WP.17

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

MODELLING THE CONSTRUCTION OF A SOCIAL ACCOUNTING MATRIX IN THE CONTEXT OF STATISTICAL MATCHING

Supporting Paper

Submitted by the Italian National Statistical Institute (ISTAT), Italy[1]

I. INTRODUCTION

1. In the last few years, interest in statistical matching problems has increased (see Rässler and references therein). This is due to the large number of datasets available and, at the same time, to the need of timely and not costly information. Actually, Statistical Matching techniques aim at combining information from different sources. In particular, it is assumed that the two sources (e.g. two samples) do not observe the same set of units, so that neither merging nor record linkage techniques can be applied.

2. There are various statistical matching applications, for example, the construction of the Social Accounting Matrix (SAM) is one of the most important. The SAM is a system of statistical information containing economic and social variables in a matrix formatted data framework. The matrix includes economic indicators such as per capita income and economic growth. In Italy, as well as in other countries, such an archive is not available; hence, the SAM is built by the fusion of the Household Balance Survey conducted by the Bank of Italy and the Household Expenditure Survey conducted by the Italian National Statistical Institute.

3. This paper will explain how this application can be modelled in the statistical matching context.

II. THE SOCIAL ACCOUNTING MATRIX

4. In the new system of National Accounts (also known as European System of the Accounts, or SEC95) a very important role is assigned to the Social Accounting Matrix (SAM). United Nations (1993), Chapter XX, defines a Social Accounting Matrix (SAM) as a system of statistical information containing economic and social variables in a matrix formatted data framework, registering entries and outlays of the economic agents grouped according to appropriate characteristics. In particular, a SAM is different to an input-output table and typical national account because more details about all kinds of transactions within an economy are reported. Generally speaking, a SAM aims at:

(i) organizing knowledge on the social conditions and economic structure of a nation in a year, and
(ii) providing the definition of a plausible economic model able to furnish a static image of the economy together with simulating how policy interventions affect the economy.

[1] Prepared by Marcello D'Orazio (madorazi@istat.it), Marco Di Zio (dizio@istat.it) and Mauro Scanu (scanu@istat.it).

5. Many countries and statistical institutes have the objective of estimating a SAM. In Italy, a SAM cannot be estimated easily, because all the necessary information is not available in a unique dataset (archive or sample) but should be taken from many different and independent datasets. In the following, we describe the problem of the estimation of just one module of the SAM, i.e. the one related to households. This problem was initially studied in Coli and Tartamella (2000a, 2000b) and finally in Coli *et al.* (2005).

6. The module of the SAM on the households is a matrix where households are distinguished according to a set of different typologies: for instance, the household area of residence and the household primary income source. The SAM organizes for these household typologies:

(i) the number of outlays (distinguished according to a very detailed list of different expenditures categories);
(ii) the number of entries (divided in compensation of employees, self-employed income, interests, dividends, rents).

7. In Italy, there are two main sources containing reliable information on respectively the entries and outlays of the households:

(i) the Bank of Italy Survey of Households' Income and Wealth (SHIW), and
(ii) the Istat Sample Survey on Household Consumption (SSHC).

The first survey studies household income and wealth, according to the different household entries components. The second survey estimates household final consumption at a very high level of detail, from the acquiring household group to the type of purchased group of products.

8. The previous two surveys are independent surveys, organized and carried out by two different institutes. They should be *integrated*, or in other words *fused*, in order to put together information on household outlays from the SSHC and information on household entries from the SHIW. This integration process can be carried out by means of information on the socio-economic characteristics observed in both the samples. This is exactly a *statistical matching problem.* The matching process consists of three steps:

(i) the consistency of the two surveys should be checked and, if necessary, the two surveys should be harmonized;
(ii) the statistical framework where the sample surveys live should be defined; and
(iii) according to the two previous steps, an appropriate statistical matching method should be applied.

9. In the following sections, the first two steps are described as in Coli *et al* (2005). The last step is detailed in a simpler way. Only total household entries and outlays are taken into consideration, with a few common socio-economic variables.

