UNECE 23 – Topic 1: "Process Automation and Efficiency"

A System-to-System Data Communication Channel for a Multi-technique Data Collection Process: the Case of Italian Agricultural Census

Claudia Fabi, ISTAT, Rome, Italy – claudia.fabi@istat.it[1]
Maura Giacummo, ISTAT, Rome, Italy – magiacum@istat.it[2]

Index

[1] Paragraphs 1., 2., 2.1., 4.
[2] Paragraphs 2.2., 2.3., 3., 4.

1. The reference context: the 7th General Census of Agriculture

Between January and July 2021, the seventh General Census of Agriculture was surveyed, the last one before the transition to the Permanent Census also for the agricultural sector. The only link with tradition, however, was the inclusion in the survey of all the farms who were compliant with the definition harmonized by Eurostat[3]. This census, in fact, had a strong innovative connotation as far as the data collection phase, the survey design, the data collection networks involved and the techniques used to fill in the questionnaires.
Specifically, for the first time in an Italian census survey, the respondent had a wide choice of how to complete the questionnaire, guaranteeing the simultaneous presence of different survey techniques: CAPI, CATI both inbound and outbound, and CAWI.
The census list, i.e. the Reference Universe for the survey, included approximately 1,700,000 units, coming from the use of Administrative Registers, also partially provided by Entities external to Istat. Before starting the survey, the census list was divided into two subsample each pre-assigned to a specific survey technique: CATI or CAPI. However, the pre-assignment was not strictly binding, but preferential, to facilitate the organization of the CAPI and CATI networks, so that they could plan their work on the basis of a predictable workload.
Furthermore, since the start of the survey, all respondents have been able to choose to fill in the questionnaire also through one of the open access techniques:

- CAWI: by self-compiling the questionnaire on a web application developed by Istat;
- Inbound CATI: requesting a telephone interview to the Istat toll-free number or by sending an SMS or a WhatsApp message to a dedicated SIM.

Table 1.1 – Data collection design for the 7th Agricultural Census

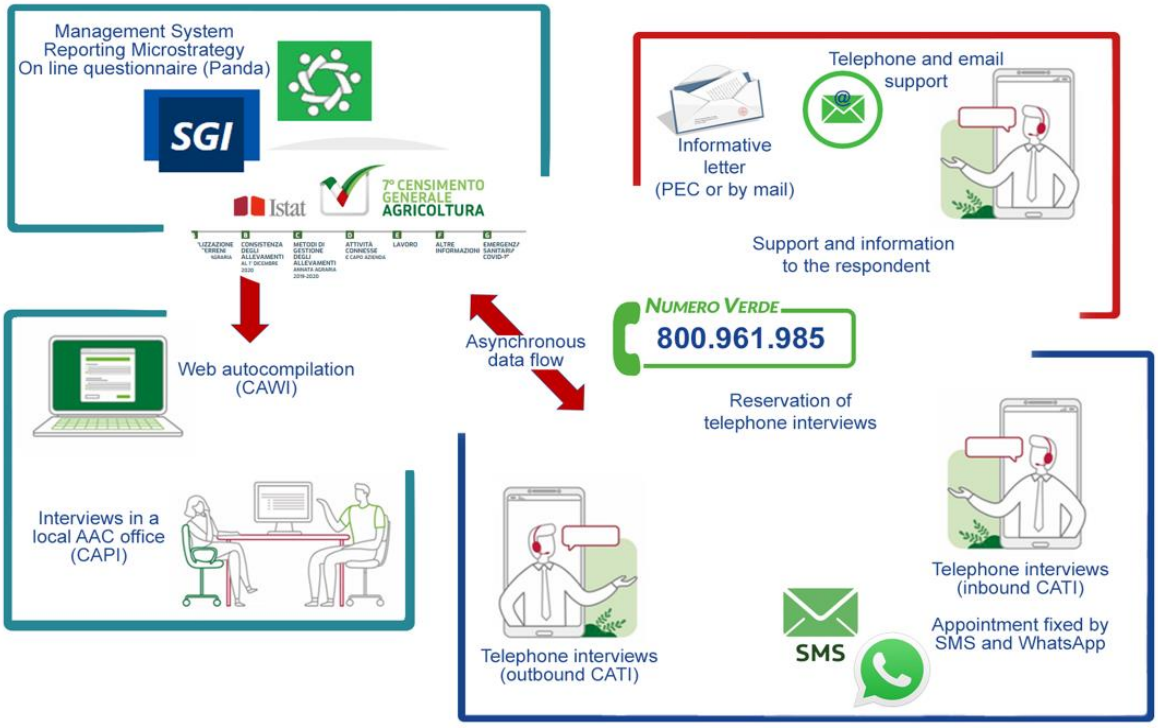
	Schedule of data collection activities	
	From 7th January 2021	*To 30th July 2021*
Techniques available on individual initiative	**CAWI**	
	Inbound CATI	
Pre-assigned Techniques	**CAPI**	
	Outbound CATI	

