders of huge volumes of new data about other persons, legal entities and events, has challenged the role of official statistics and the NSIs, see Sæbø and Hoel [1]. Statistics based on big data may be more relevant than existing official statistics by describing new phenomena, increasing timeliness and the level of details. To improve the quality of official statistics it is necessary for the NSIs to access and reuse these data. A new statistics act described in chapter 2 should ensure this. Proper legal basis is one prerequisite for the NSIs to obtain such privately held data.

Statistics Norway has during the past few years been in dialog with different owners of privately held data. One example has been the acquirement of purchase receipt data and debit card data for use in the household budget survey (HBS). Other applications of new data comprise the use of scanner data for the consumer price index (CPI), electricity statistics based on data from an electricity data hub with information from all metering points, including "smart meters". Statistics Norway has little practical experience of using position data from Mobile Network Operators (MNO) to produce official or experimental statistics. However, some methodological work has been done and Statistics Norway is part of a consortium with several European NSIs that will cooperate in a three year Eurostat grant project on on the use of MNO data in official statistics.

The paper summarises the experiences with the access to privately held data so far, exemplified by the work to develop a new household budget survey. Key aspects in addition to the legal ones are cost-benefit analyses, collaboration, and privacy protection. As a result of our experiences and in particular the General Data Protection Regulation (GDPR) data minimization requirement, a methodological framework for big data is under development, and an outline is presented in this paper.

An overview of the work with the HBS up to the summer of 2022 is given by Linnerud and Egge-Hoveid [2].

2. New statistics act

A new Norwegian Statistics Act was adopted in 2019 and came fully into force in 2021 [3]. The act emphasises quality requirements for official statistics, the coordination role of Statistics Norway for all such statistics, which shall be described in a multiannual programme, and the access to privately held data for use in official statistics.

In the new act official statistics are limited to those statistics described in the national programme. The first programme which covers the period 2021 – 2023 has been prepared by Statistics Norway in cooperation with other producers of such statistics [4]. A new programme covering the period 2024 – 2027 is expected to be adopted by the Government in 2023.

The former Norwegian statistics act from 1989 in principle ensures the access to privately held data given that secrecy is safeguarded, by a formulation that any person must provide the information which is necessary to produce official statistics. In the new act from 2019 the following sentence has been added for emphasising access to third-party data: "The obligation includes data on the party with a duty to report, and other data for which the party has a right of disposal".

One of the main changes in the new act on access to data is the inclusion of a formulation on the need to carry out a cost-benefit analysis: "Statistics Norway shall not make a decision to impose a duty to provide information until an assessment has been made of the benefit of obtaining the information, balanced against the costs incurred by the party with a duty to provide information and a determination of the extent to which the processing will impact on the data subject. The assessment shall be made public". A template for such a cost-benefit analysis has been developed, see chapter 3.

3. Cost-benefit analysis

The standard cost-benefit analysis contains the following items in addition to information on responsible entity, data provider, and description of the data in question including reporting frequency:

1. Anchoring in the national statistics program.
2. Benefits: Justify why the information is necessary for the development of new or improved statistics, or the preparation or dissemination of official statistics.
3. Give an assessment of whether Statistics Norway can achieve its purpose by using information that is available from public authorities (in case the obligation is imposed on a non-public entity).
4. Costs: Consider what burden the obligation to provide information will entail for the person or companies obliged to provide information: Consider the costs in finding and preparing the information they must report, and the costs of the reporting itself. If the respondent must establish or change technical solutions to be able to report in the desired way, this must also be included in the assessment.
5. Describe the use and processing of personal data (if relevant), and why Statistics Norway in case needs directly identifying information.

6. Sensitivity of wanted information.
7. Security measures: Describe any special information security measures beyond the general measures in Statistics Norway.
8. Justification for the scope of data required: Describe why the information that the task givers must report is necessary and relevant (data minimization).
9. Secondary use: Possible limitations on secondary use if the information has commercial value, is copyright protected or could pose a danger to the security of the kingdom.
10. Methodology: Describe the assessment of the use of algorithms, methods, etc.
11. Summary of the overall cost/benefit assessment and conclusion.

