DATA EDITING AND QUALITY OF DAILY DIARIES IN THE ITALIAN TIME USE SURVEY*

Salvatore F. Allegra, Barbara Baldazzi
Istat
allegra@istat.it, baldazzi@istat.it

Abstract

In the Italian Time Use Survey, carried out by Istat (Italian National Statistical Institute) in 2002-2003, the coding process translates the sentences reported by the respondents into codes, but association text-code is not a one to one easy linkage, because the respondents describe the activities performed using the common language. So, a process of data editing is requested in order to improve data quality. The possibility to use sentences (recorded in the data entry) has definitely improved the data editing process.

The main questions to analyze concern:

a) particular ancillary codes used by coders to point out daily diary's problems;

b) incompatibility between codes and codes and between codes and text;

c) errors in the intervals' sequences.

This paper shows methodological and contextual choices in order to build the process of data editing. In particular, it is focused on the analysis of techniques and deterministic rules adopted and on the analysis of imputation system to correct adequately the wrong codes, with reference to the entire diary.

* This paper was presented at the XXVI IATUR (International Association for Time Use Research) Conference held in Rome, 27-29 October 2004. The authors share contents and views expressed in this paper. However, the paragraphs are so drafted: Salvatore F. Allegra drew up the paragraphs 1, 2, 3, 7; Barbara Baldazzi drew up the paragraphs 4, 5, 6, 8. Salvatore F. Allegra and Barbara Baldazzi drew up the paragraph 9.

Index

1. Introduction

The multipurpose survey on households called "Time use survey" has been carried out by Istat (Italian National Statistical Institute). A sample of 21,075 families was interviewed, summing up to about 55,000 individuals, and data was collected from April 2002 to March 2003 by PAPI technique.

The daily time dedicated to different activities was reported by the respondents aged 3 and over in daily diaries. These activities (main and parallel activities) were described in the shape of free text – very often in the shape of sentences – 10 minutes by 10 minutes, as well as the locations in which the activities took place and the modes of transport used by the respondents to move from a place to another place. The respondents were asked to indicate with whom time was spent, for each interval of 10 minutes, ticking the respective box[1].

Even though the natural language is subject to shared rules, it expresses meanings which can change depending on the context. So, sometimes the description of the activities in the natural language is not enough to code the activities; since it needs to take into account the context where every activity takes place (see Bolasco, 1997; Camporese et al., 2001; Romano, 2004a). In other words, the meaning is conditioned by various dimensions: for instance, by the succession of the activities, by their purposes, by their locations and their daily schedules, by the seasons in which they are performed and by the linguistic skills of the respondent (Cappadozzi et al., 2001).

Finally, the perception of time is not the same for all respondents. Activities lasting the same time can be perceived in a different way. Time is made elastic by respondents' perceptions. Perceptions define the segmentation of time in hours and minutes, and the sequence or the simultaneity of the activities as well (Camporese et al., 2001).

The activities, the locations and the modes of transport were translated into codes by coders trained for this purpose. The system of classification codes proposed by Eurostat in the guidelines on harmonised European Time use surveys (Eurostat, 2000) was revised by Istat in order to fit it to the Italian requirements and specifics (Istat, 2002a).

Furthermore, in order to prevent non-sample errors in the process of coding, an assisted coding was carried out by the software called "Blaise" (Cappadozzi et al., 2001) and a continuous monitoring of coding work was implemented. The coding operations were centralised and constant assistance and supervision were offered to coders by a team of the Time use survey unit (Cappadozzi et al., 2003; Romano, 2004a; Romano et al., 2004b; Romano et al., 2004c).

[1] Specifically, the respondent was requested to indicate if he/she was alone, or if he/she was with household members (aged less than 10 or aged 10 and over), or if he/she was with non-household members, or if he/she was with other individuals.

This paper shows the methodological choices and the operational steps in order to make the process of data editing better and, therefore, to improve the quality of data collected by the daily diaries. In particular, this paper is focused on the analysis of the deterministic rules and non-automatic rules adopted.

2. *The importance of recording strings for the process of data editing*

Because of their complexity, the corrections have been divided into different steps: deterministic corrections and non-automatic corrections have been carried out. These steps have been integrated in a continuous process of data editing.