Therefore, from the first day of the survey and until the end of the data collection, respondents had the opportunity to connect to a dedicated Istat website using their own username and password to fill in the questionnaire or to call the toll-free number and fix an appointment for a telephone interview, at their convenience. In the meantime, however, CAPI and CATI networks have started to work trying to reach those farms that had not already taken steps independently to participate. Interviewers begun to contact respondents, scheduling appointments and proceeding with interviews.
Furthermore, the success of the CATI technique is strongly correlated to the quality (completeness and update) of the telephone numbers in the list. To avoid the increase of non-respondent farms, the design of the data collection expected to transfer from CATI to CAPI the farms who were unreachable by telephone, in particular those who were unavailable at the phone for long time (more than 30 contacts without response) or if the telephone numbers available prove to be incorrect or non-existent. This transition from CATI to CAPI was ongoing throughout the survey, both to avoid dispersing census units for reasons related to the quality of the source list, and to allow the CAPI network to receive new farms in time to try to find them on the territory.

[3] See EU Regulation 2018/1091 (art. 2 paragraph a) for the definition of farm included into census survey.

Figure 1.2 summarizes the subjects and instruments involved in the survey. Top left, the box representing the IT architecture capable of supporting and sustaining survey activities: specifically, a synergy between SGI, Istat's Survey Management System, PANDA, the data acquisition system and Microstrategy, the monitoring and reporting system.

Figure 1.2 – Scheme of the multi-technique design for the 7th General Census of Agriculture



In the lower part of the figure, the subjects involved in the data collection networks, divided between the CAPI network - supported by the Agricultural Assistance Centers on the territory, and the CATI network - centralized in a contact center external to Istat.
The two management systems (SGI with the CAWI and CAPI techniques, and the outsourcer's management and data acquisition system, with the CATI inbound and outbound techniques) were independent. The first aim was the design of a communication module between the two Systems, so that the work of the interviewers could be as synchronized as possible, even if in fact operating on separate platforms.
The synchronization process kept the archives containing the survey results updated for both IT structures, CAWI-CAPI and CATI. This has allowed, for example, to avoid to contact again all the farms that have yet chosen to fill in the questionnaire with the CAWI technique, even if they belong to the subsample assigned to the CATI technique. Similarly, the CAPI assigned farms that chose to call the toll-free number and book an interview with a telephone interviewer were reported to the CAPI network, to prevent them from being further disturbed by face-to-face interviewers.
It is easy to understand that the possibility of effective data exchange between the two software architectures was the key to the entire census operation. In the absence of an effective, timely and functional synchronization between data collection Systems, in a few days the CAPI and CATI subsamples would have been affected by duplications. CAWI and CATI inbound respondents would be subjected to continuous contact attempts even after having filled in their questionnaires with other techniques.
Furthermore, it is precisely through this system-to-system data communication channel that the farms unreachable by telephone were reassigned to the CAPI technique. In this way, the data collection process gained a further possibility of optimizing the use of survey techniques, attempting where