4. Experiences with access and use of transaction data

Since 2017 Statistics Norway has had a process of gaining access to transaction data from grocery stores (receipts) and banks (use of debit cards) to be used as a data source in the household budget statistics. Statistics Norway has received test data, tested methods, and assessed data quality for the use of transaction data in various statistics. In the spring of 2022, technical infrastructure for streaming transaction data in real time in large quantities had been established and tested. Between 2 and 3 million receipts from grocery stores were transferred to Statistics Norway daily.

Receipts and bank transactions are linked applying information on time, location and sum. Bank transactions will be linked with a public registers of bank accounts (tax directorate) and a household register based on the central population register. Thus, links between the grocery receipts and households are established. For analyses information from other public registers such as registers of education and income will be applied. The granularity and the amount of data is high, which emphasise the importance of applying privacy enhancing techniques and methods to secure disclosure controls. In addition to pseudonymization other methods has been developed such as indirect links, aggregations and sampling, to reduce the privacy consequences. Transparency of the methods used to ensure confidentiality is emphasised, see chapter 5.2.

In addition to household budget statistics, the same transaction data are crucial for new statistics on diets of the Norwegian households. There is a potential for extended or further use of similar data within areas such as consumer price indices, nutrition statistics, indices of wholesale and retail sales, statistics on business activities, establishments, enterprises, transport and tourism, accounts and private health services.

4.1 Household budget statistics

Statistics Norway (SSB) conducted its first Household Budget Survey (HBS) in 1958. Until 2009, data were collected and published annually. The survey was last conducted in 2012. The purpose of

HBS is to provide a detailed picture of Norwegian households' annual consumption of goods and services. The household budget statistics are important to monitor changes in consumer behaviour, as a weighting basis for the CPI and for the national accounts. It is used to analyse the distributional effects of tax changes and to monitor development in the Norwegian diet. It is also a source for setting the size of social benefits and child support.

Until 2012, data on household consumption were collected through sample surveys only. The response burden and the survey costs where high and sample bias and underreporting led to quality challenges. The COICOP group one, Food and non-alcoholic beverages, are especially vulnerable for low response burden and underreporting due to the high number of items to report.

As mentioned, Statistics Norway plans to use receipt data from grocery stores and bank transactions from banks to measure expenditure for food and non-alcoholic beverages in the HBS. The plan at this stage is to combine this with a more traditional survey for other goods and services.

4.2 Process

Experiments with the receipts data from grocery chains and bank transfer data started in 2017. Main steps or events the last two years have been:

- *Spring 2022*: Decision from Statistics Norway to impose duty to provide data according to the Statistics Act, valid for 2022 and 2023. Cost-benefit analyses for both the access to receipts and for the bank transactions to be used in the HBS were published on Statistics Norway's website ssb.no.
- *June 2022:* The decision was appealed to the Ministry of Finance (appeal body for decisions on the obligation to provide information pursuant to the Statistics Act) by two out of four grocery chains and by Nets Branch Norway representing the banks. Privacy concerns and lack of legal basis to impose such an intrusive duty to provide information are the main objections, in addition to requirement for data minimization.

  The streaming of data from the grocery chains was put on hold during the appeal process. Media has also shown interest (see below). In parallel with the appeal to the Ministry of Finance, the Norwegian Data Protection Authority started supervision of Statistics Norway and requested more information on possible privacy implications with reference to the Personal Data Act/GDPR. They requested more information on data minimization than what is covered in the cost-benefit analysis, better overview of other data sources that are relevant to link the data in question with, future storage, and possible limitations on information from children. Statistics Norway gave a written explanation with emphasis on purpose, privacy protection, data minimization and storage.

