Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Consumer expenditure statistics and retail transaction data

Prepared by Li-Chun Zhang and Henning Holgersen, Statistics Norway

I. INTRODUCTION

1. The most important use of the Consumer Expenditure Survey (CES) in the past has been to provide the expenditure weights for the Consumer Price Index (CPI). However, the CES is extremely burdensome – especially due to the diary component, has a very high nonresponse rate, and is known to suffer from misreporting errors for various types of consumption. Moreover, the CPI weights should ideally be available at the COICOP-V level or lower. For instance, the total food expenditure then needs to be divided into more than 100 groups, the sum of which changed only 0.2% from 2000 to 2012 according to the Norwegian CES. The traditional CES simply cannot support such a high demand of information. These are some of the reasons Statistics Norway suspended the survey in 2012 and has been investigating alternative sources of data since then.

2. There is clearly a great potential to use transaction data from retail chains as a basis for the relevant expenditure statistics, replacing either completely or mostly the collection of private expenditure data on items that are the most troublesome to collect in the traditional surveys. In fact scanner data from the same source has already been used to produce the relevant price indices for over a decade. There are essentially three types of big data that originate from retail transactions:

- scanner data which shows what is sold at which price but not to whom;
- bank transaction (card payment) which shows who spends how much but not on what;
- transaction receipt which shows the same as scanner data, and possibly who is the customer in case of card payment or if loyalty membership is registered.

There are clearly confidentiality issues related to the latter two types of transaction data. For this reason, in this paper we will not present any details of the data which Statistics Norway may or may not have access to, but only illustrate the relevant aspects in a conceptual and generic manner.

3. Now that scanner data are already being used in Official Statistics, the main added value from the retail transaction data is to provide disaggregation of the relevant statistics, either over demography or geography. For instance, it is of interest to many users to produce the CPI for specific subpopulations such as the pensioners. Insofar as transaction data can be classified by the demographic (otherwise non-personal) characteristics of the corresponding customers, using only non-confidential grouped data may facilitate the greater uptake of such data in Official Statistics.

4. There are nevertheless many methodological problems that need to be resolved in practice. To start with, in case it is necessary or helpful to combine the transaction data with other datasets, linkage can be a challenge due to the confidentiality requirements. Next, with or without linkage, the transaction data may be subjected to missing and measurement errors, just like any other types of data. Finally, as will be discussed later, regardless how good the editing and imputation (E&I) methods

are, coverage and measurement errors will remain in the processed data. *It is therefore necessary to consider the subsequent estimation and the E&I procedures in connection,* not only to make sure that the resulting statistical data can satisfy the relevant accuracy requirements, but also to provide the basis for appropriate design and evaluation of the E&I strategy. Without such a wholesome perspective, bigger data can create bigger trouble in practice.

5. In the rest of the paper we will outline some relevant E&I problems with transaction receipt data. Particular attention will be given to the discussion of their connections to the subsequent estimation. The potential linkage problem is not considered in this paper at all.

II. SOME E&I PROBLEMS OF TRANSACTION RECIPT DATA

1. Loyalty card programmes are offered by many retail chains and are usually based on an agreement between the customer and the business to let the business collect purchase details and some personal information about the customer, in exchange for discounts or cash-back programmes. In this Section we consider some E&I problems of transaction receipt data, where some of them are associated with loyalty card membership and some are not. For a systematic assessment of all the potential errors, we use the two-phase model of integrated statistical data (Zhang, 2012). However, since we do not discuss linkage to any other datasets, only the 1st-phase model is needed (Figure 1).