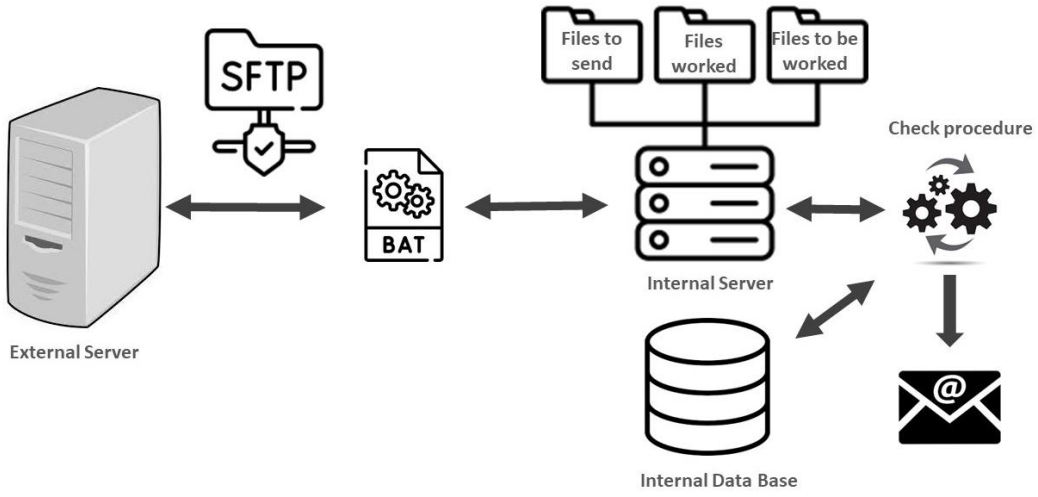
possible the telephone interview first, certainly faster and less expensive, and, secondly, switching to the CAPI technique, only when the intervention of face-to-face interviewers was really necessary.

2. Design of the System-to-System data communication channel

In the start-up phase, a great attention was dedicated to the design, the construction and the testing of a bidirectional communication flow between the two Systems. The communication, based on ASCII files with shared encodings, ran twice a day to complete the synchronization operations, in pre-set time and without any need to interrupt data collection operations during the update.

An element of complexity was represented by the need to show the results of the CATI survey in the web application, SGI, acquiring a series of basic information related to the daily telephone contacts made by the CATI interviewers. At the same time, the CATI System had to be constantly aligned with the data coming from the activities carried out by the CAPI network and from the online compilations carried out directly by the respondents.

Graph 2.1 – System to System data communication scheme adopted for the survey



The design process of an automated data exchange system, through files in a predefined format, coming from SGI to the CATI system and vice versa, lasted about 6 months before the start of the survey.
The design specifically concerned the following issues:

a) scheduling of data exchange frequency;
b) exchange file format;
c) content of the exchange files;
d) nomenclature of exchange files;
e) quality control strategies for exchange files and identification of anomalous records;
f) recovery plan, in case of failure of the exchange procedures.

In order to avoid a system slowdown, the twice a day data exchange was scheduled in hours in which it was predictable to have less data traffic on the Systems, considering that the daily work of the interviewers would probably have led to a daily recording on the order of tens of thousands records. The data exchange hours have been set as follows:
- 1st synchronization: 06:00-08:00 a.m.
- 2nd synchronization: 01.00-03.00 p.m.

2.1. The structure of the interchange files

In designing the structure and content of the interchange files, it was intended to pursue an objective of simplicity, completeness and non-redundancy, including only the information indispensable for the purposes of synchronizing the management Systems. This in order to keep the time required for automatic data transmission between the two systems to a minimum, reducing the size of scheduled sending.
The following table shows the list and characteristics of the transmitted variables.