- *August 2022:* Statistics Norway had a meeting with the Norwegian Data Protection Authority discussing the necessity and proportionality of applying the relevant data sources for the purpose of official statistics. Statistics Norway has reassessed all aspects of the cases considering the appeals from the grocery chains and Nets Branch Norway. In accordance with Norwegian Public Administration Act the appeals and Statistics Norway`s assessment of the statements has been sent to the superior department, in this case the Ministry of Finance. The ministry is the appellate instance for decisions on the duty to provide information pursuant to the Statistics Act.
- *November 2022*: The Norwegian Data Protection Authority notified a possible prohibition of the proposed data processing of receipts, claiming that decisions of Statistics Norway does not satisfy the requirements of the GDPR concerning the lawfulness of processing the data concerned. Statistics Norway was given two months to comment on the notice. Statistics Norway disagreed with the notified decision.
- *January 2023*: Statistics Norway commented on the notified decision and has asked for further dialogue with the Data Protection Authority.
- *April 2023*: The Data Protection Agency maintained the prohibition of the proposed data collection and processing of receipts. Since it is too late to use this data source for the HBS for 2022 anyway, Statistics Norway did not appeal the decision, but claimed that we disagree and asked for a dialogue. Statistics Norway will follow the case further as a matter of principle regarding our role and mandate and the practice of the Statistics Act.

Before and during the process with the Norwegian Data Protection Authority Statistics Norway has worked on and considered methodology for privacy enhancing. An overview of status is given in chapter 5.

4.3 Public reactions

In addition to the reactions from the involved parties and the Norwegian Data Protection Authority the case has been mentioned in the media, including the Norwegian Broadcasting, several newspapers also abroad and of course in social media. In general, it seems that an impression has been created that Statistics Norway and other public authorities will monitor individuals and their consumption. The message that this is only for official statistics and that no one will be able to see the (pseudomized) individual data is difficult to get through.

However, Statistics Norway welcomes the public debate and works to explain that statistics are aggregated data which cannot be used for surveillance of individuals. An NSI neither will nor can share data on persons for commercial purposes, like the private companies in question.

4.4 Technical solutions and privacy

The large amount of data from the new sources (receipts from grocery stores, bank transactions) requires new systems for receiving, storing and editing data. It is estimated that SSB will receive 1.3 billion receipts and 1.6 billion payment transactions per year. This makes it challenging to carry out statistical production in an environment based on SSB's existing "on-premises" setup.

A strongly connected and secure infrastructure is needed. SSB has developed a new cloud-based services/platform as a part of a bigger modernization process in the office. The new data platform manages all types of data that Statistics Norway collect, treat and disseminate as statistical products. Google Cloud Platform (GCP) is the service provider.

Receipt data are not personally identifiable. Only after receipts are linked to card transactions by using information on time, location and sum, they should be considered as direct personal identifiable information. Debit card transactions are personally identifiable. All data on the platform are encrypted and all direct personal identifying information is pseudonymized.

5. A methodological framework

In anticipation of potential complications with access to privately held data, SSB are exploring how privacy enhancing techniques could be developed and implemented for big data in general and the HBS in particular. Privacy preserving techniques is a growing field, both in and outside of official statistics. The methods outlined below are not exhaustive but may be suitable for Statistics Norway's statistical production.

The methodological framework under development comprises of three categories:

- Data minimization by sampling (big data sets)
- Improving confidentiality in the processing and production of statistics.
- Minimization of data storage.

Within each of these main categories there are several possible approaches. Some of these involve increased requirements for the data owner, for instance that the data owner must be responsible for performing various operations on data. Below, each category is discussed briefly.

5.1 Data minimization by sampling

The traditional and last HBS in 2012 was based on a sample of 7000 households. However, given proper technical solutions, transferring all data to Statistics Norway is cost efficient, both for the data

providers and for SSB. It is possible to draw a sample after the linkage, but that would not eliminate the privacy concerns.

In principle, data minimization can be done both outside Statistics Norway or “inside” by "throwing away" all data immediately after linking and selecting a suitable sample. However, the legal assessments of the Data Protection Authority may imply that the latter is not sufficient. If so, samples must be drawn before data reach Statistics Norway. Drawing samples outside the NSI imposes a burden on the data owner, as technical solutions that performs the chosen sampling strategy needs to be incorporated in their data systems. Also, there are quality issues that must be addressed and solved. If the selection is done outside of Statistics Norway, systems for quality assurance and documentation must be implemented to ensure that the data received have been created in line with the sampling plan and the specific design. The systems must also ensure the desired degree of reproducibility.