III. THE HARMONIZATION STEP

10. The two surveys SHIW and SSHC are affected by many inconsistencies. These inconsistencies should be solved to make them at first comparable, and consequently integrable, through harmonization of:

(i) population and unit definitions, and
(ii) variables definitions.

These steps must be performed with great caution, noting that there does not exist an “optimal” procedure. In fact, the harmonization phase consists of a kind of “simplification” of a set of key characteristics of the different surveys. This operation produces changes in the original variables meaning, changes in the definition of the population target, and as an overall result changes in the initial informative power of the samples. Statistical matching output is greatly affected by these operations. A rule of thumb can be the following: change as less as you can during the harmonization step.

11. The two surveys (SHIW and SSHC) target populations are the Italian households. However, these two surveys have a different definition of the units of the target population, i.e. they use two different definitions of "household". The SSHC definition of household consists of a set of cohabitants, linked by wedding, familiarity, affinity, adoption. The SHIW defines a household as a set of people that, independently of familiarity, joins completely or partially their entries for their necessities.

12. Hence, the unit definitions of the two surveys are inconsistent. This inconsistency is difficult to solve, because the two surveys do not contain enough information for making a SHIW household in one or more SSHC households, and vice versa. Actually, Coli *et al* (2005) consider this a minor problem. In fact, the two populations almost overlap, i.e. the set of SHIW households inconsistent with SSHC definition, as well as the set of SSHC households inconsistent with SHIW definition are very small. Furthermore, apart from the definition, the samples did not contain inconsistent households. Hence, both SHIW and SSHC were assumed to be two samples drawn from the same population defined as the intersection of the two previous population definitions.

13. As far as variable harmonization is concerned note that, although SHIW and SSHC investigate two different aspects of the household economic situation, they observe a large set of common variables. Roughly speaking, these variables can be clustered in three groups: socio-demographic variables, variables on the household outlays and variables on the household entries. These variables are usually inconsistent in their overall definition or, when the overall definition coincides, in their categorization (number and type of states of a variable). In this case, variable harmonization uses different strategies.

(a) Some variables cannot be harmonized. These variables are not useful for statistical matching of the two samples.
(b) Some new variables take the place of the original ones, by appropriate transformations.
(c) Some variables are just recoded.

14. The first group contains a very important variable: the "head of the household". Actually, this variable is very important in the SAM, because one of the socio-economic group of household organized in a SAM consists of the households clustered according to some head of the household's characteristics (e.g. age, gender, education, and characteristics on his/her work status). The justification for clustering households according to head of the household characteristics is based on the assumption that these characteristics are usually correlated with both the household outlays and entries. As a matter of fact, one survey assumes the head of the household is the one registered on public archives, while the other assumes as head of the household the one who is responsible of the household economy. The two surveys do not contain enough information for the harmonization of such definitions. Hence, the head of the household and his/her characteristics were disregarded during the matching of the two samples. Note that, once the two samples are matched, the characteristics of the head of the household of the SSHC were maintained and used for the analyses. As a matter of fact, this operation hides the notion of conditional independence between the head of the household characteristics and the variables of the SHIW not in common with the SSHC.

15. Among the variables in the second group, many describe some household characteristics, and in particular these variable describe socio-economic characteristics of the different household components. In fact, these characteristics are better used if reported at the household level, instead of single individual level. For instance, additional variables as the number of household components aged 64 years or more (categories: 0,1,2 2+), number of employed components (0,1,2,3,3+), number of graduated components (0,1,2,2+), number of females (0,1,2,3,4,4+), have been introduced. Note that these variables have been considered during the matching process. Hopefully, the head of the household characteristics, previously disregarded in the matching phase, is actually independent of income and expenditures given the socio-economic characteristics of all the components.

16. As far as the last group of variables is concerned, the two surveys contain many variables affected by different categorizations. The harmonization step consists in defining a common categorization given by the largest categories in common.