Table 2.2 – The structure of the interchange files

Variable name	Description	Notes
PROGR_REC	Progressive number of the record	Identification code of the record in the current file
COD_IDENTIFICATIVO	Identification code for the farm	Identification code of the farm included in the census list: the identification code was assigned before the start of the survey and didn't admit duplications
FLAG_CATI	CATI pre-assignment flag	Allow to recognize if the farm was pre-assigned to CATI technique
STATO	Questionnaire status in Istat SGI	
ESITO_CHIU	Definitive outcome	Allow to recognize if the outcome transmitted was or not definitive, meaning that the farm should not be contacted anymore ("0" by default, "1" means "no more contact")
ESITO_OUT	CATI outbound outcome in detail	Outcome code in detail: this code was intended to be used to update the CATI outbound database or viceversa
ESITO_IN	CATI inbound outcome in detail	Outcome code in detail: this code was intended to be used to update the CATI inbound database or viceversa
DESC_ESITO	Outcome description	Textual description of the outcome code
DATA	Date (day and time)	Day and time in which the outcome has been recorded
TECNICA	Outcome technique	A code that identifies the data collection technique in which the outcome has been recorded

The coding of each variable allowed to uniquely link the information of a contact with a farm with its own "history" of contacts, integrating it chronologically and recording for each attempt also the data collection technique used.
The need to make information on contact attempts and the respective outcomes mutually interchangeable has also led to the design of outcome table code by technique that are as consistent as possible (meaning that a coding scheme meaning "same code" = "same outcome" in each technique). This allowed to simplify the interpretation of the file content and drastically limited the need for additional data transmission recoding post processes.
Particular attention was also dedicated to the non-trivial aspects of nomenclature, which would allow the automatic interpretation of the file content by the acquisition and synchronization batches.
Specifically, the name of each file included:
- two initial letters identifying the direction of the exchange, specifically the letter "C" to mean the outbound and inbound CATI technique (therefore the outcomes recorded by the outsourcer) and the letter "S" to mean the Istat management system, for the CAWI and CAPI techniques. Therefore a file with a name starting with "CS" is a file that contains CATI outcomes intended for updating SGI, while a file with a name starting with "SC" is a file that contains CAWI/CAPI outcomes intended to update the CATI system;

- an explicit reference to the date, in the format "*yyyymmdd*", corresponding to the day on which the outcomes contained in the file occurred;
- a letter identifying the file produced and transmitted during the night "M" and the file produced in the early afternoon "P".

This nomenclature always produces not duplicated names, so that it was never possible to overwrite them. It was also easy to identify and recover any tranches of processing not performed automatically by the Systems due to technical problems.

Even the transmission of files has been automated through sFTP batch, capable of copying the files produced in predetermined destination folders, towards which the automated synchronization procedures of the respective systems have pointed.

2.2. Data quality control

The policies implemented for quality data control provided by the external outsourcer were based on three levels that intervene in different parts of the process. In details:

- 1st level of check: This level evaluated the file received from the external outsourcer as a whole and determining the acceptance or the rejection of the entire delivery;
- 2nd level of check: This level analysed every single record contained in each received file and, in case of failure, excludes non-compliant records and accepts the compliant ones;
- 3rd level of check: This level prevents the generation of the return synchronization file when the received file did not successfully pass the 1st level checks.

In general, the strategy chosen was to minimize discrepancies, reducing them to critical situations only, without compromising the quality of the transmitted information. The following are the controls performed at each of the three levels described above.

The first level of check performed the following controls:

- Correctness of the file name according to the predetermined and expected nomenclature;
- Correctness of the file structure according to the predefined and expected record layout.