Alternative sampling strategies are:

- Nano: Sampling is done at the level of an event below the level of a statistical unit. In the case of the HBS, nano sampling implies sampling receipts and bank transactions. As there are millions of receipts and transactions, an independent Bernoulli sampling with probability 0.1 for both the receipts and bank transactions, would give a sample of approximately 1 percentage of all linked records (0,1 x 0,1). This sample would still be around 300 times larger than the actual sample of the HBS (7000 Norwegian households).
- Micro: Sampling is done at the level of the statistical unit, in the case of the HBS, the household. With this approach, even with an informed consent of a sample of households, Statistics Norway would need to process all receipts and bank transfers. Once the link is established, receipts and transactions that do not belong to the sample can be deleted.
- Cluster: Sampling is done at a level above the statistical unit (household). In the case of HBS this could be a sample of grocery stores and selected dates. It could also follow a rotational plan through all stores, equalizing the response burden. This would require the stores to extract the sample receipts from their systems. Statistics Norway still needs to receive and process all bank transactions to establish the link to households.

5.2 Improving confidentiality in processing

By using methods for data minimization (section above), as well as traditional pseudonymization of personal identifiers, the individuals’ or households’ attributes (variables) are in theory available during the production process. Increased confidentiality in the production of statistics can be achieved by separating the processing of individuals' identifiers and their attributes in the production process also after pseudonymization as described by Zhang & Haraldsen (2022) [5]. A

simplified version of this method was implemented in Statistics Norway using the test transaction data that were collected in the autumn of 2021. A project has been launched to implement a generic solution of the same methodology that would allow it to be adopted more generally in Statistics Norway. Different forms of access management are other measures in this category.

5.3 Minimization of storage

Official statistics needs to be well documented and verifiable following amongst other the principles of the European Statistics Code of practice. Which data that is stored after the statistics have been produced can in various ways be minimized. How exactly the minimization should be done depends on which statistics are produced, the methods used in the production and the degree of reproducibility required. Principles of storage needs to balance confidentiality protection and possible future needs for reproducing or development of the statistics in question. Among other things, it must be considered what identifiers need to be stored (pseudonymized, de-identified, anonymous?) and how aggregated the stored data can be. Statistics Norway is currently in the process of developing for a minimization storage framework for big data to address these issues.

6. Quality aspects

It is believed that overall use of the transaction data will improve the quality of the HBS with regard to accuracy and in particular timeliness (e.g. Zhang, 2021) [6]. Besides, such use will reduce response burden and other costs. However, there are also quality challenges linked to use of such data.

As for accuracy, coverage and hence representativity of data is an issue. The market for groceries in Norway is unique, with only four large chains representing approx. 98 per cent of the market. Of all receipts, debit card transactions (which are the ones Statistics Norway receives) constitute almost 75 per cent. By linking grocery receipts to bank transactions and bank accounts, approximately 70 -75 per cent of the receipts to persons and households could be connected, see Runningen Larsson and Zhang [7]. Purchases made by credit cards, cash, via customer clubs and other payment methods cannot be linked so far. This is not regarded to have a significant impact on the accuracy of the resulting statistics today.

However, comparability of data and hence statistics over time may be a challenge. To which extent is the availability and quality of external data sets beyond the control of the statistical organisation guaranteed over time? There is uncertainty connected to the future representativity of debit card transactions. Payment methods are complicated and evolving fast. New constellations and payment

methods are continuously introduced to the market, e.g., Apple pay, Google pay and customer cards. These are all payment methods that cannot be tracked by receiving debit card bank transaction data.

7. Challenges

7.1 Data and statistics

Data is not the same as statistics, though statistics are also data. Statistics are numerical information relating to an aggregate of data on units or observations. There is a classic way of ordering statistics above data on the road to knowledge, i.e., statistics is closer to decisions than data. However, today the concept of data is widened and dominates the public discourse, the age of statistics is being replaced by the data era, [1] [8]. The emergence of data science as a discipline might have contributed to this.