IV. MODELING THE SOCIAL ACCOUNTING MATRIX

17. The statistical formalisation of the construction of the SAM as treated in ISTAT is explained in detail in Coli *et al.* (2005). With some simplification, the main objective can be summarised as the construction of Table 1, where $\mathbf{C}$=(C_1,…,C_u) are different consumption typologies (e.g. food consumption, durable goods), $\mathbf{I}$=(I_1,…,$I_?$?are different sources of income and T_1($\mathbf{X}$),…,T_m($\mathbf{X}$) address different family typologies that are function of the demographical variables $\mathbf{X}$=(X_1,…,X_k) (e.g. educational level, age,...). The two surveys SHIW and SSHC are represented respectively in Tables 2 and 3.

Table 1: SAM objective

	C_1	…	C_u	I_1	…	$I_?$
T_1(X)						
…						
T_m(X)						

Table 2: Variables observed in the sample survey on household income and wealth carried on by the Bank of Italy (SHIW)

X_1	…	X_k	I_1	…	$I_?$	C_1	…	C_{u1}	TC

Table 3: Variables observed by the Household Budget Survey carried out by the Italian National Statistical Institute (SSHC)

X_1	…	X_k	TI	C_1	…	C_k

18. In Table 2 we remark that C_{u1} is a variable composed by a subset of variables of C_1,…,C_u, for example food consumption that is determined by bread consumption, egg consumption and so on; while TC is the total consumption.

19. In Table 3, the variable TI represents the total income and it is generally registered in categories.

20. Since the two surveys are carried out with two different objectives, the first for the income analysis and the second for the consumption analysis, variables representing consumption in Table 2 and incomes in Table 3 present coarse categorisations, or even do not have a high degree of reliability. This has to be taken into account when building Table 1.

21. The construction of Table 1 can be pursued either by directly estimating the cells quantities (macro-approach) or by first building a data set of micro data obtained by the fusion of the two Tables 2 and 3 and then by computing the cells corresponding to Table 1 (micro-approach), D'Orazio *et al.* (2002).

22. In any case, it is important to remark explicitly the unavoidable hypotheses characterising the problem. In general, we can write that our objective is the probability distribution $P(\mathbf{X}, \mathbf{I}, \mathbf{C})$, that can be written as $P(\mathbf{C} \mid \mathbf{X}, \mathbf{I})\, P(\mathbf{X}, \mathbf{I})$.

23. The probability distribution $P(\mathbf{X}, \mathbf{I})$ can be estimated by the Bank of Italy survey on income, where the variables are jointly observed, i.e.

$$\hat{P}(\mathbf{X}, \mathbf{I}) = \hat{P}^{(SHIW)}(\mathbf{X}, \mathbf{I}).$$

24. More problematic is the estimate of the probability distribution $P(\mathbf{C}|\mathbf{X},\mathbf{I})$. Since the consumption variables need to be estimated on the Household Budget survey, it is required another assumption: $P(\mathbf{C}|\mathbf{X},\mathbf{I}) = P(\mathbf{C}|\mathbf{X},\mathrm{TI})$.

25. Once the unavoidable assumptions have been introduced, it is possible to follow two different approaches each based on different hypotheses.

26. The first approach is based on the independence of **C** and TI conditionally on **X,** i.e. conditional independence assumption (CIA). In this case it is stated that the statistical relationship between consumption **C** and income **I** variables is explained by the common demographic variables **X**. Under this assumption we can write $P(\mathbf{C}|\mathbf{X}, \mathrm{TI}) = P(\mathbf{C} \mid \mathbf{X})$. Thus the estimate of such a distribution can be performed using the SSHC table, i.e.

$$\hat{P}(\mathbf{C}|\mathbf{X}) = \hat{P}^{(SSHC)}(\mathbf{C}|\mathbf{X}).$$

This hypothesis is done when considering the income variables in the SSHC unreliable, or at least not directly usable. This setting was performed in the first construction of the SAM in ISTAT.