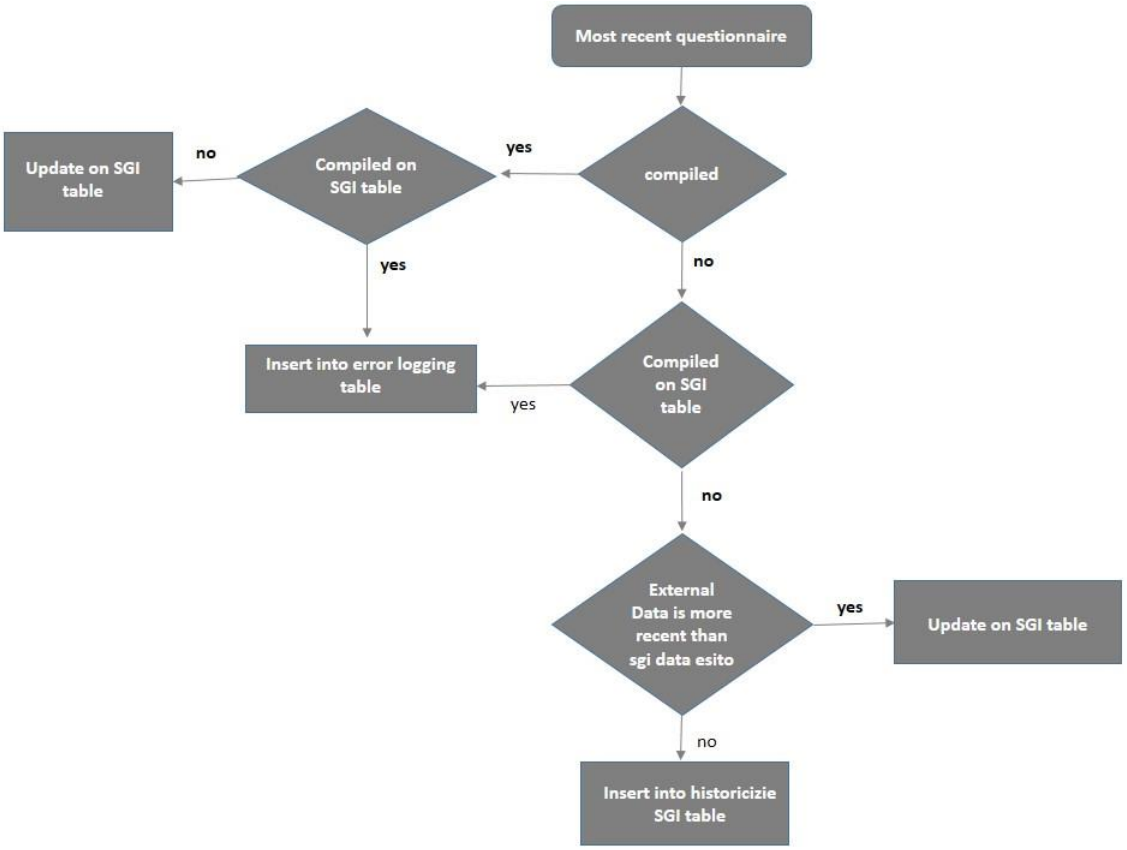
For the control of each record, it was possible to identify two additional categories of problems: one attributed to an overlap of response techniques and the other resulting from mapping errors or information coherence. The first case is a discrepancy not related to any error but simply linked to the possibility that the respondent could complete the questionnaire using more than one technique, due to the lack of perfect synchronization between Systems. The second type derives from errors attributable to incorrect content in the file received by the external outsourcer. In detail, the second level included the following controls:

- Presence of a farm for which the questionnaire was already completed using a different technique;
- Correctness of the exchange identification code;
- Coherence of the data sent by the external outsourcer;
- Consistency between the sent outcome and the compilation technique declared by the external outsourcer (inbound or outbound);
- Compatibility of the record's processing date with the data submission time window.

Finally, the third level of check managed the decision for the generation of the return synchronization file. The creation of the file, intended to update the CATI system, stopped if the first level blocked the input file. Blocking the generation of the file was necessary, if the received file was rejected during first level of check, to avoid synchronizations that did not take into account the received information (and thus the outcomes). In fact, the return file would have been formally correct but lacking in terms of completeness and level of updating since it could not consider the outcomes of the attempts recorded by the CATI technique on the previous day.

