Meeting of the Group of Experts on Business Registers

Organised by UNECE

15 – 18 October 2024

online meeting

*F. Alonzi, M. Consalvi, C. Viviano*
*the National Institute of Statistics (Istat)*

# *A Comprehensive Strategy for Implementing NACE Rev. 2.1 in the Italian Statistical Business Register*

*Keywords: NACE, economic activity classification, recoding strategy, Statistical Business Register.*

## Abstract

*Implementing a new classification of economic activities in the Statistical Business Register (SBR) is a complex task that, while offering a more accurate and up-to-date depiction of economic structures, can disrupt existing time series, necessitating either back-casting or double coding to mitigate this issue. In the context of SBRs, back-casting is generally impractical, making double coding a primary concern, to support all statistical domains to avoid inconsistencies within the system.*

*Given their central role, statistical business registers are among the first areas to implement the new classification, thereby facilitating its adoption in other business and trade statistics domains. With the NACE Rev. 2.1 classification now largely finalised, primary efforts for its implementation in the SBR domain include: 1) the release of the ATECO national version by Istat, under the responsibility of the same office as the SBR, and 2) the recoding of SBR units and management of the double coding period. Istat is responsible for developing an implementation plan that addresses user needs and the statistical production requirements of the relevant domains.*

*This document outlines an implementation plan for the adoption of the new ATECO version starting on January 1, 2025. The plan involves coordination with administrative bodies responsible for business registration and certification, such as chambers of commerce and tax authorities. Following an overview of the organizational framework established under the governance of an ATECO Committee – coordinated by Istat and comprising national classification stakeholders – the paper presents the overall strategy and various methods for recoding units.*

*The initial step involves setting up the new classification, including its structure, titles, explanatory notes, and correspondence tables. The development of the new national classification is an incremental process, with successive versions produced and shared internally with stakeholders until the final version is approved by Eurostat and legally adopted. During this refinement phase, the complexity of recoding is evaluated to determine the optimal combination of methods, sources, and expertise.*

*Several tools have been developed to facilitate the coding strategy, primarily leveraging automated techniques: text analysis applied to classification and administrative sources, conversion tables mapping the old and new classifications, and a survey to directly recode and estimate probabilistic conversion matrices. In this context, the most challenging innovation is the construction of an ATECO index.*

*The development of these methods involved multidisciplinary teams, including methodology experts, classification specialists, IT professionals, and communication support staff.*

## 1. Introduction

On 1 January 2025, the national Ateco2025 classification (directly derived from NACE Rev. 2.1) will be officially adopted.

The SBR is the first statistical domain to have to implement the new Ateco2025, maintaining the dual coding for a limited period of years since its role of backbone for the other statistical business domains.

Economic activity is a relevant variable of the SBR, mainly used for sample stratification and to describe the economic structure of an economic system. Even though the new NACE Rev. 2.1 was defined not as a total revision but as an update of the previous classification, changes introduced have a non-trivial impact on numerous activities for the adaptation of the business register. Data structures of the historical SBR database (DbAsia) have to be adapt to record the new classification; two coding systems have to be kept aligned; the interactive DbAsia update tools have to be redesigned; the methods and procedures for estimating the economic activity code, currently based on the massive use of administrative data, have to be revised as well.

The peculiarity of the Italian ATECO classification makes its implementation more complex for two reasons. From a structural point of view, the ATECO classification aims to better describe the national reality; in fact it is a hierarchical classification made up of six levels, of which the first four are fully consistent with the NACE defined at European level; moreover, it is used at national level for statistical and other purposes. In particular, in the SBR managed by Istat, the fifth level is used while the sixth level is not used for statistical purposes, but is limited to administrative functions.

The Business Register Section, within the Directorate for Economic Statistics of Istat is also responsible for the release of the ATECO classification and its updates; it represents the official Italian version of the NACE Rev. 2.1 used by both the statistical system and administrative bodies. Due to this dual function, the process of defining the classification and then its implementation is based on a new organizational model. Firstly, the set-up of the ATECO Committee under the Istat governance was formed by the main national stakeholders of the classification, and was later expanded by adding a stable network of ATECO national stakeholders.

It was also necessary to set up an *ad hoc* organisational model for the implementation stage. Since the burden of the implementation activity falls more or less on the same resources belonging to the SBR office and simultaneously dedicated to the definition of the classification, other cross-sectional Istat structures have been involved in the implementation phases. Methodologists, computer scientists, IT, data collection experts participate in the implementation activities thanks to the establishment of dedicated TFs. The role of experts in classification and register maintenance remains central, as they must guide and test the development of new methods of automatic coding, validate operational conversion tables, and address the development of new tools for the introduction of innovations into the classification support with manual reclassification of large units.

The choice of applying different recoding methods will depend upon the amount of changes we have to face in the SBR: complex cases, i.e., those that cannot be automatically recoded, the number of units involved, their size, the availability of administrative sources, etc., will steer in the approach of recoding activities and the use of the different tools and methods.

In the initial section (§2), we will provide a concise overview of the revision process, then looking deeper into the preparation of the national version of the classification, given its peculiarities and dual function as both a statistical classification and an administrative tool (§3). Afterwards we will outline the implementation plan, covering methodological and thematic aspects. This includes an assessment of the impact of the changes and the expected outputs for SBR users (§4.1). Additionally, we will explore the distinctive features and potential of various tools for recoding, such as profiling activities (§4.2.1), a special ad hoc survey (§4.2.2), the utilization of auxiliary administrative sources (§4.2.3), the mapping and creation of the operational correspondence table (§4.2.4), and the innovative automatic classification tools that employ machine learning techniques for textual analysis (§4.2.5). Finally, we will summarize the overall integrated approach to the use of the different tools and methods analysed and presented in the previous sections (§5).

## 2. Organisation and governance of the revision process: development and implementation of Ateco2025

How was it possible to manage the complex task of implementing the ATECO classification into the Business Register? To streamline this process a comprehensive governance structure was implemented. While the Business Register Section took the lead in delivering the final product, since it was ultimately responsible for the final output, it became evident that a collaborative effort involving multiple organisations and teams was indispensable. Thus, a new multi-tiered governance system was designed to harness the diverse skills and perspectives required to successfully complete this project, with the SBR Section as the central coordinating body and parallel structures (committees, task forces and other organisms) to tackle specific aspects of the project, each one operating at the same level, contributing to the overarching goal of integrating the ATECO code into the register.

This multifaceted approach allowed for a more specialised focus on each aspect of the project and was essential to address the intricate nature of the project and leverage the specialised expertise available within and outside Istat. A visual representation of the governance framework is provided below.