Before coding, the texts (sentences) concerning main activities, parallel activities, activities' locations and modes of transport used by the respondents, were recorded in a literal way by recorders trained previously. Therefore, a considerable number of strings has been obtained for each activity (main and parallel activity), activity location and mode of transport.

The possibility of using texts has definitely contributed to make effective corrections where the association between text and code was not a one-to-one linkage.

Some figures about the volume of strings obtained and the average length of strings are reported in table 1. In the same table, figures about the quantity of words produced and the average frequency of words are shown.

Table 1 – Strings and words by type of information

	Main activity	Parallel activity	Location of the activity/ Mode of transport
Strings	1,457,246	414,979	1,221,372
- different strings	*240,252*	*58,345*	*31,791*
Average length of strings	17.87	18.59	9.27
Occurrences (words)	4,973,359	1,477,047	2,887,369
- different words	*29,478*	*12,960*	*9,292*
Average frequency of words	168.71	113.97	310.74

Source: *Istat, "Time use survey 2002-2003"*

The greatest number of strings and the greatest number of occurrences (words) are reported in the first column of the table: 1,457,246 strings and 4,973,359 occurrences were recorded for the main activity. The location of the activity/mode of transport sums up to 1,221,372 strings and to 2,887,369 occurrences; the numbers of strings and occurrences are lower than the figures reported

by the main activity, but they are much higher than the ones concerning the parallel activity (414,979 strings and 1,477,047 occurrences). Main activity also shows the greatest number of different strings (240,252) and the greatest number of different words (29,478); while, the average length of strings (17.87) is a bit shorter than the one shown by the parallel activity (18.59). Finally, the highest average frequency of words – that's to say, the ratio between the occurrences and the different words – is reported in the third column, the one concerning the location of the activity/mode of transport, and is equal to 310.74.

The decision of recording texts has its source in the consideration of the benefits of the strings for the process of data editing: the strings have a fundamental role to identify – for instance – the errors between the texts (sentences) and the respective codes, as well as to identify the interruptions in the succession of the activities (the missing texts between two episodes[2]).

In order to plan and implement the correction rules, the recording of strings has been a fundamental decision both from a strategic point of view and from an operational point of view.

3. *The importance of preventing errors in the process of coding and the necessity of introducing the ancillary codes*

The aim of the assisted coding system was to allow coders to select the correct codes concerning the strings about activities, locations of activities and modes of transport reported in the daily diary. "Blaise" provided an assisted coding module based upon two types of codes' research: 1) a tree of codes; 2) a dictionary. They could be used separately or jointly (Cappadozzi et al., 2001).

Furthermore, information on the respondent – sex, age, marital status, level of education, activity status, profession, sector of economic activity, relationship to the other household members, number of household members –, as well as information about the date of filling in the diary and the day of the week, was reproduced on the template designed in order to translate the strings into codes (Istat, 2002b).

The decision in favour of an assisted coding system was originated by the complexity of the process of coding. Actually, before assigning the correct codes, coders have to understand the frame in which activities take place, using various pieces of information: the previous episodes, the following episodes, besides – of course – the location of the activity/mode of transport. So, if the objective was to prevent errors in the process of coding, it would have been dangerous the recourse

[2] An episode comprises one or more intervals of 10 minutes in which the main activity, the parallel activity – if present –, the location/mode of transport and individuals with whom time is spent by the respondent do not change (Eurostat, 2000).

to a completely automatic coding system, in which the code would have been attributed in a univocal way (Camporese et al., 2001; Cappadozzi et al., 2001).

In order to make the process of coding better, another strategic decision was taken: to point out particular coding problems by special types of codes, called "ancillary codes" (Istat, 2002a). In other words, even though the respondents had been requested to fill in the daily diary following precise rules, the description of some events would have surely produced undesirable effects, originating possible coding errors.