The external outsourcer sent all the outcomes of all contact attempts made on the farms to allow for archiving and monitoring all the work performed. It was a further issue because it was plausible to have multiple records corresponding to each telephone attempt made by the CATI network on the previous day for the same farm. In this case, the procedure updated the data by choosing to synchronize the databases with the information from the most recent attempt and archived the other contact attempts in the contact history. The tables updated were the same ones used by SGI. Since these operations were carried out while keeping the application online and without interrupting the work of the survey networks, it was necessary to manage data concurrency. The figure below shows the expected workflow with the data and tables.

Graph 2.3 – Data Workflow



2.3. Failure recovery plan

The adoption of a system-to-system data communication tool, based on a synchronization protocol, required a recovery plan that guaranteed a quickly restore of data in case of errors.

The first action taken was activing a critical failure events alerts (such as empty or incorrect deliveries) using an automatic system that sent emails to a control team composed of Istat personnel and external outsourcer personnel. This allowed having a real-time notification to operational staff about the issues encountered without the need to access logging systems or monitoring applications.

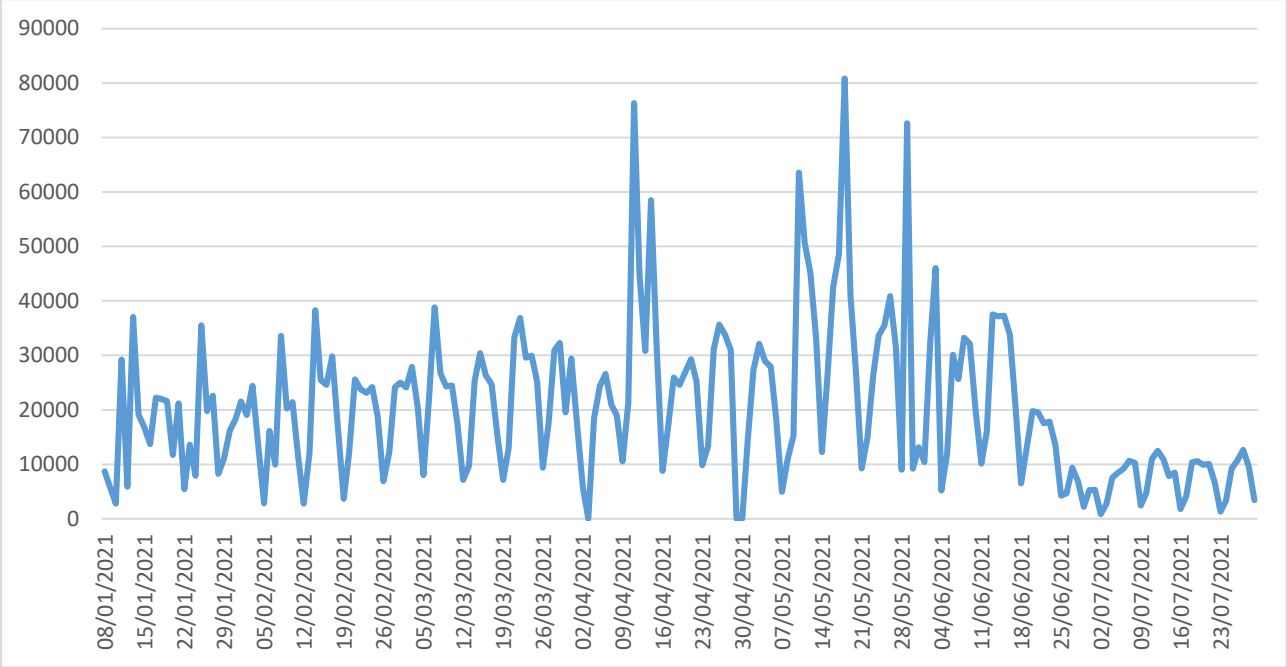
Furthermore, all control procedures were designed to repeat the processing of any incorrect deliveries, both entirety or partially.

