

Automated Coding System: An Application To The ISTAT School-to-work Transition Survey

*Codifica Automatica: Un'applicazione Sull'Indagine ISTAT
Sull'Inserimento Professionale Dei Maturi*

Manuela Murgia, Simona Pace, Paola Ungaro
ISTAT, Roma, Via C. Balbo 16 (00184)
murgia@istat.it, sipace@istat.it, ungaro@istat.it

Riassunto: Il presente documento analizza le problematiche riguardanti la codifica della variabile “Professione” rilevata nell’indagine Istat su “Percorsi di studio e di lavoro dei diplomati”. In particolare viene analizzata la convenienza in termini di tempo-costo e qualità del passare dalla codifica manuale alla codifica automatica

Keywords: Automated coding, manual coding, occupation

1. Description of the survey

ISTAT currently conducts three surveys on post-compulsory school and university first level graduates, in order to study the participation of different educational levels in the labour market. These three surveys belong to an integrated system: they use questionnaires similar in structure and content and they are based on the same methodological approach. They are carried out every three years on a single students cohort and three years after graduation. The main section of the questionnaire focuses on the respondent’s current employment, which is described by many variables such as “occupation”, “professional position”, “income”, “required qualification”, “industry”, etc.. Information regarding parents’ employment is collected in the family background section. “Occupation” is a significant variable used to compare the performance of graduates from different fields of study in the labour market. It is collected by means of an open question. The textual answers are manually coded according to Istat Occupational Classification System (*Classificazione delle Professioni – CP91*).

This paper reports the main results of the application of automated coding by the ACTR system to the variable “occupation” in the “Upper Secondary School Graduates” survey. Interviewees’ and their fathers’ occupations are separately analysed. A comparison with manual coding is also performed to evaluate the quality of the two processes.

2. ACTR v3 system

The logic behind of ACTR was inspired by methods originally developed at the US Census Bureau (Hellerman 1982). ACTR uses matching algorithms studied by Statistics Canada researchers (Wenzowski 1988). ACTR is a generalised coding system as it is independent of the language and the classification used. The automated coding activity is preceded by a sophisticated text standardisation phase, called “*parsing*” which

provides 14 different “parsing functions” able to remove grammatical or syntactical differences so that any two different descriptions, with the same semantic content, become identical. The parsed response to be coded is then compared to the parsed descriptions in the dictionary. If a perfect match is found in this search, then a unique code is assigned, otherwise the software runs an algorithm to find the most suitable partial matches. In the latter case, the software selects from the dictionary all the descriptions that have at least one parsed word in common with the input verbal phrase and assigns them a score calculated as a function of the weight of each word in common (the weight of each word is inversely correlated to its frequency of occurrence in the dictionary). It then orders the extracted descriptions in descending score rank, and compares them with user-defined threshold parameters. As a result the software returns:

- unique matches, when a unique code is assigned to a response phrase: unique winner
- multiple matches, when several possible codes are proposed: multiple winners
- possible matches, when several possible codes are proposed but whose level of similarity to the input description is not so high as to guarantee the presence of a correct code among them: possible winner
- failed matches, when no matches are found: no winner.

While in the first case there is no need for human intervention, all the other cases need to be evaluated by expert coders. This can be done by the “on-line” coding function that helps the operator to search for a winner code among a list of dictionary descriptions that match with the input response. To use ACTR for a given classification, it is necessary to create a dictionary following two steps: *i)* revision of the official classification manual that aims to transform each description associated to each code, so as to be absolutely synthetic, analytic and unambiguous and therefore suitable to be processed by a computerised system, *ii)* enrichment of the manual by selected descriptions derived from empirical responses of previous surveys and already coded by experts of the classification, in order to make it closer to the spoken language.

3. Main results

Table 1: *Automated coding results*

	Respondents’ job		Fathers’ job	
Unique codes	8358	55,0%	14391	64,2%
Multiple codes	1020	6,7%	1106	4,9%
Possible codes	1562	10,3%	2259	10,1%
Failed codes	4235	27,9%	4660	20,8%
Empty	8	0,1%	9	0
Total records	15183	100	22425	100

The table 1 shows a good success – percentage of unique codes – of ACTR for both cases and especially for “fathers’ job” where a higher percentage of unique codes as well as a smallest number of failed codes are obtained. This could be due to: a) the input descriptions to be coded are more synthetic for fathers than for respondents and therefore closer to the building criteria of the dictionary (see above): this results in a higher percentage of success; 2) respondents’ jobs are more innovative than fathers’ and are not always included in the dictionary based on the 1991 occupational classification: this increases the percentage of failed matches for respondents’ jobs.

Referring to unique codes, a comparison between automated and manual coding output has been carried out, as shown below.