Table 2 – Some examples of ancillary codes

Ancillary code	Coding problem to be pointed out
11	Main activities in succession reported in the same episode
13	Simultaneous activities reported in the box of the main activities
21	Activity and travel activity reported in the same episode
31	Very short events related to activities
32	Very short events related to travel activities
62	Activities in contrast with the location of the activity/mode of transport
63	Activities in contrast with individuals with whom time is spent

Source: *Istat, "Time use survey 2002-2003"*

For instance, critical events could have been: main activities in succession reported in a single box, that's to say in the same episode; as well as, simultaneous activities with no distinction between main activity and parallel activity, but reported in the box of the main activities.

Moreover, sometimes, the respondents did not make a distinction between the activity in itself and the travel activity, describing everything together in the same episode and reporting in the respective box neither the location of the activity nor the mode of transport, or reporting only one of these. The respondents frequently described very short events related to activities too: these events lasted less than 10 minutes and they often indicated the beginning or the ending of an activity; very short events are also related to travel activities, indicating the beginning point of a successive travel episode or its ending.

Finally, it was very important to point out that the reported activities (main or parallel activities) were in contrast with the location of the activity or the mode of transport; or those activities were in contrast with individuals with whom time was spent by the respondents.

In presence of crucial events – like the previous ones – coders were trained to choose the correct ancillary code from a list previously defined and to put it in an apposite box, in order to

point out the particular coding problem and to let the researcher find the best solution in the data editing successive phase. Some examples of ancillary codes are shown in table 2.

4. *The errors and the checked incompatibilities*

The variety of errors is wide; yet it is very difficult to make a classification. The easiest errors to discover are:

- ✓ typing errors (presence of codes with no significance);
- ✓ redundant information (same code on main activity and parallel activity);
- ✓ presence of episodes with partial non-response in activities or in location/mode of transport or in people involved.

Also, easy enough to be found is:

- ✓ incongruity between main activity and parallel activity, between activity and location/mode of transport or between activity and people involved.

More difficult is to understand when there is:

- ✓ an incongruity in the sequence of activities;
- ✓ an improper use of codes (for example, the use of the code 2 - study - for who works at school);
- ✓ a coding error.

These last types of errors are the majority. In these categories are included context errors: attribution of codes reserved to adults instead of the ones for children (for example, the code concerning a child playing with someone is 73, but the code concerning an adult playing with a child is 38); attribution of activity codes reserved to adults instead of activity codes reserved to children (for example, the code 39 that indicates help to an adult household member, used instead of the code 38 – childcare); codes of activities carried out with other people done by oneself (for example, the code 511 – socializing with household members used by singles).

Last but not least, we can consider the following errors:

- ✓ lack of coherence in all the diary (for example, diaries with changes of location without recording any travelling).

5. *The correction process for improving the data quality*

The correction process – still in progress – can be divided into two sections:

1) a deterministic correction based on SAS procedures;
2) a non-automatic correction based on a SAS/AF template.

In the first section, a series of SAS procedures automatically determines corrections on the episodes. When the conditions are exhaustive and they are not ambiguous to determine a correction, the SAS programs edit the correction on the data base and, at the same time, print a before-after corrections' report. Only the accurate analysis of this report permits the validation of the correction rules. The report is organized to point out the episodes with errors and the successive corrections, and it also shows the two previous and two following episodes , in order to keep under control not only the singular correction, but also the accuracy of the episodes' sequence.

When the correction rules create ambiguous corrections (for example, the singular correction can appear right, but the episodes' sequence is contorted), or when the same error indication creates two or more different types of correct editing, or when the correction is based on the analysis of the contextual information, a SAS/AF template is used for editing non-automatic corrections.

In figure 1 the correction process for a code X is illustrated: the SAS procedures for deterministic rules are two and more for every code and the program sequence is based on the analysis of the rules' priority: some rules need to process before others; some rules are consecutive to others (for example, the corrections based on the location of the activity/mode of transport must follow the correction of the location/mode). Every step of the deterministic correction produces a corrected database and prints a before-after corrections' report.

6. *How can we implement a rule?*

The first step to design a rule is the implementation of a checking plan to highlight errors. As we have said in the paragraph 4, there are many types of error; some of these concern the completion of the diary – that's to say the partial non-response –, other errors concern the coding of activities, the coding of location/mode of transport, and the coding of the individuals with whom time is spent by the respondents.