If it was necessary to recover the acquisition of more than one failed delivery simultaneously, it was planned to process them sequentially, starting from the oldest and moving on to the most recent one, updating the database with the latest information and archiving all the others. All files sent by the external outsourcer, once processed, were stored in a dedicated file system.

3. Results

The two main purposes of the communication and synchronization architecture were: the need to update the outcomes of the agricultural companies' questionnaire, with all the attempts done, and the possibility to reassign to the CAPI technique those farms that couldn't be reached through CATI technique due to issues with the quality of telephone contacts in the census list.
Let's start by reporting the amount of information exchanged daily. This data can be useful to evaluate the size of the architecture and the data space dedicated if a similar procedure would be re-used for other survey.
During the survey, 412 files were exchanged, regarding over 6 million records. The average number of records processed daily was around 10,000, with peaks of 80,000 records during particularly intense fieldwork periods. The graph below shows the daily trend of the quantity of records exchanged, in sending and receiving files.

Graph 3.1 – Number of record processed daily by the System



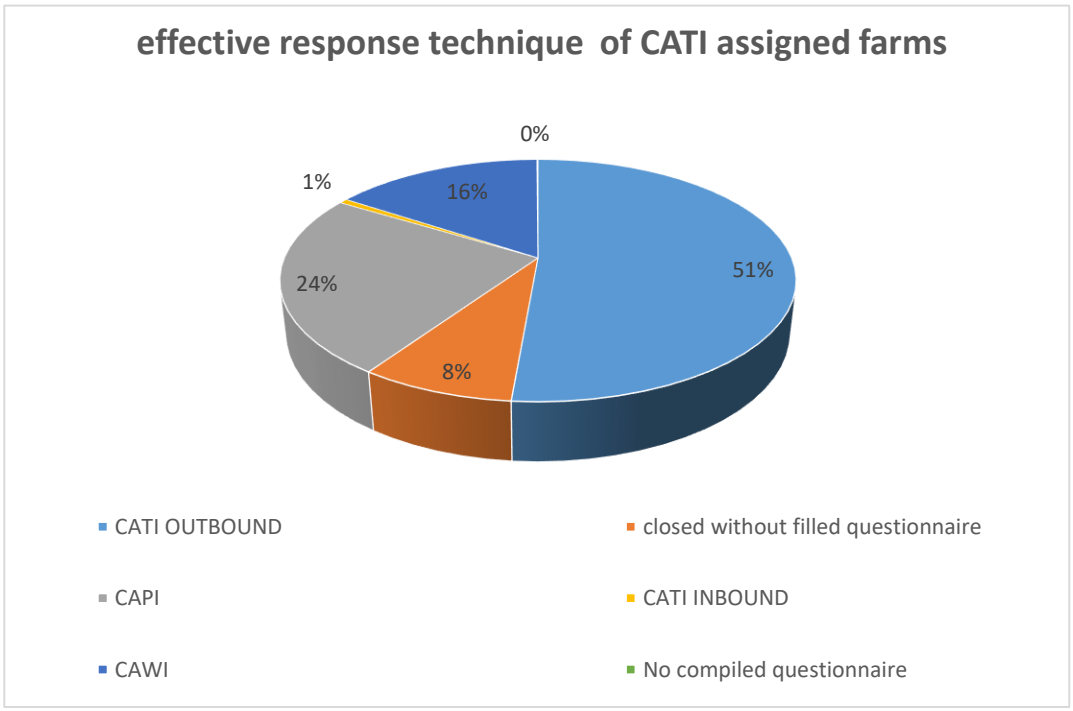
Regarding of data quality, the first level check rejected only three files, recovered by the outsourcer with new files very quickly.
The second-level check traced during the processing phase amounted of a total of 45,351. The record attributed to the change of technique were 33,065. These, corresponding at the 73%, indicated that the discrepancies were not due to processing errors but rather to questionnaires completed using other techniques. The percentage of true errors is 27% of the total number of records rejected, due to two specific data coherence issues:
- the operation date was not compatible with the date entered in the file name;
- the outcomes were not compatible with the technique, depending on an inconsistency between the transmitted outcome code and the data collection technique (e.g., an outbound outcome code for an inbound attempt, etc.).

