USING RETAIL TRANSACTIONS FOR CONSUMER PRICE INDEX AND EXPENDITURE STATISTICS

Li-Chun Zhang [1] [2]

[1] Statistisk sentralbyrå, Norway

[2] University of Southampton, (e-mail: L.Zhang@soton.ac.uk)

ABSTRACT: Scanner data arising from retail transactions have replaced survey of food price observations for the consumer price index (CPI) for more than a decade. The same data source can provide the expenditure weights needed for the CPI as well, when combined with population data using secure linkage and processing techniques that protect confidentiality. This would alleviate the most burdensome part of diary collection for the Consumer Expenditure Survey that collects expenditure data from households. Due to the sheer amount of transactions, automatic classification of the consumption subclasses of the goods requires natural language processing techniques, as long as there does not exist a catalogue that covers all the goods. Statistical theories pertaining to these big-data expenditure weights and classification are discussed.

KEYWORDS: audit sampling inference, evaluation coverage, entity resolution, maximum entropy classification, entity forest.

1 Big-data proxy expenditure weights

In some countries, scanner data arising from retail transactions have replaced survey of food price observations for the consumer price index (CPI) for more than a decade. The data are typically available on a weekly basis, in the form of unit value (average) price for each consumption goods or *item*. Scanner data constitute a promising source of price data for CPI, which is being expanded to other consumption subclasses such as clothes, electronics. For the price index methodology based on scanner data, we refer to the website of Ottawa Group.

Provided one can connect the transaction items of different consumer subpopulations, it is possible to calculate *proxy* CPI expenditure weights for any specific subpopulation. Unlike the survey-based weights, these proxy weights can be considered to have virtually zero sampling variance for practical purposes because of the sheer amount of data that can be made available. But they are generally biased due to a number of errors that are unavoidable in reality. In particular, these include coverage errors caused by the discrepancy between

the available transactions and the entire consumption of the population, and selection errors from the available transactions because, for various technical reasons, one is not able to code and classify all the items.

In such a situation, where bias completely dominates variance, modelling the intrinsic variability of the proxy weights would be fruitless, as long as it cannot capture the bias. Additional observations of expenditure are necessary to investigate the extent to which the proxy weights may be biased. Zhang (2021) propose and develop audit sampling inference for big-data statistics, which consists of the following elements:

I. clarify the *validity condition* for unbiased big-data statistics,
II. derive *tests* for the unbiasedness of big-data statistics,
III. *measure* the accuracy of the big-data statistics.

The theory of audit sampling inference is applied to the Norwegian data in following setup. First, fully anonymised food expenditure data are obtained for a single weekday in September of 2016, based on extractions provided by the largest debit card payment service and some of the largest supermarket chains. The proxy expenditure weights are calculated from 0.8 million transactions, broken down to four groups according to the age of the cardholder.

Next, take the Consumer Expenditure Survey 2012 as the audit sample, where the survey-based expenditure weights are treated as unbiased estimates of the true CPI food weights for 2012. Setting aside the coverage and selection errors of the available transaction data with respect to all household purchases, the proxy CPI weights do not refer to exactly the same subpopulations as those identified in the survey, because the transaction data refer to a different time point and the proxy weights are broken down by the age of the cardholder instead of the age of the household head. In other words, these proxy weights are *necessarily* biased for the true CPI weights in 2012.

Applying the tests developed for this setup, one is unable to reject the null hypotheses that the proxy-weights CPI for the different subpopulations are unbiased, despite the high power of the tests. However, since one can be certain that the proxy weights are not exactly equal to the true weights, it is sensible to treat these non-rejection results as indications for the usefulness of the resulting big-data CPI, and accuracy measures are still necessary.

Mean squared error (MSE) is a common choice of accuracy measure where bias is known to exist. However, MSE estimation can easily produce negative (hence unusable) results, where the audit sampling variance is large compared to the bias of the big-data statistic. It is unattractive to simply increase the

audit sample size in such situations, which means audit sampling would be more costly in a relatively favourable setting for adopting big-data statistics.

Zhang (2021) proposes and develops *evaluation coverage* as a novel accuracy measure for any big-data statistic, which is generally applicable based on audit sampling and overcomes the problem of limited audit sample size. Whereas the estimation of MSE runs into troubles in the said application, the evaluation coverage provides meaningful results. Indeed, to reach the same evaluation coverage of the proxy-weights CPI, the survey sample size would need to be increased approximately by a factor of 80 in some cases, which is unrealistic in practice.