When a type of error has been found, a checking plan, developed with SAS procedures, prints episodes with errors. A careful analysis shows salient points of incongruence between the activity and its location/mode of transport, between the previous activity and the following one, and so on.

The reports of the check SAS procedures are analysed in order to pinpoint the rules. Every interaction of codes with codes, or codes with sentences, is used in order to formularise rules. Some instances are: a) the use of the previous activities' codes and/or the following activities' codes; b) the use of the sentences (the previous ones and/or the following ones); c) the use of the location/mode of transport (the previous ones and/or the following ones); d) the use of the

individual characterizations (for example, knowing the age of the respondent and knowing that some types of codes have to be used only for coding children activities, we are able to make the adequate correction); e) the use of household information (for example, knowing that the childcare code is different if this care is given by parents to help their children or given to help children of other households, we are able to make the adequate correction); f) the use of ancillary codes. They can be used separately or jointly.

At the beginning of the correction process, we utilised some general rules in order to correct some typical errors which could be found in each group of activity codes: presence of sentences with no significance, presence of codes with no significance, missing texts and typing errors too.

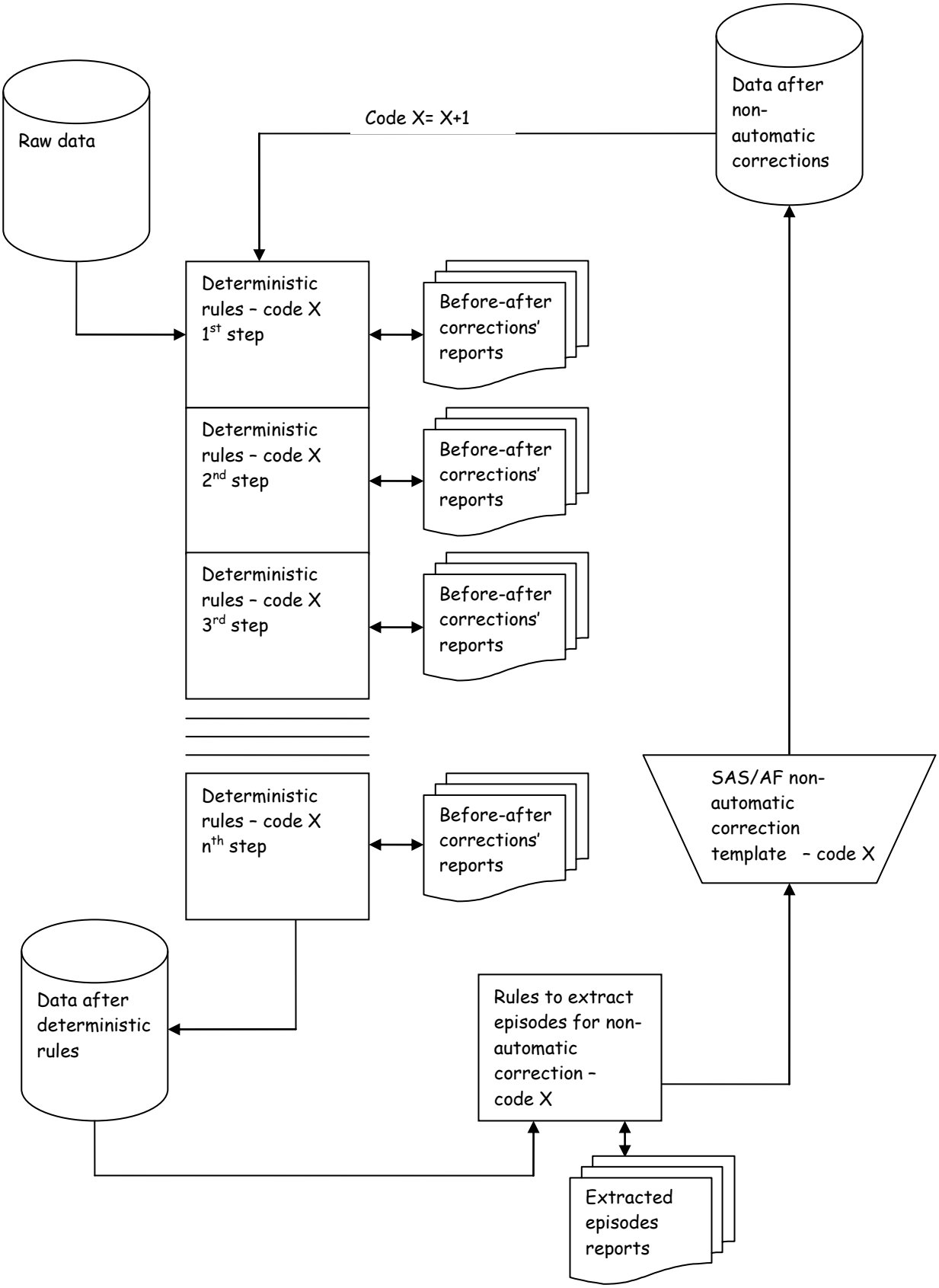
For example, because the code 821 – "watching television" – and the code 812 – "reading books" – are similar, sometimes, the coders reversed 1 with 2 and 2 with 1. Recording strings previously resolves this kind of error: indeed, when there was code 821 without the keyword "television" or "TV" and, in the same sentence, there were the keywords "read" and "book", we could correct, putting the adequate code. Reading the before-after corrections' report, we found out that writing sentences can often produce errors. Indeed, coding the sentence "I've switched off TV and I'm reading a book" with '821' is wrong. Therefore, the rule has been made sharper by looking for the keywords "switch off". This is only an example about the difficulties met in order to make perfect correction rules.