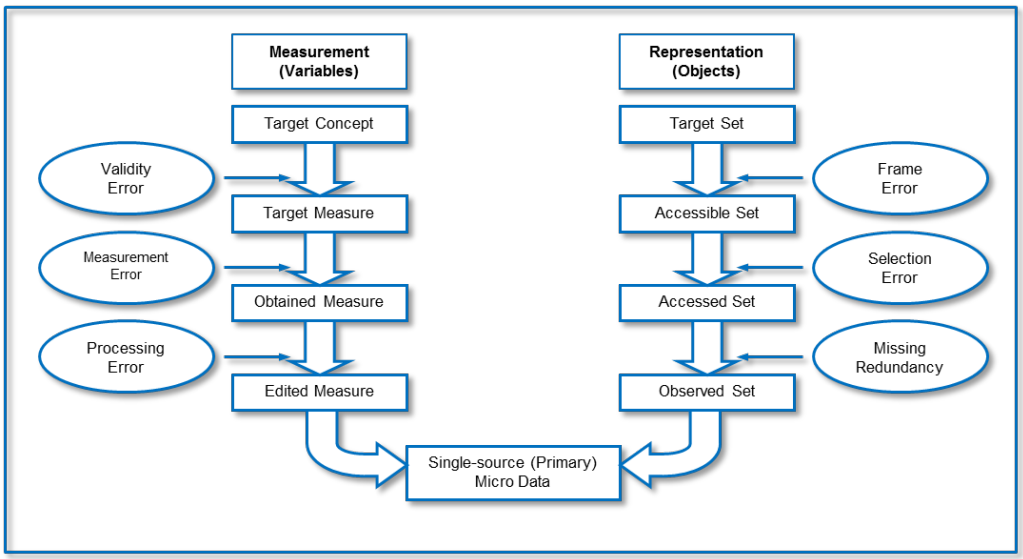

FIGURE 1. Total error framework of single-source data

2. To apply the model one needs to decide the object and variables as the row and columns of a data matrix. Let each row correspond to a transaction. The columns can contain the membership and related identifiers, which is missing if no loyalty card is registered for the transaction. They can further contain the associated geographic and possibly demographic variables. Finally, they contain three vectors of items, one each for the corresponding prices, quantities and monetary totals. It is possible to define the data matrix differently, but we will use this specification to fix the idea.

3. Take first the Representation side. Given the specification above, the target set is the collection of relevant transactions for the CPI weights. Let it be all the transactions by domestic private

consumers. The accessible set are now the transaction receipt data from those retail chains that deliver data to the NSO, according to the scope specified by the data agreement. In general this will contain purchases by foreign and business customers in addition to domestic private consumers. The accessed set contains the transactions the NSO actually receives. The observed set is the retained transactions as the result of E&I process, which is a proper subset of the accessed data, because some transactions are deemed out-of-scope (e.g. containing only garden furniture items in data from supermarkets) or unusable (e.g. containing only items that cannot be correctly classified). It is thus clear that the final observed set will always suffer from both over- and under-coverage errors, and the selection mechanism of the retained set does not follow a probability design.

4. Take next the Measurement side. The target concept differs for different variables. Consider e.g. the variable age. The target concept is the age of the consumer, but one can only observe in some cases the age of the customer, which can be considered the target measure (i.e. operational but not necessarily ideal). Next, the obtained measure can only be the age of a loyalty card member, which is not always that of the customer. The edited measure may differ from the obtained measure for various reasons. For example, a card member can have a registered age that is outside of the realistic range of living customers. Or the card may be associated with several persons, in which case one can at most make a guess at picking one of them. Similarly for the other variables. In particular, for the items purchased, only a subset of all the obtained items will be retained, e.g. because one is not able to classify all of them. Moreover, the retained items may be misclassified. It is thus clear that the final edited variables will always be subjected to missing and measurement errors.

5. As an example of micro-data E&I methods, consider the classification of purchased items. To start with, one needs to map the catalogue of goods (GTIN) from the retain chains to the COICOP catalogue at the NSO. This has problems of its own, which can differ from the situation when diary data are collected in the CES. Complete mapping is by and large impossible in practice. Some text mining techniques, such as Bag of Words, will be needed to make the most of the obtained data.