*Table1 – The governance framework for the creation of Ateco2025 classification and its implementation in the SBR* [1][2]

**The ATECO Committee and the network of stable users**

- Definition of the new version of the ATECO classification.
- Collection of users' need and contribution to the development of the NACE/CPA classifications.
- Coordinated by ISTAT and formed by national stakeholders of the classification including administrative bodies and statistical domains.

**Project SMP-ESS-2022-NACE-H1-5994-IBA Preparing the Implementation of NACE Rev. 2.1 Classification**

- Preparing the implementation plan of NACE Rev. 2.1 in the Business Register.
- Defining the Guidelines for the translation of the explanatory notes into the national language.
- Running an *ad hoc* special survey supporting the new classification implementation.

**Task Force to support the Survey of Economic Activities (SEA) for the Implementation of Ateco2025**

- Responsible for the entire data production process, from data collection to dissemination.
- Design and management of technical support activities for respondents, monitoring of data collection and non-response recovery strategies.
- Design of the sampling strategy, developing ML methods for the use of non-sampling and sampling data for the recoding of SBR units; sampling estimation and quality assessment of results.
- Questionnaire design, optimisation, pre-testing and functional testing; IT development and Business Portal.

**Task Force to re engeneering the official automated coding system CIRCE**

- Feasibility study to decide if and how updating the CIRCE classification tool.
- Design and development of a two-faced environment: deterministic vs AI approach.
- Formed by IT specialists, methodologists and classification experts.

**Improvement of IT support infrastructure (BR, Business Portal and Ateco management tool)**

- Re engeneering of the SBR informative system to support the double coding and the updating of the new metadata.
- New functionalities in the BP to collect feedback from enterprises on economic activities.
- Development of a new database to store the classification (SISMA).

**Cross-media communication campaign and involvement of the stakeholders**

- DEM (Direct Email Marketing) package to support the *ad hoc* survey and to facilitate the implementation of the new classification.

[1] **SISMA** is an Italian acronym standing for *Sistema Informativo di Supporto per la Manutezione della classificazione ATECO*, which translates to *Information System to Support the Maintenance of the ATECO classification*. Ateco2025 is housed within this new database, developed as part of Istat's 2020 project for an innovative information system. It is designed to provide users with a dynamic and electronic ATECO classification. The primary users of SISMA will be those involved in maintaining, updating, and revising the classification. SISMA is the result of a collaborative effort between Istat's classification experts and IT specialists. It is designed to support regular updates of ATECO due to national requests and future major revisions of its parent classifications (like NACE). Within the SISMA database, each explanatory note in Ateco2025 is accompanied by metadata, including the source of the information and the type of explanatory note. Correspondence tables are also stored in SISMA.

[2] The **Business Portal (BP)**, known as "*Statistica&Imprese*", is currently the system of statistical services dedicated both to businesses and to Istat statisticians for the collection and return of information of the main surveys of official statistics. This

## 3. Preparing a national version of the classification

### 3.1 Peculiarities of the Italian ATECO classification

From a structural point of view, the NACE includes four levels (sections, divisions, groups and classes). It is the subject of legislation at the European Union level, which imposes the use of the classification uniformly within all the Member States; based on the NACE regulation, Member States' statistics presented according to economic activities are to be produced using the NACE or a national classification derived therefrom.

The national classification may introduce additional headings and levels and a different coding may be used. Each of the levels, except for the highest, shall consist of either the same headings as the corresponding NACE level or headings constituting an exact breakdown thereof.

Italy, as several Countries, such as Austria and France have national versions of the NACE. Also Switzerland has adopted a national version. The reason lays on the evidence that the NACE does not completely satisfy national needs and requires an adaption to meet Countries specific conditions.

Sometimes, national versions of the NACE, developed primarily for the production and presentation of statistics, are utilised also for other purposes e.g. legal purposes thus further breakdowns are necessary to identify economic activities in a more delimited way.

In general, but not always, national versions of the NACE are more detailed than their European version and include at least one more level.

The Italian version of the NACE, namely ATECO is fully consistent with the four-digit structure defined at European level, but it has two more digits. More specifically, while the first four levels are inherited from the NACE, two more levels have been added at national level; they are called “categories” and “subcategories” and they constitute an exact breakdown of the NACE classes. In most cases, ATECO categories and subcategories consist of a conversion of the notes of inclusion described in the NACE into national headings.

Since the release of the NACE Rev. 2 currently in use, the corresponding ATECO classification in force has been used at national level both for statistical and non-statistical purposes. It is worth noting that the sixth level is not used for statistical purposes, but it is limited to administrative functions. More specifically, in the Statistical Business Registers maintained by Istat economic units are classified according to the ATECO categories and not the subcategories. On the contrary, in administrative sources (such as fiscal sources or the register of enterprises managed by the Chamber of Commerce) enterprises are registered at the highest level of detail according to the ATECO subcategories. ATECO is also widely used by the Government and local agencies, trade associations and other organisations.

The fact that the same classification is used for different purposes, other than statistical, implies the need to change more frequently the classification in order to satisfy policy requirements. Such need has increased starting from March 2020 when the Italian Government had mandated the first shutdown of non-essential businesses and defined the list of all economic activities that would have been interrupted by policy mandate according to the ATECO classification; such a catalogue, and the following ones, were included in national regulations. The same *modus operandi* was applied later in the months when the Government started to adopt a set of economic measures to support enterprises active in those industries that were affected the most by the pandemic.

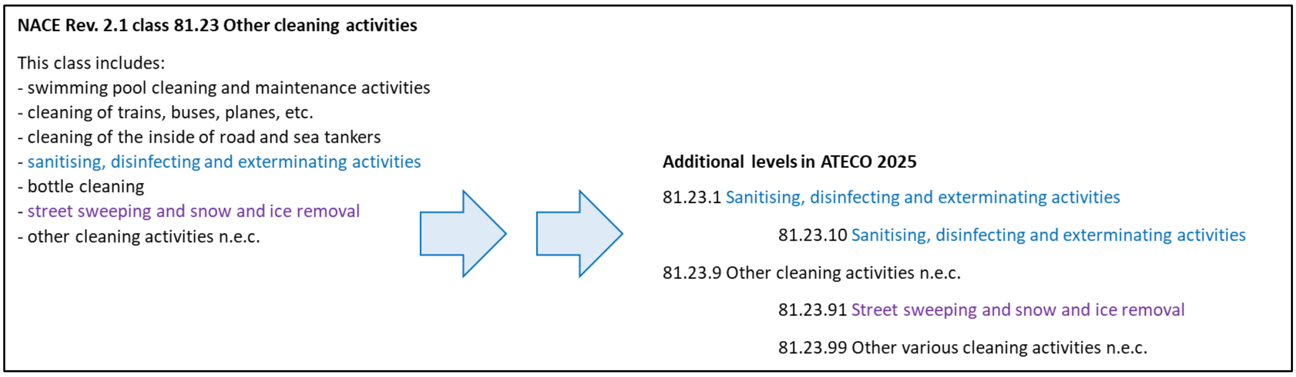
single point of access allows bi-directional communication between Istat and the business world: using this tool all enterprises involved in business surveys can quickly and easily fulfil their information obligations receiving back a number of significant advantages, like customized statistical information feedback. Most of the updates of the SBR due to statistical sources – Structural Business Statistics (SBS) and Short-Term Statistics (STS) – are mainly obtained via BP, simplifying not only the procedures by which businesses provide statistical information, but also those collected by SBR experts.