The correction rules become more complicated when it is necessary to correct specific codes. At this time, the recourse to the analysis of the previous episode and/or the following episode is fundamental. By means of this analysis we have been able to edit a previous/following code for each episode: for example, when the location/mode of transport code is missing, but the activity code is the same of the previous activity code, we have edited the previous location/mode of transport code.

The use of sentences, concerning the main activity, the parallel activity and location/mode of transport, is very wide in order to make the correction rules: not only the sentences concerning a specific episode, but also the sentences of the previous episodes or the sentences of the following ones. The use of the sentences has been possible by using keywords – only one keyword or the association of more keywords – that must not be ambiguous.

Also, we have analysed the presence/absence of one code (or one word) in the entire diary: for example, in order to pinpoint diaries without meals, or in order to define travels in the right way.

Figure 1 – The correction process



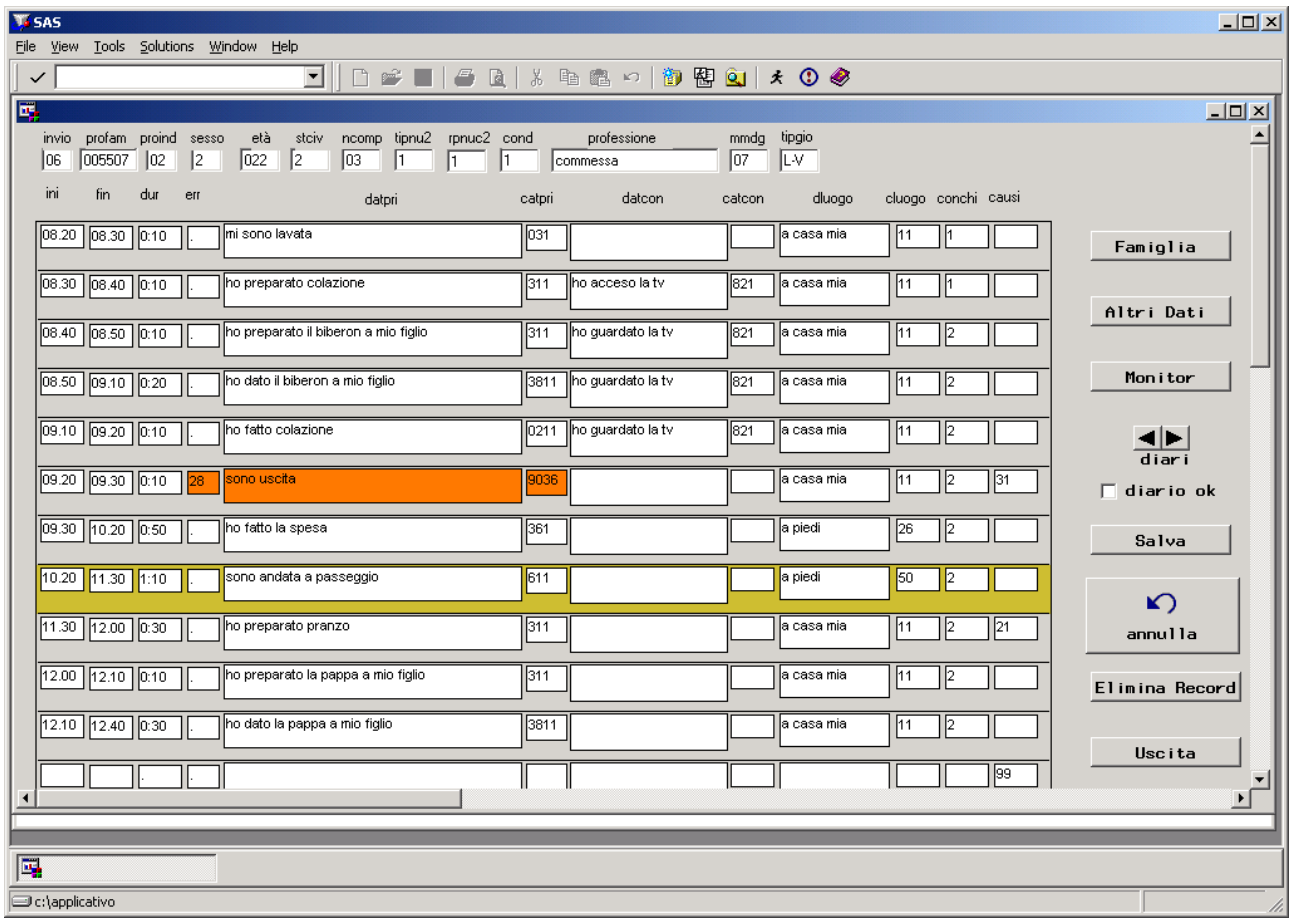
7. *The non-automatic corrections: the template*

Relating to the SAS/AF template used by a trained staff of coders, it needs to be reminded that the non-automatic correction has been possible because of the recording of the respondents' sentences. It needs to be also precised that the objective is to make corrections in those episodes extracted according to certain rules individuated after the deterministic correction.

Figure 2 shows an example of the structure of the SAS/AF template.