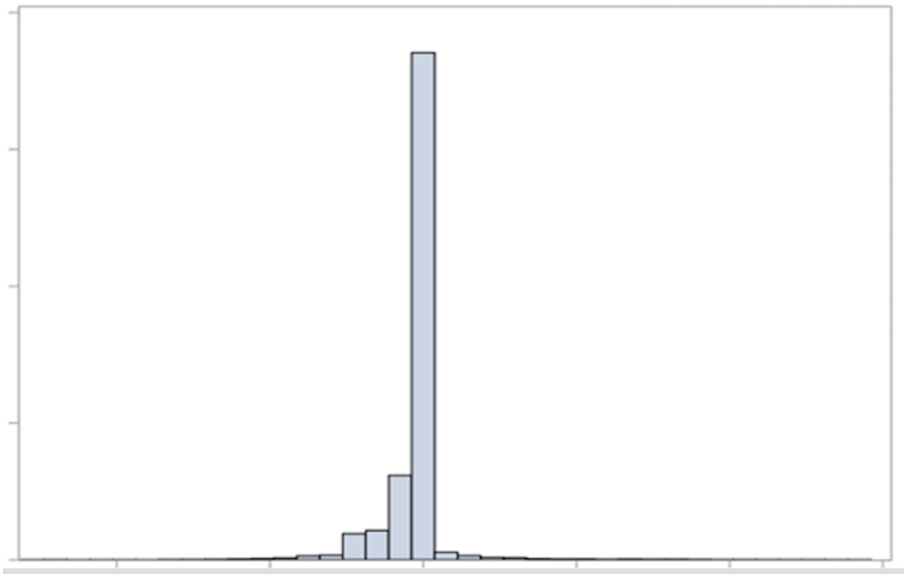

FIGURE 2. Illustration of possible source effects

6. As an example of macro editing, suppose scanner data are delivered from the same retail chain that also supplies the transaction data. One may e.g. compare the expenditures derived from the

scanner data and the transaction receipt data of the same period, respectively, since in principle the two are based on the same transactions. The histogram in Figure 2 illustrates the possible outcome of the comparison, where most of the item-specific expenditures may differ little between the two sources, while some may have a relatively large discrepancy. There can be many reasons for such source effects, from the specification of data agreements to the processing by the data owner and the NSO.

7. Finally, as an issue of overall importance, we would like to draw one's attention to potential problems related to the metadata. The quality of the supplied metadata can e.g. affect the identification of the transactions and associated items. They can e.g. cause problems regarding how refunds and discounts can be appropriately handled, which are sometimes subtracted from the monetary total sometimes from the price, possibly leading to negative values of both kinds.

III. E&I FOR ADJUSTMENT AND ESTIMATION

1. It can be seen form above that direct tabulation for the statistics of interest may fail to convince, *regardless of how good the E&I process is*. In this Section we consider some possible adjustment and estimation methods. To reiterate, the key message we would like to convey is that one needs to consider them together with the E&I methods, not only to make sure that the resulting statistical data can satisfy the relevant accuracy requirements, but also to provide the basis for appropriate design and evaluation of the E&I strategy.

2. Suppose one is interested at CPI expenditure weights. For the transaction receipt data, let us for the moment disregard the coverage and measurement errors in the processed data, and focus on the fact that the retained transactions are not a random sample from all the transactions relevant for the CPI, even if they form a proper subset of the target set. The overall expenditure patterns of card holders and non-card customers can be compared to each other. Suppose that based on the edited data, different patterns are observed. For instance, on average, card holders may be substantially older or younger than the resident population of Norway. Now, under the assumption that, conditional on age, the expenditure pattern is the same for card holders and non-card customers, one may adjust the observed overall expenditure patterns by post-stratification. The $\mathbf{X}_{age}$ be the vector of age-specific expenditure weights, and $\mathbf{X}_{age,I}$ the corresponding observed vector from the card holders. Let

$$W_{age} = N_{age}/N$$

be the proportion of people of a certain age in Norway. The post-stratified estimator of the overall weights is given by

$$\widehat{\mathbf{X}} = \sum_{age} W_{age}\mathbf{X}_{age,I}$$

Figure 3 provides an illustration of the possible outcome, when the directly observed $\mathbf{X}_I$ and the post-stratification adjusted $\widehat{\mathbf{X}}$ are compared the expenditure weights actually used in the CPI. The main point we want to illustrate here is that the adjustment may seem 'unsatisfactory' in two aspects: (i) it may have little effect on the directly observed pattern, and (ii) both the unadjusted and adjusted patterns may differ quite much to the results one expects. One may draw the conclusion that there are other confounding factors which are unavailable in this dataset.