During the latest years Istat has collected several proposals for changes, both in the form of changes to existing codes (e.g., greater detail or merging) and in the form of assigning new codes to specific economic activities currently included in residual classes and thus treated as “not elsewhere classified”.

All these inputs have been taken into account in the initial phase of development of the structure of the national version of the new NACE Rev. 2.1 that is Ateco2025. As its predecessor, Ateco2025 is a hierarchical classification made up of six levels. Several efforts have been made to ensure the consistency of the Ateco2025 classification with the four-digit structure of the NACE Rev. 2.1 defined at European level especially in order to correct a few mistakes introduced during the last revision process in 2007.

In order to avoid introducing misalignments with the NACE Rev. 2.1, when possible the titles of Ateco2025 categories and subcategories consist of a conversion of the notes of inclusion described in the NACE into national headings as provided in the example presented below.

*Picture 1 – NACE Rev. 2.1 vs Ateco2025: some examples*



**NACE Rev. 2.1 class 81.23 Other cleaning activities**

This class includes:
- swimming pool cleaning and maintenance activities
- cleaning of trains, buses, planes, etc.
- cleaning of the inside of road and sea tankers
- sanitising, disinfecting and exterminating activities
- bottle cleaning
- street sweeping and snow and ice removal
- other cleaning activities n.e.c.

**Additional levels in ATECO 2025**

81.23.1 Sanitising, disinfecting and exterminating activities

81.23.10 Sanitising, disinfecting and exterminating activities

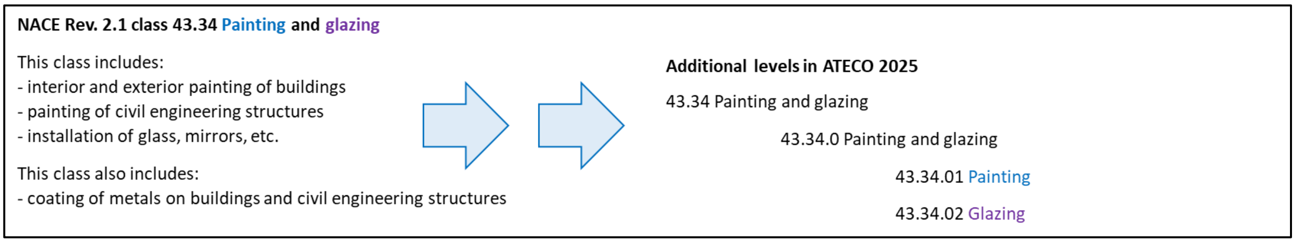
81.23.9 Other cleaning activities n.e.c.

81.23.91 Street sweeping and snow and ice removal

81.23.99 Other various cleaning activities n.e.c.

In other cases, they are derived by splitting the titles of NACE Rev. 2.1 classes as the example below shows.

*Picture 2 – NACE Rev. 2.1 vs Ateco2025: some examples*



**NACE Rev. 2.1 class 43.34 Painting and glazing**

This class includes:
- interior and exterior painting of buildings
- painting of civil engineering structures
- installation of glass, mirrors, etc.

This class also includes:
- coating of metals on buildings and civil engineering structures

**Additional levels in ATECO 2025**

43.34 Painting and glazing

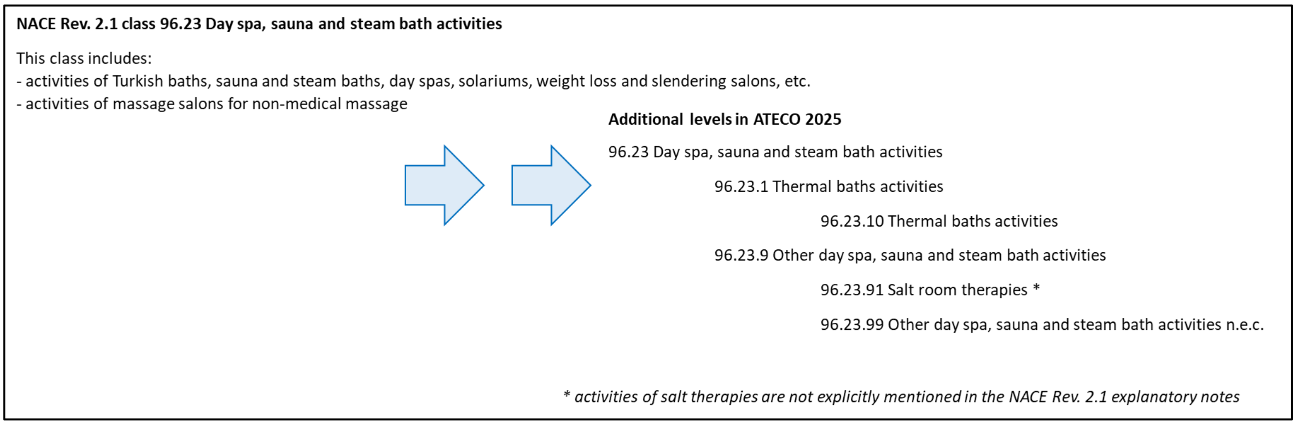
43.34.0 Painting and glazing

43.34.01 Painting

43.34.02 Glazing

In a few residual cases, there are titles concerning economic activities which are not explicitly mentioned in the NACE Rev. 2.1 explanatory notes (see example below) but implicitly included there according to the ATECO classification team’s knowledge.

