Workshop *"2009 European Establishment Statistics"*, Stockholm, September 7-9, 2009

A linkage experiment between survey business data and administrative data

Daniela Ichim, Cristina Casciano, Giovanni Seri
Italian National Statistical Institute - Istat
Via Cesare Balbo, 16 - 00184, Roma, Italy
e-mail: ichim@istat.it, casciano@istat.it, seri@istat.it

1. Introduction

The exploitation of administrative data is one of the greatest challenges facing the NSIs. The administrative data may be useful in the sample design stage or in the estimation phase. Auxiliary information may also be useful in data validation and editing. This paper reports several experiments undertaken to identify an optimal matching strategy for business data. The problem of linking survey and administrative data is addressed. The business survey data is the Small and Medium Enterprises (SME) survey while the administrative data source is the Balance Sheets (BIL).

In section 2 a brief data description is provided. The data integration settings are detailed in section 3, with special emphasis on pre-processing, blocking and the choice of the matching variables and the corresponding distances. The main linkage results are presented in section 4, while some ideas for future work are given in the final section.

2. The administrative and survey data

Small an Medium Enterprises sample survey (SME) is carried out annually by sending a postal questionnaire with the purpose of investigating profit-and-loss account of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulation n°58/97. The main variables of interest are "Turnover", "Value added at factor cost", "Employment", "Total purchases of goods and services", "Personnel costs", "Wages and salaries", "Production value", etc.

The frame for the SME survey is the Italian Statistical Business Register (ASIA). ASIA results from the logical and physical combination of data from both statistical and administrative sources. The "Business Fiscal Turnover" is provided from the Fiscal Register, this variable being a good proxy of the "Turnover" collected in SME.

SME sample survey's population of interest is about 4 millions of active enterprises. Both the selection and estimation phases are based on the information available in ASIA, but a time lag exists between the reference years of SME and BR. The sampling design of SME is a one stage SRS, with the strata derived from the principal economic activity, size classes and administrative region. The sample size is about 120.000 units, selected according to multi-variable and multi-domain sample allocation methodology (Bethel, 1989). The correction factors for the initial sampling weights for unit non-response and under-coverage are calculated by applying a methodology based on calibration estimators (Deville and Sarndal, 1992).

The Italian limited enterprises are obliged to fill their financial statements according to the standards speciefied in the EEC fourth Directive and to transmit them to the Chambers of Commerce. The resulting database is called Balance sheets (BIL). This data source is actually the most used in the production of SBS estimates. In industry and services sectors there are about 500,000 limited enterprises which account for one half of the total employment. BIL data's coverage is 11.3% among 1-19 persons employed size class, it reaches 80.7% in the size class 20-99 and it is 96.2% among larger enterprises.

Referring to the economic variables, the analysis of the relationships between SBS variables and those included in BIL has highlighted in some case differences concerning the definitions. Istat actually buys the BIL data set from the Chambers of Commerce. The data set contains about 180 variables from profit-and-loss accounts, balance sheet and part of the explanatory notes. This source has been judged positively, both from the point of view of the information quality and of its stability. Also the rules governing the construction of the financial statements look stable. In this work we used the SME and BIL data that referring to the year 2002.

3. Description of the linkage experiment

Record linkage aims at identifying whether two (or more) records represent the same entity. The record linkage process may be decomposed in several phases: 1) pre-processing of the input files, 2) choice of the matching variables, 3) choice of the comparison function, 4) creation of the search space of link candidate pairs, 5) choice of the decision model and selection of unique links and 6) record linkage evaluation. For the probabilistic record linkage , we used the software Relais 2.0 (Cibella *et. al.* 2008).

The pre-processing phase aims at converting the input data in a well defined format, resolving the inconsistencies, to reduce the errors derived from an incorrect data registration. Since the enterprises with less than 19 employees are not obliged to present their balance sheet at the Chambers of Commerce, from both BIL and SME data sets only the enterprises with more than 19 employees were selected. Further, only the enterprises having a non missing *NACE* value were considered. Finally, in this preliminary experiment, only three regions (NUTS2) Veneto, Toscana and Sardegna were considered. The distribution of the observations in BIL and SME over the three regions is presented in Table 1.

Region	**BIL**	**SME**
Veneto	4762 (61.57%)	1428 (54.01%)
Toscana	2429 (31.41%)	935 (35.36%)
Sardegna	543 (7.02%)	281 (10.63%)
Total	7734 (100%)	2644 (100%)

Table 1. Frequencies (and percentages) of the observations over the three regions.