Table 2: *Unique codes – comparison with manual coding*

	Respondents' job			Fathers' job		
Unique codes = Manual codes	6651	79,6%	(100)	11190	77,8%	(100)
• <i>matching score</i> = 10	5270	63,1%	(79,2%)	8861	61,6%	(79,2%)
• <i>matching score</i> < 10	1381	16,5%	(20,8%)	2329	16,2%	(20,8%)
Unique codes ≠ Manual codes	1707	20,4%	(100)	3201	22,2%	(100)
• <i>matching score</i> = 10	739	8,8%	(43,3%)	1787	12,4%	(55,8%)
• <i>matching score</i> < 10	968	11,6%	(56,7%)	1414	9,8%	(44,2%)
Total of unique codes	8358	100		14391	100	

The great majority of unique codes (79,6% and 77,8%) perfectly corresponds to the manual ones. Besides, most of them have a matching score equal to the maximum - score 10 – i.e. the input description is identical to the dictionary one (63,1% and 61,6%). To evaluate ACTR performance versus manual coding, unique codes that do not coincide with manual ones have been analysed. A first measure is soon obtained by the percentage of non-coinciding codes with score 10, that are correct by default (De Angelis & Others, 2000). This represents the 43,3% and the 55,8% of the total amount of non-coinciding unique codes respectively for respondents and for fathers (revealing a wrong manual coding).

To measure the quality of non-coinciding codes⁽²⁾ with score less than 10 we analysed the input textual descriptions of a sub-set of them⁽³⁾ represented by automated codes with a number of digit less than or equal to the corresponding manual code. In fact, we assumed that ACTR performs better than manual coding in providing a more complete information when it produces codes with a greater number of digit.

We found that ACTR correctly coded the input description for 39,2% for respondents' occupations and 38,8% for fathers'.

Considering both the unique codes with matching scores 10 and the latter results, ACTR produced better output than manual coding for 56,6% interviewees' occupations and for 67,0% fathers' occupations.

4. Conclusions

The main results of the analysis are the following:

- Human intervention is limited to a restricted number of cases: the “non-unique codes” (i.e. multiple, possible and failed codes) which amount to the 45% of respondents' and 35,8% of fathers' jobs. The descriptions that have to be manually coded, being a restricted number, can be treated by expert coders. This increases the quality of the results reducing the risk of assigning different codes by different coders to the same occupation descriptions.
- In a high percentage of cases (43,8% of the total amount of respondents'

⁽²⁾ We considered as coinciding those codes with at least three coinciding digits.

⁽³⁾ The analysis has been based on 581 records among 968 with score less than 10 for respondents and 919 among 1414 for fathers

occupations and 49,9% for fathers' jobs) ACTR performs as well as manual coding, i.e. unique codes and manual codes are the same. From the organisational point of view the advantages of adopting automated coding are clear: in this survey the time spent in interviewer training (about 15 days) and coding activity (approximately estimated in 5 minutes each occupation) is higher than that necessary for the automated coding system that codes the majority of records in batch process.

- c) In some cases (6,4% for respondents' occupations and 9,6% for fathers' jobs) ACTR performs better than manual coding, improving the quality of results.
- d) In the 4,9 and 4,7% of cases manual coding performs better than ACTR. These small percentages could be further reduced.

Table 3: Main results

	Respondents' job		Fathers' job	
Non unique codes	6825	45,0%	8034	35,8%
Unique codes = Manual codes	6651	43,8%	11190	49,9%
ACTR performs better	967	6,4%	2144	9,6%
Manual coding performs better	740	4,9%	1057	4,7%
Total	15183	100,0%	22425	100,0%

This experience shows that an automated coding system can provide higher quality results rather than manual coding in terms of accuracy, efficiency and standardisation of the process. To increase the percentage of success of ACTR system it is necessary *i)* to organise an *ad hoc* training for the interviewers (to allow the collection of shorter input descriptions that could better match with the dictionary ones), *ii)* to update the dictionary, training it with other school-to-work transition survey data which includes occupations of specific targets within the Italian population.

References

- De Angelis R., Macchia S. and Mazza L. (2000) Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale, *Istat Quaderni di ricerca – Rivista di statistica Ufficiale*, 1, 29-54.
- Hellermann E. (1982) *Overview of the Hellerman I&O Coding System*. Internal document, US Bureau of the Census, Washington.
- Micali A., Ungaro P. (1998) Il sistema di indagini Istat sulla transizione scuola-lavoro. *SIEDS Rivista italiana di economia e statistica*, LII, 311-324.
- Tourigny J.Y. and Moloney J. (1995) *The 1991 Canadian Census of Population experience with automated coding in Statistical Data Editing*, 2, United Nations Statistical Commission.
- Wenzowski M.J. (1988) ACTR – A Generalised Automated Coding System. *Survey Methodology*, 14, 299-308.