*Picture 3 – NACE Rev. 2.1 vs Ateco2025: some examples*



**NACE Rev. 2.1 class 96.23 Day spa, sauna and steam bath activities**

This class includes:
- activities of Turkish baths, sauna and steam baths, day spas, solariums, weight loss and slendering salons, etc.
- activities of massage salons for non-medical massage

**Additional levels in ATECO 2025**

96.23 Day spa, sauna and steam bath activities

96.23.1 Thermal baths activities

96.23.10 Thermal baths activities

96.23.9 Other day spa, sauna and steam bath activities

96.23.91 Salt room therapies *

96.23.99 Other day spa, sauna and steam bath activities n.e.c.

** activities of salt therapies are not explicitly mentioned in the NACE Rev. 2.1 explanatory notes*

### 3.2 Setting up the new classification

### 3.2.1 Roadmap to define the final version

The new structure of the Ateco2025 classification is the result of a process officially started in 2020 when a new organisational model was set up under the governance of the already mentioned ATECO Committee coordinated by Istat and formed by the main national stakeholders of the classification.

In the latest years, a wide network of national stakeholders of the ATECO classification has been set up to involve also minor users in the process.

Istat has adopted a transparent, collaborative and inclusive approach to the whole revision process by applying the basic principles of statistical classifications, e.g. conducting regular hearings and offering opportunities for discussions with major users of the ATECO classification, determining users' requirements, balancing users' needs.

From a general perspective, the following actions have been undertaken to set up the new classification:

- a dedicated e-mail address to collect proposals of change from national users;
- launch of a public consultation to collect proposals of change;
- organisation of about 150 virtual meetings to define the new ATECO structure and contents with single users (bilateral meetings) or within the ATECO Committee (more than 100 members) or the stable network of ATECO national stakeholders (more than 800 members);
- sending official informative notes to the various Italian Ministries to collect their requirements;
- launch of several written consultations addressed to national stakeholders on specific issues to acquire suggestions and positions by specific groups of stakeholders.

Istat has collected more than 600 proposals for change.

All the proposals have been analysed one by one by classification experts and accepted if considered valid from a methodological point of view. When possible, the relevance and impact of the different proposals of change have been analysed, too.

The proposals were of different types: from requests to introduce new explanatory notes or suggestions to improve translations from the NACE Rev. 2.1 classification but also instances for the restructuring of entire branches of the ATECO classification.

In the process of defining the new classification, several versions have been produced; in this incremental process, the successive versions incorporate corrections and updates, according to a shared approach with the entire network of stakeholders.

The final version of the Ateco2025 classification consists of:

- a national structure consistent with the structure of the new NACE Rev. 2.1 classification;
- a rich list of explanatory notes consistent with the contents of the NACE Rev. 2.1 classification;
- a bilateral correspondence table between Ateco2025 and Ateco2022 currently in use;
- informative documents presenting the ratio of the new classification and its methodology.

Thus, the new version of the ATECO classification is the result of the revision process of the NACE Rev. 2.1 and of several consultations with national stakeholders, as described in the following roadmap towards the final version of the Ateco2025 classification.

*Table2 – Roadmap to the final version of Ateco2025 classification*

| When | What and who |
|---|---|
| 17 Oct 2022 | Launching of an **early written consultation** to collect information on all needs for change expressed by the different stakeholders. |
| 16 Dec 2022 | Deadline of the written consultation. |
| 29 Dec 2022 | Release of a very first draft of the **structure** of the Ateco2025 classification. The document has been shared within the ATECO Committee and some other statistical domains within Istat. |
| 12 Sep 2023 | Launching of a **public written consultation** to collect proposals for change and users' needs. |
| 31 Oct 2023 | Deadline of the public written consultation. |
| 16 Nov 2023 | Release of a very first draft of the Ateco2025 classification: structure and explanatory notes (**Ateco2025 V0 version**). The document has been shared with the main administrative bodies (Chambers of Commerce and Tax Office) and with some other statistical domains within Istat. |
| 29 Dec 2023 | Release of the first official draft of the Ateco2025 classification: structure and explanatory notes (**Ateco2025 V1 version**) aligned to the NACE Rev. 2.1 version 1.01. The document has been shared within the ATECO Committee, the network of stable users of the ATECO classification and all statistical domains within Istat. |
| 31 Jan 2024 | Deadline for comments, suggestions and changes to Ateco2025 V1 version. |
| 23 Feb 2024 | Release of the second official draft of the Ateco2025 classification: structure and explanatory notes (**Ateco2025 V2 version**) aligned to the NACE Rev. 2.1 version 1.01. The document has been shared within the ATECO Committee, the network of stable users of the ATECO classification and all statistical domains within Istat. |
| 18 Mar 2024 | Deadline for comments, suggestions and changes to Ateco2025 V2 version. |
| 08 Apr 2024 | Release of the third official draft of the Ateco2025 classification: structure and explanatory notes (**Ateco2025 V3 version**) aligned to the NACE Rev. 2.1 version 1.02. The document has been shared within the ATECO Committee, the network of stable users of the ATECO classification and all statistical domains within Istat. |
| 24 Apr 2024 | Deadline for comments, suggestions and changes to Ateco2025 V3 version. |
| 4 Jun 2024 | **Meeting of the ATECO Committee** to finalise the structure of Ateco2025 before sending it to Eurostat for approval. |
| 01 Aug 2024 | Release of the fourth official draft of the Ateco2025 classification: structure and explanatory notes (**Ateco2025 V4 version**) aligned to the NACE Rev. 2.1 version 1.03. The document has been shared within the ATECO Committee, the network of stable users of the ATECO classification and all statistical domains within Istat. |

### 3.2.2 Developing explanatory notes and ensuring EU consistency

Explanatory notes in a statistical classification offer detailed descriptions to clarify the meaning of labels and titles of the codes. While titles alone may be insufficient, these notes provide context and support for understanding the activities covered by a classification category. However, explanatory notes are more than just definitions; they must follow specific guidelines, some of which are implicit but well-understood by experts in the field.

During the process of defining Ateco2025, Istat developed some guidelines for drafting the explanatory notes of the new classification and chose to structure them into five types: '*Central'*, '*Inclusion'*, '*Also includes'*, '*Exclusion'*, and '*Implementation'* rules. ***Central notes*** provide a general overview of a code's content, often starting with phrases like '*This section/division includes*'. They are particularly useful for high-level categories (sections, divisions, groups, classes). ***Inclusion notes*** offer examples of activities covered by a code, but they are not exhaustive and usually they are used for the lowest level of the classification. ***Also includes notes*** identify activities that are conventionally classified in a category but could also fit elsewhere. ***Exclusion notes*** list activities that are excluded from a category due to their belonging to others. ***Implementation rules*** provide guidance for applying explanatory notes correctly.