In short, by the proposed approach of audit sampling inference, one can conclude from the study that proxy CPI weights derived from the transaction data can replace the relevant diary component that is the most burdensome part of the traditional expenditure survey.

2 Classification based on text

The consumption items are classified into subclasses called COICOP groups. Automatic classification of COICOP groups of the large amount of transaction items requires natural language processing techniques, as long as there does not exist a COICOP-catalogue that covers all the items.

Denote by $i = 1, ..., N$ the *items* to be classified. Let $U = \{1, ..., N\}$. Denote by $y = 1, ..., K$ the *groups* to which items are classified. Let $\Gamma = \{1, ..., K\}$. Denote by x any *term* that can be used in item description, e.g. jasminris, toalettpapir, etc. Denote by $\boldsymbol{x}$ the collection of terms in *item description*, possibly in vector-representation , e.g. $\boldsymbol{x} = \{\text{coop, jasminris}\} = (1, 1, 0, 0, ..., 0)_{1\times p}$. Denote by Ω the *corpus* of item description, i.e. $\Omega = \{\boldsymbol{x}_i : i \in U\}$.

For each $i \in U$, let y_i be its group classification. Group classification can be viewed as an entity resolution problem, where Γ are the (known) entities and U the records. The records U_y, $U_y = \{i \in U : y_i = y\}$, are considered to be matched (to each other) via co-reference to the entity y. The *resolution* we seek is the partition $U = \bigcup_{y\in\Gamma} U_y$, denoted by $\mathbb{C} = \{U_y : y \in \Gamma\}$.

One can distinguish generally the *discriminative* or *generative* machine learning approach to classification or entity resolution problems. By the discriminative approach, classification of y_i for any $i \in U$ is based on

$$f(y|\boldsymbol{x}; \Omega) = \Pr(y_i = y \mid \boldsymbol{x}_i = \boldsymbol{x}; \Omega)$$

where the different terms in an item description $\boldsymbol{x}$ are used as distinct features

for $f(y|\boldsymbol{x};\Omega)$. Let $f_U(y|\boldsymbol{x};\Omega)$ be the model function given the corpus Ω and the true resolution $\{U_y : y \in \Gamma\}$. As long as there exists any term x, e.g. $x =$ ekstra, which appears in multiple item descriptions not all belonging to the same group, classification of any $\boldsymbol{x}_i$ that contains this x-term may be incorrect by the *discriminative classifier*

$$y_i = \arg\max_{y\in\Gamma} \; f_U(y|\boldsymbol{x}_i;\Omega)$$

By the *generative approach*, one would focus on the model function

$$f(\boldsymbol{x}|y;\Omega) = \Pr(\boldsymbol{x}_i = \boldsymbol{x} \mid y_i = y;\Omega)$$

Let $f_U(\boldsymbol{x}|y;\Omega)$ be the model function given the corpus Ω and the true resolution $\{U_y : y \in \Gamma\}$. The corpus Ω is *free of entity-duplication* provided, for any $\boldsymbol{x} \in \Omega$,

$$\sum_{y\in\Gamma} \mathbb{I}\big(f(\boldsymbol{x}|y;\Omega) > 0\big) \equiv 1$$

This *admissibility* condition is in fact necessary for any well-defined mapping from the item descriptions Ω to the groups Γ. Notice that it allows for multiple items with the same $\boldsymbol{x}$ as long as they all belong to the same group. Given any admissible corpus Ω, classification of any $i \in U$ based on $\boldsymbol{x}_i$ would always be correct by the *generative classifier*

$$y_i = \arg\max_{y\in\Gamma} \; f_U(\boldsymbol{x}_i|y;\Omega)$$

even if there are terms belonging to item descriptions in different groups.

Thus, perfect classification is conceptually possible and easily achievable only by corpus engineering under the generative approach. We develop a generative approach to item classification based on text descriptions. Entity resolution and maximum entropy classification are adopted as the formal framework. In situations where only a subset of all the items have known classifications, we develop supervised learning of an *entity forest* model and associated classification method (based on item descriptions) for the rest of the items.

References

Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, **184**, 571-588.