3. What then? Does it mean that the 'proxy' weights based on the transaction data cannot be used to *replace* the CES? Now, even when the proxy weights are not perfect, they can still represent a considerable cost saving opportunity, both for the respondents and the NSO, as long as they are 'better' than the CES-based weights. So the appropriate question is how the transaction-data weights

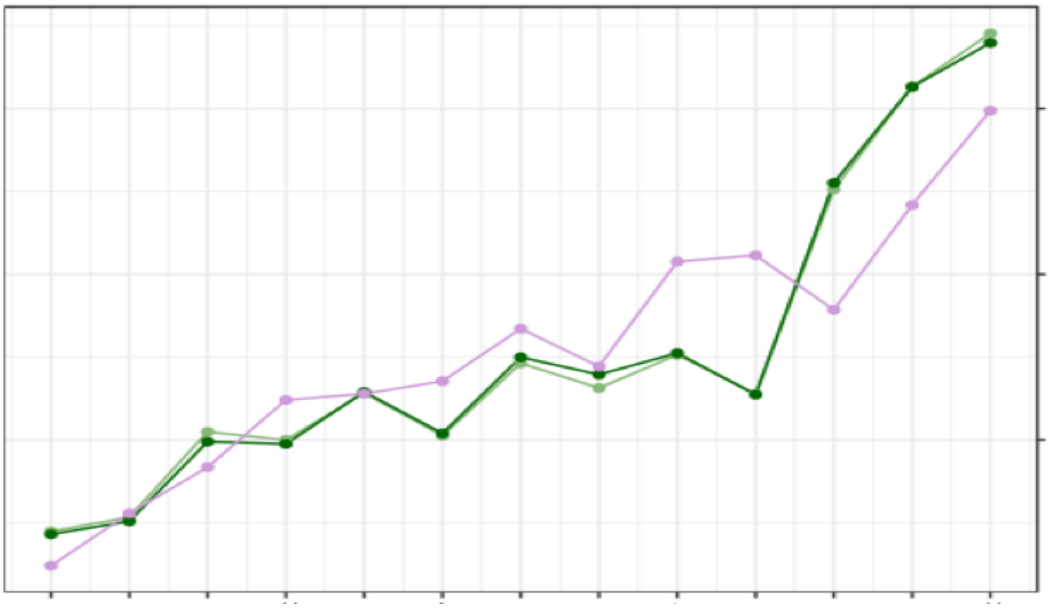

FIGURE 3. Illustration of possible expenditure weights based on card-holders (direct or post-stratification adjusted), compared to expenditure weights used for CPI (pink)

compared to the CES-based weights. Theoretically, let the CPI for food be given by

$$P = \sum_{g=1}^{G} w_g P_g \qquad \text{where} \quad \sum_{g=1}^{G} w_g = 1$$

and $g = 1, ..., G$ denote the groups for which the CPI weights need to be calculated. The actual source effect from replacing the CES by the transaction data can then be defined as

$$\Delta = \sum_{g=1}^{G} (w_g^{proxy} - w_g^{CES}) P_g$$

It follows that there will be *no* source effect, provided

$$Cov(b_g, P_g; \tilde{w}) = \sum_{g=1}^{G} \tilde{w}_g b_g P_g = -\sum_{g=1}^{G} (w_g^{proxy} - w_g^{CES}) P_g = -\Delta = 0$$

where $b_g = w_g^{CES}/w_g^{proxy} - 1$ and $\tilde{w} = w_g^{proxy}$. That is, if the covariance between the price index and the relative difference between the two weights is zero, where the covariance is evaluated with respect to the probability mass function $(w_1^{proxy}, ..., w_G^{proxy})$. We make two obsrevations.