All explanatory notes from the NACE Rev. 2.1 have been translated into the national version. Additionally, new explanatory notes specific to the national context have been included, ensuring they align with the European classification. This combined set of notes forms the complete national classification, integrating both the translated European notes and the newly created national ones.

As for the new explanatory notes introduced in the national version, which were promoted by national stakeholders to better clarify the location of certain economic activities within the classification, in some cases they were inherited from the previous version of the ATECO classification. They may focus on economic activities that are particularly widespread or important in Italy compared to what happens in other European countries.

Most of the explanatory notes were instead translated. Each Member State has adopted a different strategy to translate explanatory notes of the NACE Rev. 2.1 into its national language. An accurate and consistent translation is crucial to preserving the integrity and usefulness of the NACE classification at the national level. Clear explanatory notes are essential for the intelligibility of the classification; they can help reduce errors that may arise when coding descriptions of economic activities or when businesses seek the most appropriate code for their economic activities.

In Italy the translation process was a collaborative effort involving Istat experts, Swiss colleagues and national stakeholders. The developed guidelines ensured accuracy in the translation, maintaining consistency across different sections, addressing synonyms and homonyms, and balancing colloquial and technical language. Continued refinement of the explanatory notes is still ongoing, based on feedback and evolving needs.

Istat decided to use a cooperative approach, preferring a collaborative translation technique by exploiting different kinds of expertise and a multi-level strategy. The process involved several stages. Initially, statistical experts from each section of the classification proposed translations, guided by a structured manual outlining linguistic guidelines. Given the shared use of Italian in Switzerland, a collaborative effort was undertaken with the Swiss Federal Statistical Office to evaluate the translations and ensure their effectiveness and the first issue addressed was the establishment of common guidelines for the translation. To further enhance the quality of the explanatory notes, experts and stakeholders in specific sectors were consulted for their opinions and suggestions on technical terms. This open call for contributions was launched through the ATECO Committee, established to support the overall revision process. Finally, extensive consistency checks were conducted both within and between sections of the classification to address any remaining inconsistencies and ensure the final product's accuracy.

First of all, the **main criteria** or the subset of rules and standards that represent the general core of the drafted guidelines were defined, which are "general" in the sense that they apply to all types of explanatory notes. The main instructions emphasize that while translating, it is essential to maintain the original meaning and ensure consistency with the English version. If necessary, new notes can be added to clarify the original, but they should always be in line with the English content. Attention should be paid to style details such as punctuation, abbreviations and word usage; avoid quoting specific brands or places; and focus on describing economic activities with clear and consistent language. Although the use of a colloquial style can be helpful, technical terms may be necessary in some cases. The goal is to provide clear and accurate explanations that maintain the integrity of the original classification.

In addition, a set of guidelines regarding **editorial standards** was also drafted before the translation work began, which were later enriched during the translation phase. They emphasize clarity and consistency. For *Central explanatory notes*, often consisting of multiple statements, they recommend to divide them into shorter segments for easier understanding and suggest some Italian words that are preferable to use when the translation could be done in many different ways. *Inclusion notes* and *Also includes notes* must be introduced by a hyphen and should not end with a full stop and additional national notes can be inserted between the existing European ones, maintaining a logical order. In cases where inclusions occasionally contain exclusions, the recommendation is to indicate them in parentheses, to facilitate the reading and the application of automatic coding algorithms. *Exclusion notes* shall be organised according to the order of the ATECO code to which they refer (ascending order), they should start with a hyphen and do not end with a full stop. Abbreviations should be avoided or accompanied by their full descriptions. Hyphens are limited to specific cases and may be reduced in the future. Finally, indefinite articles are generally recommended over definite articles. Overall, these guidelines aim to ensure that explanatory notes are clear, consistent, and easy to understand, facilitating the effective use of the Ateco2025 classification.

In a similar way, the **translation process** also needed to define guidelines and make operational decisions agreed upon by the classification experts. The Italian language, rich in synonyms, presents challenges in translating technical terms. While different signifiers may convey similar meanings, homonyms can create confusion. Istat has focused on using consistent Italian translations for English words to avoid reader confusion. Translating a classification section requires translating related concepts in other sections to maintain linguistic coherence. This can be challenging when dealing with exclusions or inclusions in different industries. Istat's initial approach of assigning sections to SBR individual experts led to variations in translations. To address this, linguistic harmonisation was implemented to select the most suitable translations. The collaborative effort among different translators also sparked valuable discussions on effective translations. For example, the terms "caravan" and "motorhome" were initially translated differently but were later standardised for consistency. Collaboration with the Swiss Federal Statistical Office has been instrumental in this process. Translating technical texts requires meticulous attention to detail, especially when it comes to selecting the most appropriate Italian words. The translator must carefully consider the connotation of each word, its frequency of use in the Italian language, and how it fits within the broader context of the text to ensure that the translation is accurate, clear, and consistent. In some cases, it may be useful to retain the English term, especially if it is widely recognised or if there is no exact Italian equivalent, thus maintaining consistency with international standards and avoiding potential misunderstandings. When a concept can be expressed in more than one way in Italian, offering two or more translation options can provide flexibility to the reader. This is particularly useful when the choice of translation may depend on the specific context or intended audience. Furthermore, sometimes two or more English words may have similar meanings or be used interchangeably in a particular context. In these cases, a single English translation can be used to simplify the text and avoid redundancy. Moreover, sometimes the same English word can have different meanings depending on the context in which it is used. Therefore, you must carefully consider the surrounding text to determine the most appropriate Italian equivalent. This may involve using different translations for the same English word in different sections of the text.

Once the classification (or at least a draft version of the classification) has been developed, it has to be implemented in the SBR.

## 4. Implementing the classification in the SBR

### 4.1 The impact of the changes and the expected outputs for the SBR's users

The analysis of the impact of changes is part of the overall plan of implementation of a new classification in the SBR, giving the idea of the burden of recoding activities and consequently the type of tools and methods that the recoding process can use.