A light standardisation (*light std*) was applied on *NAME* and *ADDRESS* in both datasets. The characters judged "unusual" (',", =, $, # and double spaces) were deleted. The most frequent strings were standardized, too. For example, "S R L" or "S.R.L." were replaced by "SRL". In a second stage, more standardisation (*more std*) was applied to the streets typology and the types of the enterprises (SRL, SPA, SCARL, etc). Such standardisations depend obviously on the language and national legislation. Few tests were performed using only the first 10 characters of *NAME* and *ADDRESS* (*light std 10*). These tests were based on the assumption that the longer the string is, the greater the probability of misspelling/misspecification/etc is.

The matching variables should be chosen among the consolidated variables with a high identification power. In our experiments, we tested the usage of the following matching variables: *NAME*, *ADDRESS*, *MUNICIPALITY*, value of production (*VAL*), turnover (*RIC*), other income (*ARI*), use of third party assets (*GOD*), increase of fixed assets (*IMM*), costs of production (*COS*), purchases (*ACQ*), personnel costs (*PER*), wages and salaries (*SAL*), services (*SER*), value adjustment and depreciation (*AMM*), changes in work in progrESS (*LAV*), changes in stock of raw material (*MP*), changes in stock of finished product (*RIM*), profit/loss (*UT*). Three types of experiments were performed: a) on categorical variables, b) on numerical variables and c) on a mixture of variables. The categorical variables *NAME*, *ADDRESS* and *MUNICIPALITY* are generally non missing and they are quite heterogeneous. Consequently, a good identification power should be expected. In order to select the most reliable numerical matching variables some descriptive statistics have been analysed (see Tables 2 and 3).

Considering the economic relevance, completeness (number of zero values), presence of negative values and correlations, we first tested *VAL*, *COS* e *UT*. Afterwards *RIC* and *ARI* were used as components of *VAL*, and *ACQ*, *PER* and *SAL* as components of *COS*. Each variable was tested at least twice.

BIL/SME	***VAL***	***RIC***	***ARI***	***GOD***	***IMM***	***COS***	***ACQ***	***PER***	***SAL***	***SER***	***AMM***	***LAV***	***MP***	***RIM***	***UT***
VAL		1.00	0.34	0.13	0.08	1.00	0.94	0.38	0.20	0.51	0.35	-0.01	0.13	-0.02	-0.01
RIC	1.00		0.31	0.13	0.08	1.00	0.94	0.37	0.20	0.50	0.34	-0.04	0.13	-0.04	-0.01
ARI	0.78	0.77		0.06	0.03	0.34	0.32	0.21	0.10	0.24	0.25	0.00	-0.01	-0.04	-0.06
GOD	0.67	0.67	0.57		0.00	0.13	0.05	0.13	0.06	0.13	0.07	0.02	0.01	0.00	0.01
IMM	0.23	0.21	0.36	0.07		0.08	0.07	0.05	0.06	0.07	0.12	-0.58	-0.06	-0.02	-0.01
COS	1.00	1.00	0.77	0.67	0.22		0.94	0.37	0.20	0.50	0.34	-0.01	0.14	-0.02	-0.01
ACQ	0.91	0.91	0.69	0.39	0.16	0.91		0.27	0.15	0.29	0.27	-0.03	0.11	-0.02	0.00
PER	0.81	0.81	0.65	0.90	0.21	0.80	0.53		0.58	0.37	0.38	0.04	0.07	0.04	0.01
SAL	0.80	0.80	0.65	0.91	0.20	0.79	0.51	1.00		0.15	0.25	0.03	0.04	0.05	0.02
SER	0.79	0.79	0.62	0.67	0.27	0.80	0.53	0.81	0.80		0.27	0.05	0.10	-0.02	0.02
AMM	0.68	0.68	0.54	0.69	0.27	0.70	0.41	0.79	0.78	0.83		-0.05	0.02	-0.01	-0.10
LAV	0.12	0.08	0.06	0.05	0.01	0.13	0.08	0.05	0.04	0.21	0.17		0.07	0.01	-0.01
MP	-0.39	-0.38	-0.35	-0.06	0.01	-0.39	-0.54	-0.08	-0.07	-0.12	-0.06	-0.20		0.06	-0.01
RIM	0.05	0.03	-0.01	-0.01	-0.04	0.06	0.05	0.00	-0.01	0.05	0.10	0.27	-0.17		0.00
UT	-0.04	-0.04	0.00	-0.10	0.02	-0.09	0.04	-0.11	-0.12	-0.28	-0.50	-0.09	-0.07	-0.01	

Table 2. Correlations between the numerical variables in the BIL dataset (below the diagonal) and the SME dataset (above the diagonal).