On the template, each episode showing one or more incompatibilities (for instance, between the main and the location/mode of transport; or between the main activity and the parallel activity; or between the main activity and the person with whom time is spent by the respondent) is marked by an orange strip which highlights, respectively: the error's type, pointed out by a code standing for the rule that must be applied (column "err"); the recorded string concerning the main activity (column "datpri"); the code of the main activity (column "catpri").

Figure 2 – The SAS/AF template used for non-automatic correction



Moreover, on the template, the previous five episodes and the following five ones are always shown, as well as information on the duration of the episode (column "dur"), information about the respondent (sex, age, marital status, activity status, profession, number of members of his/her own household, relationship to the other members of his/her own household), eventual number of children of his/her household (button "altri dati"). Information about the month of filling in the diary and the type of the day of the week (working or non-working day) are reproduced in the template too.

In addition and if necessary, the coder may look at the diaries of the other members of the respondent's household (button "famiglia"), in order to check the activities carried out in the same time intervals by the respondent with someone of them, so that the coder is able to solve particularly complex cases and to correct them adequately. Sometimes, the coder can be also helped by the ancillary codes, indicating coding problems (column "causi").

At the end of every correction session, the template shows the corrections' situation at the moment: specifically, the number of all diaries, the number of corrected diaries, the number of diaries to be corrected.

8. Monitoring corrections: some results

The correction process relating to two groups of activity codes is shown in the table 3.

It reports the number of deterministic rules and the number of rules of extraction for non-automatic corrections concerning, specifically, the code 1 (employment) and the code 2 (study).