Assessing the impact of the new classification starts by comparing Ateco2022 with Ateco2025 in order to identify both the changes and the more complex cases that occur when old codes cannot be automatically transformed into new ones. To do this, the first step is to rely on a stable **theoretical correspondence table** between Ateco2025 and Ateco2022[3], since correspondence tables between different versions of the same classification are very useful to describe the detailed changes that have taken place in the revision process. From the theoretical correspondence table between Ateco2022 and Ateco2025, the following types of correspondences could be identified:

*a) 1-to-1 correspondences: one code in Ateco2022 corresponds exactly to one code in Ateco2025 and vice versa;*
*b) 1-to-M correspondences: one Ateco2022 code is split into two or more codes in Ateco2025;*
*c) N-to-1 correspondences: two or more codes in Ateco2022 correspond to one code in Ateco2025;*
*d) N-to-M correspondences: two or more codes in Ateco2022 correspond to two or more codes in Ateco2025.*

These types of correspondences indicate different levels of complexity in the reclassification of SBR units from the old to the new classification; for example, cases referred to in a) or c) are simpler (Total correspondence) than cases referred to in b) and d) are complex cases (Partial correspondence).

At five digit level the current Ateco2022 classification is adopted for all economic units, both those in the scope of Structural Business Statistics – SBS (considered the core of the SBR) and the other units represented by private and public institutions and the agriculture sector. By analysing the SBR units according to the types of correspondence between the old and the new classification, it is possible to assess the significant impact of Ateco2025 on the SBR (table 3).

*Table3 – Type of recoding by type of units*

| **Type of recoding** (a) | **N. of codes (Ateco2022)** | **SBR in SBS scope** | | **SBR not in SBS scope** | |
|---|---|---|---|---|---|
| | | **% units** | **% persons employed** | **% units** | **% persons employed** |
| *Partial* | 205 | 46% | 40% | 21% | 16% |
| *Total* | 715 | 54% | 60% | 79% | 84% |
| **Total** | 920 | 100% | 100% | 100% | 100% |

*(a) The adopted version of Ateco2025 classification is the version V4*

The SBR in the SBS scope requires complex reclassification procedures respectively for the 46% in terms of units and 40% of persons employed. This impact is less evident for the rest of SBR not in SBS scope (21% of units, 16% of persons employed).

Focusing on the units in the SBS scope and involved in multiple correspondences to recode, the distribution in terms of economic structure (table 4) shows how the impact affects 95.2% among the very small units; although they represent only 0.2%  large units account, in terms of persons employed this figure become more significant (23%). The Ateco2022 sectors most affected by the complex recoding process are wholesale and retail trade, transportation and storage and accommodation and food service activities (approximately 49% in terms of persons employed).

[3] The two-way correspondence table between Ateco2025 and Ateco2022 is consistent with the NACE correspondence table between NACE Rev. 2.1 and NACE Rev. 2 and also includes national correspondence relationships.

*Table4 – Units and employment by macro-sector and size*

| Macro-sector<br>(a) | Units | | | | Employment | | | |
|---|---|---|---|---|---|---|---|---|
| | Size | | | | Size | | | |
| | 0-9 | 10-99 | 100+ | Total | 0-9 | 10-99 | 100+ | Total |
| *C+D+E* | 112,783 | 22,519 | 1,474 | 136,776 | 270,827 | 532,407 | 446,520 | 1,249,753 |
| *F* | 307,697 | 15,626 | 240 | 323,563 | 555,767 | 296,553 | 49,259 | 901,579 |
| *G+H+I* | 603,104 | 43,552 | 1,820 | 648,476 | 1,285,318 | 875,758 | 830,637 | 2,991,713 |
| *J* | 10,507 | 564 | 66 | 11,137 | 14,587 | 13,487 | 29,050 | 57,124 |
| *K* | 14,929 | 86 | 9 | 15,024 | 5,508 | 2,297 | 1,740 | 9,544 |
| *L* | 197,918 | 490 | 14 | 198,422 | 225,160 | 8,426 | 2,288 | 235,874 |
| *M+N* | 396,582 | 9,345 | 807 | 406,734 | 529,815 | 217,773 | 280,717 | 1,028,305 |
| *P+Q* | 168,480 | 2,014 | 99 | 170,593 | 204,424 | 42,663 | 31,532 | 278,619 |
| *R+S* | 237,898 | 3,653 | 123 | 241,674 | 399,274 | 74,020 | 30,754 | 504,047 |
| **Total** | 2,049,898 | 97,849 | 4,652 | 2,152,399 | 3,490,679 | 2,063,383 | 1,702,497 | 7,256,559 |
| *%* | *95.2* | *4.5* | *0.2* | *100.0* | *48.1* | *28.4* | *23.5* | *100* |

*(a) The adopted version of Ateco2025 classification is the version V4*

In the Italian SBR it is planned to carry out the double-coding for two years, RY 2024, disseminated until the first quarter of 2026, and RY 2025, disseminated until the first quarter of 2027.

During 2025, the Italian administrative bodies will start using the national version of the NACE Rev. 2.1 classification in their production processes and above all in the collection of tax and chamber of commerce data. Using this information from administrative data – together with statistical information from STS surveys and ad hoc *Survey of Economic Activities for the Implementation of Ateco2025* – it will be possible to adapt the SBR by the end of 2025 (RY 2024) and then officially disseminate it in the first quarter of 2026 in double-coding NACE Rev. 2 (Ateco2007) and NACE Rev. 2.1 (Ateco2025).

According to the European Implementation Plan[4], envisaged to harmonize all statistical domains that make use of classification for the dissemination of their statistical outputs, the SBR must give support to all of them by providing data with the appropriate coding and within the necessary timeframe.

For the needs of SBS, which will have to disseminate its data in double-coding during 2027 (RY 2025), support was requested from the SBR by providing its data in double-coding also during 2027 for RY 2025.

Although the first dual coding of the SBR is scheduled by December 2025, there are many reasons why recoding should be expected at the end of 2024. In fact, all enterprises involved in statistical surveys can consult and update their identification and structural data, contained in the SBR, through access to the Business Portal and therefore also those concerning the economic activity carried out. At the beginning of 2025 they will be informed of the adoption of the new Ateco2025 classification and should also be able to consult the code assigned to them through recoding. Companies will be able to confirm this classification or provide the SBR with important information about their actual economic activity carried out. This means that, in order to support surveys and users accessing the BP, the RY 2023 version of the SBR will also need to be available in the dual coding, although higher quality will be assured to the subset of firms involved by SBS and STS surveys.

[4] Commission Implementing Regulation (EU) 2024/1840 of 27 June 2024 amending Commission Implementing Regulations (EU) 2020/1197, (EU) 2022/918 and (EU) 2022/1092, as regards references to the statistical classification of economic activities NACE Revision 2 established by Regulation (EC) No 1893/2006 of the European Parliament and of the Council.

### 4.2 Automatic and manual reclassification tools

### 4.2.1 Profiling activity

As well known, profiling is a method to analyse and maintain the legal, operational and accounting structure of an enterprise group at national and world level, in order to establish the statistical units (enterprises) within that group, their links, and the most efficient structures for the collection of statistical data.