	Min		1st Quartile		Median		3rd Quartile		Max		STD/MEAN	
	SME	BIL	SME	BIL	SME	BIL	SME	BIL	SME	BIL	SME	BIL
VAL	185	204	2758	5477	5486	10252	11356	22021	375619	5512071	1.82	4.16
RIC	26	0	2644	5260	5261	9920	10958	21237	372916	5366888	1.84	4.17
ARI	0	0	5	30	39	104	142	328	15089	154102	3.44	5.63
GOD	0	0	34	71	115	196	286	466	43745	539568	3.83	10.47
IMM	0	0	0	0	0	0	0	0	25088	36840	16.39	9.21
COS	192	502	2570	5262	5196	9793	10711	20930	373849	5240127	1.87	4.19
ACQ	0	0	418	1512	1867	4375	5325	11454	322429	4291727	2.63	4.90
PER	54	27	682	919	1061	1483	1677	2842	19811	1289835	0.67	4.84
SAL	0	0	252	637	440	1041	747	2005	2793	1015419	0.76	5.22

Table 3. Summary statistics on the numerical variables in the BIL and SME dataset

For each matching variable, a comparison function measures the “similarity” between two fields. Many comparisons are proposed in literature; the functions used in this work are described below, except for the equality function. For each comparison function, a threshold was chosen converting the results in binary elements, treating all the results above the threshold as 1 (match) and all the results below the threshold as 0 (non-match). More details may be found in Cibella *et. al.* (2008).

The numerical comparison of two values (strings) Nx and Ny is given by: $\mathrm{NC}(\mathrm{S}_x, \mathrm{S}_y) = \frac{\min(|\mathrm{N}_x|, |\mathrm{N}_y|)}{\max(|\mathrm{N}_x|, |\mathrm{N}_y|)}$.

The Levenshtein function is the basic edit distance: the minimum edit operations (copy, delete, insert and substitute) which transforms one string into the other. The Dice comparison function is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings. The Jaro function considers the characters s’ in s that are “common with” t, and the characters t' in t that are "common with" s. Let $T_{s',t'}$ measure the number of transpositions of characters in s' relative to t'. Then the Jaro similarity metric for s and t is defined by: $Jaro(s,t) = \frac{1}{3}\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{2|s'|}\right)$. The q-grams function is generally used in approximate string matching by “sliding” a window of length q over the characters of a string ,s, to create a number of 'q' length substrings for matching. We set q = 3. The Soundex function is a phonetic indexing scheme which generally focuses on individuals names. Each term is given a Soundex code consisting of a letter, the first one of the string, and five numbers between 0 and 6.

For the categorical variables, the threshold was always set to 0.8, except for the equality distance. Different combinations of the distance functions were tested to identify the best setting of the record linkage experiment.

To reduce the numerical complexity, it is generally necessary to reduce the number of pairs to be compared. Two reduction methods were applied in this experiment: blocking and sorted neighbourhood (SN). When testing only the categorical variables, several blocking schemes were applied. The region *REG* (3 categories) was always used as a blocking variable since this information was considered very stable. Tests were also performed using the province *PROV* (21 categories) and the principal economic activity NACE 2-digit (53 categories) as blocking variables. The regional blocking for numerical variables did not lead to any results. We then applied the SN approach using *RIC* as sorting variable as it was considered as the most relevant economic numerical variable. When using both categorical and numerical, the SN approach was applied as a reduction method, while *PROV* was included among the matching variables, using the equality distance.

The applied probabilistic model follows the Fellegi-Sunter approach (Fellegi, 1969). For each pair, the matching variables comparison is summarized in the comparison vector γ. Then the probability distribution of the comparison vector is assumed to be a mixture of two distributions: one for matches (m) and one for non-matches (u). Generally, the conditional distribution estimates can be obtained via the EM algorithm (Jaro, 1989). According to the Fellegi and Sunter theory, once the composite weight r = m(γ)/u(γ) is estimated, a pair is classified as link if the corresponding weight r is above a certain threshold T_m, and as a non-link if the weight is below the threshold T_u. In this work, the thresholds were derived from the probabilities of false non-match (0.90) and false match (0.95). Finally, the reduction one-to-one was solved as a linear programming problem (Jaro, 1989).

In both BIL and SME datasets, there is a unique identifier, namely the fiscal code. Even if the fiscal code may be subject to some errors, it was used for evaluating the quality of the record linkage through *precision* and *recall*. With respect to the fiscal code, there are 1540 common units (true matches) in BIL and PMI. 55.1% of these units are located in Veneto, 35.3% in Toscana and 9.6% are located in Sardegna.