- The source effect is observed based on the two sets of weights and the price indices.
- The effort of processing the age variable in the transaction data can result in a smaller source effect, if the corresponding covariance is closer to 0 when calculated based on the adjusted proxy weights derived from $\widehat{\mathbf{X}}$ instead of the directly observed from $\mathbf{X}$ without using age.

This provides an example why it is important to connection the E&I process to the subsequent adjustment and estimation.

4. As mentioned earlier, a main added value of the transaction data is to provide disaggregation of the CPI not supported by other means. For example, suppose one is interested in age-specific CPI expenditure weights. Under the assumption of the post-stratification adjustment above, one can use

the weights $\mathbf{X}_{age,I}$ observed directly from the card holders. Another idea is to impute (or estimate) the age of non-card customers first. For the transaction receipt data, let us for the moment disregard the conceptual difference between consumer and customer mentioned earlier, and assume that age is observable for loyalty card holders but not otherwise, but the observed age-specific transactions are *not* a random sample of all the age-specific transactions. A generic description of the available data is given in Table 1. More specifically, expenditure is observed together with age in the case-I, but expenditure is only observed marginally in the case-II where age is missing.

TABLE 1. Generic availability of data

	Age	Expenditure
I. Card holder	Yes	Yes
II. Non-card Customer	No	Yes

5. Under the assumption age is independent of expenditure given I or II, denoted by

$$f(\text{age}|\text{exp.}, \text{II}) = f(\text{age}|\text{exp.}, \text{I}) \tag{1}$$

one can impute the age for in the case-II. This requires building a prediction model of age given expenditure in the case-I, for which supervised machine learning (ML) techniques such as kNN can be used. However, suppose that the result is 'unsatisfactory' using the chosen ML technique, because expenditure seems to have very little 'explanatory power' on age, denoted by

$$f(\text{age}|\text{exp.}, \text{I}) \approx f(\text{age}|\text{I}) \tag{2}$$

Under (1) and (2), one would impute age for the non-card customers (case-II) using simply the age distribution among the card holders (case-I), which is unattractive. Instead, aiming at the target distribution $f(\text{exp.}|\text{age})$ directly under the assumption (1), we have

$$\begin{aligned}
f(\text{exp.}|\text{age}) &= f(\text{exp.}, \text{I}|\text{age}) + f(\text{exp.}, \text{II}|\text{age}) \\
&= f(\text{exp.}, \text{I}|\text{age}) + f(\text{exp.}|\text{II}, \text{age}) f(\text{II}|\text{age}) \\
&= f(\text{age}|\text{exp.}, \text{I}) \frac{f(\text{exp.}, \text{I})}{f(\text{age})} + f(\text{age}|\text{exp.}, \text{II}) \frac{f(\text{exp.}, \text{II})}{f(\text{II}, \text{age})} f(\text{II}|\text{age}) \\
&\overset{(1)}{=} f(\text{age}|\text{exp.}, \text{I}) \frac{f(\text{exp.}, \text{I})}{f(\text{age})} + f(\text{age}|\text{exp.}, \text{I}) \frac{f(\text{exp.}, \text{II})}{f(\text{age})} \\
&= f(\text{age}|\text{exp.}, \text{I}) \frac{f(\text{exp.})}{f(\text{age})} \\
&\neq f(\text{age}|\text{exp.}, \text{I}) \frac{f(\text{exp.}, \text{I})}{f(\text{age}, \text{I})} = f(\text{exp.}|\text{age}, \text{I})
\end{aligned}$$

i.e. a different estimator of the age-specific expenditure than that directly observed from the card holders. In this way, the analysis shows one *how* to work with and use the variable age under the assumption (1). There is no need to impute age of the non-card customers, whether or not the extra assumption (2) may be empirically established. This provides yet an example why it is important to connection the E&I process to the subsequent adjustment and estimation.

References

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