Manual profiling activities usually provide useful and precious information on the economic activities carried out by large and complex enterprises, but it is costly activity in terms of time and resources.

Such information can be found by investigating annual reports (including consolidated and sustainability reports) as well as corporate websites. In order to support the implementation of the new Ateco2025 by recoding the most important units in the SBR according to the new classification, profilers were asked to update or assign the NACE code to these units according to the current classification (Ateco2022) and its new version (Ateco2025) simultaneously. Specifically, starting from March/April 2024, when the 2024 profiling cycle started (reference year 2023), the profilers registered information in an in-house tool developed to store data on economic activities according to Ateco2025, because at that time the Ateco2025 national metadata had not yet been implemented in the SBR information system. Indeed, at the time of writing, the definition of the Ateco2025 classification has not yet been finalised. As mentioned earlier, several drafts of the Ateco2025 classification have been produced over the months in order to provide users and profilers with an increasingly final version; the very final draft must be approved by the European Commission in accordance with the European Regulation. Once the new Ateco2025 is approved, its metadata will be implemented in the SBR and the information on economic activities collected by profilers and coded according to the new classification will be automatically transferred in the SBR information system after a validation phase aimed at checking that the economic activity code assigned by profilers is in line (still from a metadata perspective) with the final version of the Ateco2025 classification.

### 4.2.2 Planning a sample survey and preliminary results

Starting in April and running until July 31, 2024, Istat has conducted the Survey of Economic Activities for the Implementation of Ateco2025 (hereafter SEA). The objective is to detect the economic activity carried out by enterprises to use the information acquired to reclassify the units in the SBR according to the new Ateco2025 classification.

The survey targets a sample of enterprises (in SBS scope) drawn from the SBR consisting of about 150 thousand units, a limited set after evaluating all organisational aspects of the survey and the necessary accuracy of the estimates. Although the entire population of enterprises is affected by the recoding process, however, a selection criterion was adopted in choosing the sample that covers only those economic activities most affected by the changes. Operationally, the target population consists of enterprises in the SBR that are classified in Ateco2022 codes whose correspondence with Ateco2025 is not 1:1 (or Total) but is 1:M (or Partial). These codes with multiple correspondences are 171, resulting in 1.9 million of units in the SBR[5]. The sample size for each Ateco2022 class is determined so that their total does not exceed the above sample limit.

To formalize the *sample allocation*, the method of determining the sample size for a single Ateco2022 stratum is as follows (then the overall sample is obtained as the sum of the strata):

$$1-\alpha = \Pr(\hat{p}_i - d_i \leq\ p_i \leq \hat{p}_i + d_i,\ \forall i = 1, \dots, k)$$

[5] Since the sample selection operations were carried out several months before the data collection time, the survey used a provisional Ateco2025 classification (version V2).

where $k$ is the number of possible Ateco2025 codes, which can vary from 2 to m depending on the Ateco2022 considered and $\alpha$ is the highest acceptable risk that one or more frequencies of the probability distributions to be estimated for a given class of Ateco2022 are outside a fixed $2d_i$ interval of confidence; that is, at least one frequency of interest falls outside its confidence interval and then 1-$\alpha$ is the probability that no interval is violated. Then the overall sample is obtained as the sum of the strata.

If we define $\alpha_i$ as the risk that the interval $2{*}d_i$ does not contain the value of the corresponding $p_i$, we can consider the following relationship: $\alpha \leq \sum_{i=1}^{k} \alpha_i$ The inequality is derived from the total probability theorem, due to the fact that two or more intervals can be violated simultaneously. Using this inequality, we calculate the sample size $n$ when considering the uniform distribution $p_i = p = 1/k \ \forall i$ for the k Ateco2025 classes of an Ateco2022 stratum:

$$n = \left( \frac{m}{1 + \frac{m-1}{N}} \right) \quad \text{with} \quad m = \frac{z_{\alpha^*}^2 \hat{p}(1-\hat{p})}{d^2} \alpha \leq \sum_{i=1}^{k} \alpha_i$$

An attempt was made to determine $n$ with a maximum risk $\alpha$ that one or more confidence intervals of the Ateco2025 frequencies would be violated. In other words, we wanted to ensure a probability of 1-$\alpha$ that the estimates would meet all the respective intervals.

The working hypotheses were the following: the semi-confidence interval $d_i$ is set to the minimum between the values 0.03 and $1/2k$, $\alpha$ is fixed at 0.1 when $k$ is greater than or equal to 20, otherwise it is fixed at 0.05.

Operationally: 1) we considered that we had to estimate a uniform distribution; 2) we set the $d_i$ semi-intervals so that the CV of the estimates never exceeds 25% (it happens when p<6%). For example, if p=0.5 the CV is 3%, if p=0.2 the CV rises to 7%, if p=0.1, CV=15%. 3) we set the risk at 5%, except for the 7 strata of Ateco2022 where the existing 'possible' Ateco2025 were more than 20. We also calculated $n$ even in the case of a plausible non-uniform distribution and chose the larger sample of the two cases.

After finishing the sample drawing phase, we moved on to the Data Collection, based on Computer-Assisted Web Interviewing (CAWI), which entailed the development of an electronic questionnaire. This was a collaborative process involving the classification experts and computer scientists, who worked together to create the questionnaire, which was then subjected to a series of tests before being administered to the enterprises.

As far as ***the questionnaire*** is concerned, it is a short questionnaire (form) to collect information on the economic activities carried out by the enterprise. It was optimised to reduce redundancy in the information requested, harmonise concepts and definitions, and make it easier for respondents. In particular, the respondent is guided in the choice of the economic activities performed according to a top-down logic (from the general to the particular) and is therefore asked to choose, in order: the main macro-sector of economic activity (high level aggregation of economic activities); the main sector of economic activity (intermediate level aggregation of economic activities) and finally the specific economic activity or activities carried out by the company. The questionnaire does not have any pre-printed ATECO codes (neither Ateco2022 nor Ateco2025), nor does it require any codes to be indicated during compilation. Internal Istat experts are working on the reclassification of the activities indicated in the forms, only for the necessary statistical purposes.