4. Main results

The exact matching (merge) using only *NAME*, *ADDRESS* and *MUNICIPALITY* gave 218 units (61% in Veneto, 29.8% in Toscana and 9.2% in Sardegna).

The results obtained using NACE 2-digit as blocking variable are summarized in Table 4. The descriptive statistics were computed over the tests defined by varying the comparison distance on *NAME*, *ADDRESS* and *MUNICIPALITY*. Even if the precision exceeds 80% in more than 50% of tests, since the maximum recall is about 62%, it means that there are 40% of true matches that are not identified. Over the 1540 common units in BIL and SME, 282 have a different NACE 2-digit value. Generally we noticed that in about 4% of cases there is a difference between the NACE 2-digit value observed in SME and the corresponding field in the previous version of BIL. We then concluded that NACE 2-digit could not be a reliable blocking variable.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Light Std (26 tests)	Precision	63.4	72.4	83.4	80.2	84	92.9
	Recall	21.3	31.7	43	41.9	48.4	57.4
Finer Std (19 tests)	Precision	69.9	75.1	80.5	79.9	83.8	90.3
	Recall	38.2	50.8	54.1	53.9	60.6	61.8

Table 4. Descriptive statistics of precision and recall (%) over the tests using NACE 2-digit as blocking variable.

We also performed a series of tests using *PROV* as blocking variable. As before, we varied the distance function on *NAME*, *ADDRESS* and *MUNICIPALITY*. The results are presented in the Figure 1.

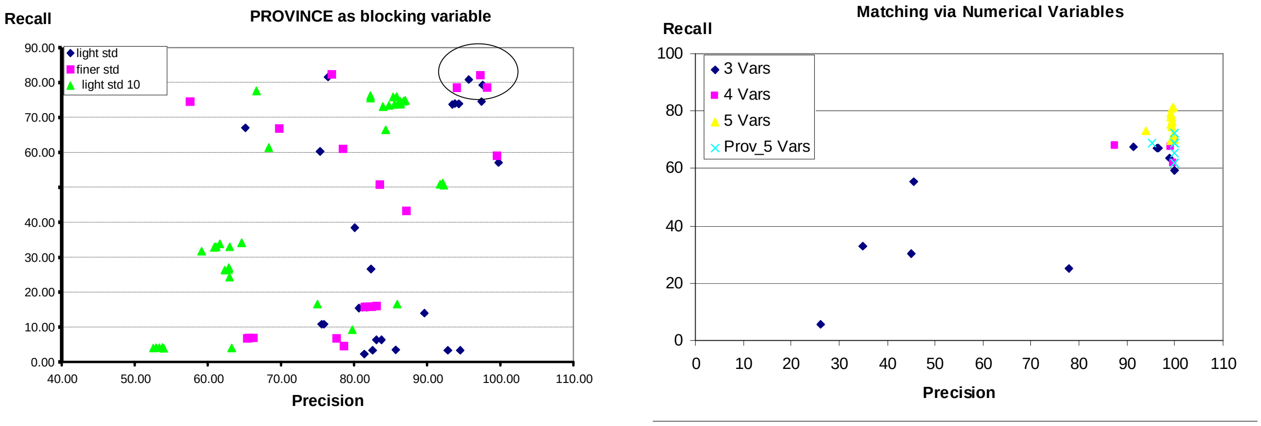

Figure 1. Results obtained using *PROV* as blocking variable (left) and numerical variables (right).

From figure 1, several issues may be observed. First of all, only the first 10 characters of *NAME* and *ADDRESS* are not sufficient to identify an enterprise. Anyway, since the recall is improved, this idea could be further investigated. Second, the results obtained using more standardisation are generally better than the ones obtained using the light standardisation, both in terms of recall and precision. Third, the distribution of the best results (circled area) deserves some attention. In all the best 5 tests, the Soundex comparison function was used for both *ADDRESS* and *MUNICIPALITY*.

We conducted more the 80 tests on numerical variables. Generally, we focused on the SN approach using the Numerical comparison distance. Different windows sizes (10, 20 and 50), thresholds (0.8 and 0.95), and matching variables have been tested. Finally, some tests on a single Province were performed. Increasing the window size or reducing the threshold worsens the results. Relaxing the similarity criteria (reducing the threshold) or increasing the cardinality of the search space (increasing the window size), the number of false matches increases. A too narrow window or an extreme threshold value could reduce too much the search space or make difficult any match identification. We then adopted the window size 20 and the threshold 0.95.