The conflation of data and statistics is probably one of the reasons why it is challenging to explain to the public that official statistics do not threaten privacy.

7.2 Information security and protection of privacy

Information security and protection of privacy is crucial in the production of official statistics. It comprises having adequate security measures in place and applying appropriate statistical disclosure policies. All quality frameworks for such statistics comprise this as principles and requirements. In the recently adopted core values of official statistics one of six such values is "Respects confidentiality" [9]. Technical solutions may still be a challenge. In the case of privately held data the granularity of privately held data may be extraordinarily high, and data points may refer to single events, transactions, encounters or movements. Sound disclosure prevention methods must be tuned to the specifics of the data source.

However, NSIs have extensive experiences in processing of personal data necessary for statistics, data based on both surveys and administrative registers. In many cases, data from different sources are linked. Still, improvements in both technical solutions and methods to enhance privacy are crucial. Some alternative promising methods are outlined in chapter 5.

7.3 Communication

It is obvious a challenge to communicate the need for reusing privately held data for official statistics, both the difference between data and statistics and the challenges linked to big amounts of almost real time data and the protection of privacy.

Statistics Norway works to establish better understanding of safety and privacy concerns. It is crucial to make it clear that statistics are a purpose for gathering data that is not intrusive. It is also relevant to mention that a national statistical institute already has vast amounts of data stemming from public registers, including linked data.

Openness is probably a key, that is also the reason to have cost-benefits analyses available for everyone.

8. Conclusions

Privately held new data pose both opportunities and challenges for the NSIs. Such data will transform both the way official statistics are produced as well as the statistical products. Statistics could be produced more cost efficient and with lower burden on respondents, and NSIs could respond better to user demands for more relevant and timelier statistics of higher quality.

A modern legal basis is a prerequisite for ensuring statistical institutes' access to privately held data for the development and production of official statistics, but far from enough. There are still obstacles such as perceived and real risks linked to data security and scepticism from private data owners and in media. The correspondence between statistical legislation and the GDPR can be debated, and what constitutes sufficient data minimization depends on discretion anyway. Good technical and new methodological solutions are necessary, and open communication of purpose, advantages and processes is crucial. Collaboration with stakeholders and in particular data holders, is essential. The processes towards satisfactory solutions may take time, but that is well justified.

9. References

[1] Sæbø, H.V. and Hoel, M.: *Official statistics: Quo vadis?* European conference on quality in official statistics, Vilnius 7 – 9 June 2022. Available at: https://q2022.stat.gov.lt/scientific-information/papers-presentations/session-37

[2] Linnerud, K. and Egge-Hoveid, K.: *Big Data for HBS – Gains and Lessons Learned.* Nordic statistical meeting, Reykjavik, 23 – 24 August 2022. Available at: https://www.nsm2022.is/facts-on-the-fly

[3] Statistics Act (2019), Act relating to official statistics and Statistics Norway (the Statistics Act). Available at: https://www.ssb.no/en/omssb/ssbs-virksomhet/styringsdokumenter/statistikkloven

[4] Statistics Norway: *National programme for official statistics 2021 – 2023*. Available at: https://www.ssb.no/en/omssb/nasjonalt-program-for-offisiell-statistikk

[5] Zhang, L.-C. and Haraldsen, G. (2022). Secure big data collection and processing: Framework, means and opportunities. J R Stat Soc Series A. 2022:1541-1559.

[6] Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. J R Stat Soc Series A. 2021: 571-588.

[7] Runningen Larsson, M. and Zhang, L.: *Using Non-Survey Big Data to Improve the Quality of the Household Budget Survey*. Nordic statistical meeting, Reykjavik, 23 – 24 August 2022. Available at: https://www.nsm2022.is/process-and-analyze

[8] Radermacher, W. J. (2021b), *New and Emerging Methods. Standardisation and Statistics*. The Survey Statistician, 2021, Vol 84, 24-31. Available at: http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2021_July_N84_04.pdf

[9] UNECE (2022), Core values of official statistics. Available at: https://unece.org/statistics/documents/2022/06/working-documents/core-values-official-statistics