Reducing response burden by using administrative cash machine data in Hungarian retail trade statistics

# Introduction

Retail trade and catering turnover statistics is one of the most important and popular indicator of the Hungarian economy. It is also a necessary data providing commitment of the Hungarian Central Statistical Office (HCSO) towards the Eurostat according to the prescribed timeline, frequency and details.

New data sources play a key role in official statistics, since more and higher quality data are needed parallel with reducing response burden. Due to a new legislation introduced in 2014, enterprises operating cash machines involved in the online cash register system are obliged to send online information about sales in retail chain to the Hungarian tax office. It was an obvious demand to change the methodology of the retail trade statistics in such a way that exploits the potential in these data.

# Methods

*Combining* the two *data sources* – the administrative data and the current survey data – is necessary for several reasons. First, it is obvious that the data coming from the online cash machines cannot be the one and only data source since the legislation does not cover all the population of the retail trade statistics. Also, we need to track the data coming from the tax office in order to build information in the estimation. Nevertheless, the case when something is going wrong with the reception of administrative data should not be excluded, so we need to establish a balance between response burden and information loss.

We have changed the *sampling design* according to the above. The population is divided into two parts: the full scope part involves the significant enterprises with all their shops, whereas the rest of the shops were surveyed by simple random sampling. We still keep this division while there will be two types of questionnaires: a „simplified” one for those enterprises whose shops (all of them) are in the scope of the legislation and do not carry on any other activity; and the original „full” questionnaire for the rest of the full scoped enterprises and for the simple random sample of the part of the population not covered by the legislation. Data for the rest of the population for which the simple random sampling was used before are coming from the TAX office.

Since the estimation of the retail selling is based on shop level in the HCSO, matching shops and cash machines was necessary. The major part of the *data editing* is then moved to this matching direction since it is inevitable to find the proper statistical NACE code for each cash machine. We also plan to handle some mistakes made by cashiers. Since we cannot reach all the data providers directly, we receive a separate data file from the TAX office including all the turnovers above a certain threshold. In case of the non full scope surveyed shops we subtract these amounts from the total turnover of the shops. This data editing step is automated.

*Missing data* are handled in the following way. We do not impute any missing data in the administrative data source. In the full scope part we impute with the previous-year-same-period data if they exist. If there is no such data but previous period data is available, the product of this and the average dynamique (the rate of the average of the period and the previous period) will be the imputed data. The same imputation will be used in the case of the sampled units. If no previous data is available we impute with a reduced average value.

The *estimation* will change as follows. In case of the full scope surveyed enterprises for which we receive the simpler questionnaire, we prefer the data sent through the questionnaire and use the rate of this data and the data from the administrative source when producing detailed statistics (e.g., for activity or region). For the estimation of the non full scope part we use a monthly correction for the cash machine data in order to compensate late response due to technical reasons. The new data source will not affect the estimation of retail sales of tobacco and of mail order and internet. The estimation for the enterprises/shops not in the scope of the legislation will unchanged.

# Results

By using the cash machine register data and the simplified questionnaires we can eliminate response burden for most of the small enterprises and shops and reduce response burden for the big enterprises (altogether burden reduction close to 80%). Moreover, we are able to cover more shops than before (from 64% to 95% of total retail sales) without increasing the response burden.

With the newly developed method we backcast the data for 2015 since we do not have reliable data for that year. Analysing the results of the new system within the period from January 2015 to March 2018 we get monthly retail trade turnover 1-4.8% less than with the existing method.

# Conclusions

Using administrative data source and handling its problems yield definitely reduced response burden and hopefully result also in more reliable data.

Quality measurement considerations are based on the principles of the European Statistics Code of Practice for statistical processes and products. The statistics to be produced have not been changed, so the new method does not affect relevance, accessibility and clarity. By assigning the statistical activity (NACE) we can also provide the appropriate groups so the principle of coherence and comparability is not damaged either. The sampling error is reduced to 0 for the retail trade sector (NACE 47) using cash register data. Turnover coverage will increase and the new method is able to handle the possible data loss. With the new method, timing is not compromised, statistics are being produced according to the Eurostat and the national requirements for the timeliness of data.

The HCSO is going to use the newly developed methodology for estimating retail trade and catering turnover from the beginning of 2019.