This result is encouraging because it demonstrates that the real-time synchronization issues between techniques did not pose a significant problem. Furthermore, the errors derived from coherence checks

rather than non-existent data in the outcomes or identification codes. This allowed for quick recovery of the rejected records.
The achievement of this result was the outcome of an intensive startup effort aimed at making the external outsourcer as independent as possible in managing its own outcomes and configurations.

Regarding the synchronization efforts that facilitated the transition of agricultural companies assigned to CATI technique but not reachable by phone, to the CAPI technique, the following data give a comprehensive overview of the role of CATI technique in the Agricultural Census.
The initial sample consisted of 1,699,942 farms, the units selected for the CATI technique based on the previously mentioned criteria and delivered to the external outsourcer before the start of the survey amounted to 550,000 units, representing 32% of the census list.
At the end of the survey, interviews completed using the CATI technique were 282,536, slightly over 50% of the assigned units. The remaining units moved towards other techniques based on conscious choices by respondents, who opted to complete the questionnaire via self-administered CAWI or by referring to the CAPI network.
The number of farms with incorrect or non-existent phone contacts was particularly high, with 116,890 companies, approximately 21% of those assigned to the CATI technique. This represents a significant amount that strongly affected the effectiveness of the CATI technique throughout the survey period. However, thanks to the possibility of synchronization between management systems, concurrent multi-technique utilization was made possible, optimizing the almost "real-time" use of techniques and, most importantly, avoiding abandoning units with incorrect phone contacts even if initially assigned to the CATI technique. This ensured the opportunity to mitigate the impact of the poor quality of phone contacts in the census list, allowing agricultural companies to participate in the survey using other techniques immediately and in a completely transparent manner for the respondents.
The graph below illustrates the distribution of units initially assigned to the CATI technique: as shown, in 16% of cases, respondents opted for self-administered CAWI, while another 24% were captured through the CAPI network.

Graph 3.2 –effective response technique of CATI assigned farms



4. Conclusions

Despite the limitation of representing an approximation of real-time synchronization, the asynchronous update, scheduled to occur automatically at predetermined times without the need to stop data collection operations during the update, has ensured a satisfactory smoothness in the data collection process both of CAPI and CATI networks, while offering respondents wide discretion to use autonomous or assisted compilation tools.

In general, both the possibility of establishing computerized dialogue between different Systems and the ability to acquire data from any physical location where the interviewer is located represent forms of adaptive evolution of survey instruments. They are increasingly necessary as data collection activities must meet the respondents' needs, their preferences for one communication channel with Istat over another, their availability of time, and their geographical distribution. These adaptations are essential for successfully maintaining contact with the respondents and obtaining their indispensable cooperation.

The integration, although it did not allow for the real-time import of CATI contact outcomes or the immediate export of CAPI contact attempts or self-completion accesses, constituted an unprecedented innovation for the multi-technique surveys carried out by Istat. These surveys are typically designed to allow either sequential or concurrent multi-technique approaches but on predefined and non-permeable subsets of the population.

Unfortunately, the "near real-time" synchronization, while representing an important technological and organizational innovation for census surveys at Istat, is still an approximation of what would constitute the optimal approach for synchronous multi-technique surveys. The optimal approach would involve centralizing technical and operational management in a single computerized instrument developed by Istat.

Over time, the continuous implementation of modules and functional structures for managing Istat surveys will likely lead to the availability of a fully integrated Management System. This System will be available not only to Istat users but also to outsourcers who will be required to operate on it in perfect synchronization with other techniques and data collection networks.

For the 7th General Agricultural Census, a single System architecture was not yet available. This certainly resulted in a greater deployment of human and technological resources to compensate for the lack of complete synchronization between the various systems. However, it marked the beginning of a direction towards the surveys of the future.