27. Another hypothesis investigated in the application performed in ISTAT relaxes the previous assumption and try to exploit as much as possible all the information.
Here we assume that not the values observed on the income variables in the two surveys, but the relative position of the unit (household typology) is reliable. For instance, if there are ? % units in the first class of TI of units with an income in (0,100) in SSHC, the category to be considered in the income variable in SHIW is not the class formed by all the units in the interval (0,100) but from all the first ? % units, roughly who belongs to a "poverty" class in the first survey, belongs to the same "poverty" class in the other survey even if the values characterising the "poverty" class are different.

28. Based on this assumption we can discretise the variable $\mathrm{TI}^{(SHIW)} = I_1 + \ldots + I_?$ in SHIW, in order to reproduce the k quantiles $q_i(\mathrm{TI})$, for i=1, …, k, of the variables TI in SSHC, and thus we need to estimate

$$\hat{P}(\mathbf{C} \mid \mathbf{X}, q_i(\mathrm{TI})) = \hat{P}^{(SSHC)}(\mathbf{C} \mid \mathbf{X}, q_i(\mathrm{TI}^{(SHIW)})), \quad i=1,\ldots,k.$$

29. Once the hypotheses have been clarified, the final step consists of the construction of Table 1. However, the cells are not directly estimated, but they are formed after the construction of an integrated data set where all the variables needed for estimating Table 1 are present. This data set is obtained through imputation techniques following one of the two hypotheses so far discussed.

30. The construction of the dataset having all the variables jointly present has been performed by imputing the SHIW missing variables (completely non-observed) with values taken from the SSHC data set. In other words, the SHIW is the recipient file and the BH is the donor file. The imputation is done through the distance hot-deck stratified, which means that the missing values have been imputed with the conditional mean (conditional to the parameters used in the computation of the distance). If we assume the first hypothesys (CIA), the strata and distance variables chosen for the nearest neighbour are all into the variables **X**, this corresponds to impute with the non-parametrically estimated conditional mean of the distribution $\hat{P}(\mathbf{C} \mid \mathbf{X})$. On the other hand, assuming the second hypothesys, i.e. relaxing the conditional independence assumption, the distance hot deck is performed choosing strata in **X** and $q_i(\mathrm{TI})$, for i=1, …,

k, and distance variables in **X**. In particular, the deciles were chosen, i.e. q_i(TI) , for i=1, …, 10. Analogously to the previous case, this is equivalent to impute through the non-parametrically estimated conditional mean of the distribution $\hat{P}$ (**C** | **X,** q_i(TI)), i=1,…,10.

31. The choice of the SHIW as recipient file, and the SSHC as donor file is essentially due to the fact that the observations in the SSHC are much more than that of the SHIW data set. This is justified by the fact that the good behaviour of non-parametrc estimates are essentially in terms of asymptotic properties.

32. Both the hypotheses of conditional independence of **C** and TI given **X**, and of preservation of relative positions of the Households in the surveys according to the observed income are non testable from the data sets at hand. Nevertheless, the latter hypothesis provides results which are more consistent with economic theory as far as the relationship between **C** and TI is concerned. In fact simulations have shown that, even though a large number of covariates **X** is used, statistical matching procedures under CIA gives correlation coefficients between TI and the imputed **C** not larger than 0.30, while the latter approach gives correlation coefficients around 0.65.

References

Coli A and Tartamella F (2000a) The link between national accounts and households micro data. Proceedings Meeting of the Siena Group on Social Statistics. Maastricht (The Netherlands), 22 - 24 May 2000.

Coli A and Tartamella F (2000b) A pilot social accounting matrix for Italy with a focus on households. 26th General Conference of the International Association for Research in Income and Wealth. Cracow (Poland), 27 August - 2 September 2000.

Coli A, Tartamella F, Sacco G, Faiella I, Scanu M, D'Orazio M, Di Zio M, Siciliani I, Colombini S and Masi A (2005) La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane (in Italian). Technical report, work group Istat- Bank of Italy.

D'Orazio M., Di Zio M., and Scanu M. (2002), "Statistical Matching and Official Statistics", *Rivista di Statistica Ufficiale*, 1/2002, 5-24.

Rässler S. (2002), *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*, Lecture Notes in Statistics, New York: Springer Verlag.

United Nations (1993), *System of National Accounts*. United Nations, New York.