The deterministic rules used for correcting data are 123, but, the rules concerning employment are more than those rules concerning study.

The main process indicators concerning the employment activity are reported in table 4. They are calculated both on the total corrections and separately on the deterministic corrections and on non-automatic corrections.

Table 3 – Deterministic rules and rules of extraction for non-automatic corrections by two groups of activity codes

	Code 1 – Employment	Code 2 – Study	Total
Number of deterministic rules	79	44	123
Number of rules to extract diaries for non-automatic correction	49	39	88

Source: *Istat, "Time use survey 2002-2003"*

Out of 86,267 episodes, at this time, 20,385 episodes (23.6%) have been corrected with at least one code. Really, the number of corrections is 26,278, equal to 1.29 corrections per episode: that's why one or more corrections can be made on the same episode. In other words, one or more rules can point out a wrong episode and the correction acts on the main activity code, the parallel activity code or on the location/mode of transport code; moreover, some rules are necessarily consecutive to other ones.

Distinguishing deterministic corrections from non-automatic corrections, the table 4 shows that the corrected episodes by non-automatic corrections are much more than the deterministic corrections: 13,581 versus 10,408, respectively 15.7% and 12.1% of episodes[3]. The deterministic correction acts on the episode in a sharp way, vice versa the non-automatic correction can be broader than the first one, because it takes into account the entire context of the diary. Therefore, also the average value of corrected episodes per rule is very different: 132 episodes have been corrected deterministically (by using 79 rules), versus 277 episodes corrected non-automatically (by using 49 rules).

Table 4 – Main indicators concerning the code 1 (employment) by type of correction

	Total corrections (deterministic + non-automatic)	Deterministic corrections	Non-automatic corrections
Number of all episodes (A)	86,267		
Number of corrected episodes (B)	20,385	10,408	13,581
% of corrected episodes (B)/(A)	23.6	12.1	15.7
Number of rules (deterministic and non-automatic correction) (C)	128	79	49
Average value of corrected episodes per rule (B)/(C)	159	132	277
Number of all diaries (D)	11,857		
Number of diaries with at least one corrected episode (E)	7,211	5,297	4,856
% of diaries with at least one corrected episode (E)/(D)	60.8	44.6	41.0
Average value of corrected episodes per diary (B)/(D)	1.7	0.9	1.2
Average value of corrected episodes per diary with at least one corrected episode (B)/(E)	2.8	2.0	2.8

Source: *Istat, "Time use survey 2002-2003"*

The number of diaries with at least one corrected episode is 7,211 (60.8% of diaries). The average value of corrected episodes is 1.7 per diary, and it is 2.8 per diary with at least one corrected episode.

Table 5 highlights the results concerning the study activity. The number of corrected episodes is 9,682 (24.6%).

The corrected episodes by non-automatic corrections are much more than the deterministic corrections: 7,334 (18,7%) versus 3,544 (9%) of episodes. The average value of corrected episodes

[3] The sum of the corrected episodes by deterministic corrections and the corrected episodes by non-automatic corrections must not be equal to the number of all corrected episodes, because some episodes are corrected both by deterministic and by non-automatic corrections.

per rule is also very different: 188 episodes have been corrected non-automatically (by using 39 rules), versus 81 episodes corrected deterministically (by using 44 rules).

Table 5 – Main indicators concerning the code 2 (study) by type of correction

	Total corrections (deterministic + non-automatic)	Deterministic corrections	Non-automatic corrections
Number of all episodes (A)	39,319		
Number of corrected episodes (B)	9,682	3,544	7,334
% of corrected episodes (B)/(A)	24.6	9.0	18.7
Number of rules (deterministic and non-automatic correction) (C)	83	44	39
Average value of corrected episodes per rule (B)/(C)	117	81	188
Number of all diaries (D)	7,070		
Number of diaries with at least one corrected episode (E)	3,273	1,964	2,457
% of diaries with at least one corrected episode (E)/(D)	46.3	27.8	34.8
Average value of corrected episodes per diary (B)/(D)	1.4	0.5	1.0
Average value of corrected episodes per diary with at least one corrected episode (B)/(E)	3.0	1.8	3.0

Source: *Istat, "Time use survey 2002-2003"*

The number of diaries with at least one corrected episode is 3,273 (46.3% of diaries). The average value of corrected episodes is 1.4 per diary, and it is 3.0 per diary with at least one corrected episode. The number of corrections is 10,770, equal to 1.11 corrections per episode.