Increasing the number of matching variables, the results get better, because false matches are more easily identified as such. In this sense, the identification power of the numerical variables is extremely significant. Given the 'best' combination of 4 matching variables we varied the fifth one. The comparison between the fifth variables gave almost the same results in terms of precision and recall (see Figure 1).

Compared with merging and nearest neighbour strategies, the probabilistic approach seems to be more promising. Merging, for the numerical variables, is quite unreasonable while the nearest neighbour approach is comparable in terms of recall but it results in lower levels of precision as it does not identify the non-matches.

In a next step, instead of varying the thresholds on the comparison functions, we concentrated on a strategy based on both categorical and numerical variables. For the mixed strategy, we used as starting point the settings derived from the best results obtained in the previous approaches. Namely, we used the Soundex function for *ADDRESS* and *MUNICIPALITY*, the SN on *RIC*, window size equal to 20, the threshold on the Numerical Comparison function for the variables *ARI*, *ACQ*, *PER*, *UT* and *MP* equal to 0.95. The results are shown in figure 2. It may be noticed that the mixed strategy improves both the precision of the strategy based on categorical variables and the recall of the strategy based on the numerical variables. It is hard to believe that a precision better than 99.9% could be obtained. From the recall point of view, we obtained good results, but further improvements might be obtained using a more suitable standardisation on the categorical variables. It should be observed that, when reducing the number of numerical matching variables, a better standardisation of the categorical variables produces the aimed results: increase the recall.

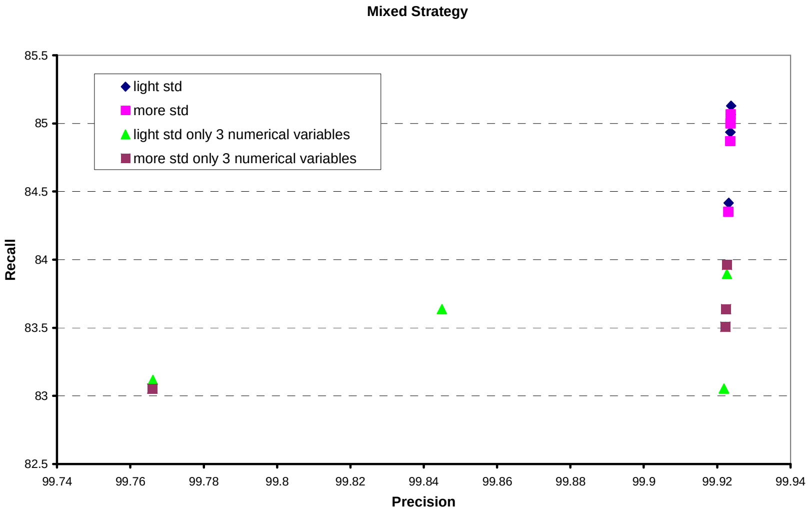

Figure 2. Results obtained using a mixture of categorical and numerical matching variables.

5. Conclusions and further work

We reported the results of a linkage experiment using the survey and administrative data stemming from the Italian SME and BIL archive, respectively. Several tests have been conducted in order to identify an optimal linkage strategy based on: search space reduction, selection of matching variables, selection of similarity distances and relative thresholds. The Fellegi-Sunter probabilistic approach has been applied. The best results have been obtained by a mixed strategy using categorical and numerical variables. As far as textual variables are concerned, the data pre-processing seems of great impact. Numerical variables show a high identification power. Consequently, if available, they might be considered very reliable.

The obtained results are a preliminary step in looking for an optimal strategy of linkage. Therefore, further experiments are planned. Our attention will be focused on: (i) improving methods of data parsing and standardisation; (ii) settings of thresholds and selection of numerical matching variables; (iii) bench-testing the approach (comparison with deterministic approach, assessment of the quality of fiscal code).

6. References

1. Bethel, J. “Sample allocation in multivariate surveys”. *Survey methodology*, 15 (1989): 47-57.
2. Deville, J.C., e C.E. Särndal. “Calibration Estimators in Survey Sampling”. *Journal of the American Statistical Association*, 87 (1992): 376-382.
3. Fellegi I. and Sunter A. (1969) “A Theory for Record Linkage”. *Journal of the American Statistical Association*, 64, 1183-1210.
4. Cibella N., Fortini M., Scannapieco M., Tosco L., Tuoto T., “*Theory and practice of developing a record linkage software*”, Workshop "Combination of surveys and administrative data" Vienne 2008.
5. Jaro M. “Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa, Florida”. *Journal of the American Statistical Society*, 84 (406), 414 – 420, 1989.