9. *Some final remarks*

Before ending, it is necessary to highlight some remarks in order to underline the main results that have emerged up to this point. They are about the strategies concerning the data editing, which is still in progress.

It is likely that there are a lot of approaches leading to an improvement of the data quality: each of these is surely good, but there isn't any approach better than another. Likely, a real improvement is the consequence of theoretical and technical choices (see Rydenstam, 2000). Therefore, it needs to balance the bearable costs and the benefits which the researcher wants to obtain, considering that the objective is the quality of data.

The first remark concerns the characteristics of the daily diary. It is a particular survey instrument, a very different instrument from the traditional questionnaire, and the only one that is able to collect information on how individuals use their times up to the maximum detail level. So, it is fundamental to improve the process of data editing in order to improve quality.

The second remark is related to the first one and it concerns the importance of recording strings. If the aim was to allow the researcher to plan and to implement the correction rules, in the successive phase of data check and correction, the analysis of recorded strings – often written in the

shape of sentences – has considerably increased the potential of formalizing rules and, therefore, the correction strategies.

The ancillary codes are the subject of the third remark. In order to make the process of coding better, they are another strategic decision. They point out particular coding problems concerning critical events and they let the researcher find solutions in the successive phase. The critical episodes, indicated by ancillary codes, have decreased after the correction process.

Finally, the fourth remark concerns the SAS procedures for making deterministic corrections and the SAS/AF template for making non-automatic corrections. The interaction between the two steps of correction is new and the aim is to make the process of data editing better in order to improve the quality of data.

References

Bolasco S., 1997, *L'analisi informatica dei testi*, in Ricolfi L. (a cura di), "La ricerca qualitativa", La Nuova Italia Scientifica, Roma: 165-206.

Camporese R., Ranaldi R., 2001, *Time use activities: translation from sentences to codes*, paper presented at CLADAG, Palermo, July, 5-6.

Cappadozzi T., Romano M.C., Camporese R., Vitaletti S., 2001, *Assisted coding process: an experiment in progress*, paper presented at IATUR Conference, Oslo, October, 3-5.

Cappadozzi T., Baldazzi B., 2003, *I dati testuali dell'Indagine multiscopo "Uso del tempo" 2002 - 2003: strategie di codifica*, paper presented at Giornata di studio su Applicazioni di analisi testuale, Roma, December, 16.

Eurostat, 2000, *Guidelines on harmonised European Time use survey*, working paper.

Istat, 2002a, *Indagine multiscopo sulle famiglie: Uso del tempo 2002-2003. Istruzioni per la codifica delle informazioni testuali rilevate tramite il diario giornaliero*, working paper.

Istat, 2002b , *Indagine multiscopo sulle famiglie: Uso del tempo 2002-2003. Istruzioni per l'uso della maschera di codifica del diario giornaliero*, working paper.

Romano M.C., 2004a, *Le indagini Multiscopo dell'Istat sull'Uso del Tempo*, in Fraire M., "I bilanci del tempo e le indagini sull'uso del tempo", CISU, Roma: 104-115.

Romano M.C., Cappadozzi T., 2004b, *Il processo di codifica dei dati testuali dell'indagine Multiscopo "Uso del tempo",* proceedings of the 7th Intenational Conference on Textual Data Statistical Analysis (Louvain la Neuve, 10-12 March), UCL, Press Universitarie de Louvain

Romano M.C., Vitaletti S., Camporese R., 2004c, *Improving time use data quality. Recent experiences in Italy*, paper presented at RC 33 Sixth International Conference on Logic and Methodology, Amsterdam, August, 16-20.

Rydenstam K., 2000, *The Eurostat project of harmonising Time use statistics. Proposal on harmonised basic statistics and other actions for promoting international comparisons of Time use statistics*, draft.