The ***data collection process*** was meticulously planned and executed. It consisted of several crucial stages, in the preparatory and the data collection phases, with innovative and supportive services to guarantee information accessibility and transparency. A comprehensive participant information sheet was created, outlining the survey detail to complement the questionnaire and processed sample contact information were meticulously crafted. Additionally, a support tool was implemented to manage participant assistance requests through the Contact Centre, whose staff underwent our training (both thematic on Ateco2025 and non-thematic) to acquire the necessary skills to assist participants effectively. Furthermore, a monitoring report was prepared to track daily/weekly survey progress and response rates. During the data collection phase two reminder emails were sent to non-responding units (May 20th and June 10th), followed by phone reminders starting June 11th, a planned outbound activity to reduce nonresponse and targeted the most relevant non responding enterprises. Throughout the data collection period, questionnaire completion rates, response rates, and non-response rates were closely monitored. This intricate process aimed to maximize survey participation and ensure the quality of the collected data.

No publication of the survey results is envisaged, as this is a survey whose primary objective is to allow the reclassification of the units in the SBR according to the new Ateco2025 classification, thus improving the quality of the information content.

Using the survey data presupposes carrying out some activities to make the results usable. In fact, in some cases it was not possible to automatically assign a code from the respondents' answers, and the output consists of a description of the activity performed, a textual string that needs to be coded. What is more, there are cases where the responses indicate that the starting Ateco2022 code was incorrect (either due to an assignment error in the SBR or an actual change of activity) so this involves changing the starting stratum of these units.

Therefore the steps followed in the ***Data quality management plan***, to check and correct the survey data, were the following, which will be performed recursively until codes are assigned to each unit in the sample.

The first step is to code all the strings and assign the main Ateco2025 code to the units. This involves the selection and validation of an automatic coding system, to be used in a massive way and using timely checks by the SBR's classification experts, whose work was planned with an appropriate plan of assignments.

For the automatic coding an **algorithm of matching** between the SEA survey strings and the preliminary Ateco2025 classification was experimented. It considers text pre-processing with word weighting and similarity metrics based on the intersection of words in common. Strings are treated differently depending on their length; specifically:

- long strings, high similarity → match
- short strings, relatively high similarity → match
- short strings, low similarity, but high if synonyms are also applied → match
- otherwise → no match

The association of the Ateco2025 for each string is done by the following process:

a. extraction of the one or more Ateco2025 codes by the similarity of the words in common, taking those with the highest similarity (multiple choices are possible);

b. choosing the final (mot reliable) Ateco2025 code from those selected with a transformer-type deep learning model (BERT). The choice of first making a selection of Ateco2025 and then using the deep learning model on the shortlist was dictated by the fact that the outputs of the model had to be somewhat driven.

This algorithm is not yet optimised and need to be improved and refined.

Given the use of a provisional classification and correspondence tables in the survey, a further step is needed to convert the findings to the final Ateco2025. This entails constructing comparison tables between provisional and final versions and manually modifying the survey template.

Next, respondents with coded Ateco2025 and those with Ateco2025 to be coded/resolved can be identified and for each Ateco2022 stratum their absolute and relative frequencies calculated for the two groups of units. At this point the initial completeness rate (respondents with coded Ateco2025/total respondents) and response rate (total respondents/total sample) for each Ateco2022 can be calculated and used for the following analysis.

Then follows a compatibility check of the newly detected/assigned code against the starting activity, i.e., it is checked whether the Ateco2025 code belongs to the set of codes that in the theoretical correspondence table corresponds to the Ateco2022 code under which the unit was registered in the SBR. If yes, the correspondence is considered “valid” and the group of units “coded and consistent” continue to the next steps, otherwise, assuming that the new code detected is correct, a different Ateco2022 code must be assigned and the unit moved to the new corresponding stratum, as if it had been the starting one in the survey. The latter operation is also carried out partly automatically (using the reverse correspondence tables, from new to old classification) and partly with the support of manual expert review, at least for the most relevant enterprises.

On the “coded and consistent” subset, which will gradually increase until it contains all the units in the sample, the following steps are taken: for each Ateco2022, the absolute and relative frequencies of detected Ateco2025 are calculated; a target value (p=0.6) is chosen above which the percentage of detected/coded cases in a stratum is considered sufficiently high or acceptable to elect that Ateco2025 code as the choice to be associated with that Ateco2022 code even for cases not detected by survey.

In order to optimize the control process, a preliminary analysis of the distribution of Ateco2025 codes will be carried out, to assess whether it is uniform or has a single peak (unimodal). Then all confidence intervals will be calculated to identify the areas with the highest risk of error. For each Ateco2022 code, the percentage associated with the most frequent Ateco2025 mode will be used as the value of p. The width of the resulting confidence intervals will guide the selection towards the most critical cases among the uncoded cases, to be subjected to manual verification, thus increasing the effectiveness of controls. Each time the strata with the highest error rate will be monitored and reported to the auditors in order to intensify targeted controls.[6]

Currently, the activities described above are still ongoing and at a different stage of progress. It is however possible to **present the main results**. The response rate of the survey is about 43% that is a good result considering the fact that the survey has no obligation to respond and was affected by some organisational problems in the data collection phase. Almost 71% of the respondents have a principal Ateco2025 code correctly assigned; while for the other 29% of units the multiple responses are expressed in a mix mode, some are codified others are strings of descriptions of economic activities. The number of strings amounts to 36,077 record and the process of automatic coding is still ongoing.

### 4.2.3 Using auxiliary admin sources

Using administrative data as auxiliary sources to scale up survey data can significantly enhance the precision and reliability of estimates. Using survey data, as already seen, we have estimated a transition matrix between old and new economic activity codes, that is, for each code x of the old classification we have the probabilities of transition to codes y1, y2 ... yn of the new classification. To allocate new codes (y1, y2, ..., yn) to all units in the SBR with an old code (x), we tried to leverage the auxiliary variables available for each unit.

The methodological experts investigated various possibilities and one approach involves creating a propensity score for each possible new code, based on the auxiliary variables. This score represents the probability of a unit transitioning to a specific new code given its characteristics, that is an estimated probability that a unit belongs to the groups having a new specific activity code based on its observed characteristics. Then, you can randomly

[6] A threshold is set to determine if an Ateco2022 category has enough data to make a reliable prediction about the Ateco2025 code (a target value set at 60%) and the confidence interval helps assess the uncertainty in the prediction. Depending on how the confidence interval relates to the threshold, the Ateco2022 category is marked as 'go', 'ok', or 'ko', indicating whether more data is needed, assignment is possible with potential errors, or assignment is unreliable. If the confidence interval includes the threshold, the Ateco2022 stratum is marked as 'go' and additional units need to be classified (it means we are not certain if the distribution of Ateco2025 codes within that Ateco2022 category will remain stable if we classify more companies. Therefore, more data is needed to make a definitive conclusion). If the lowest possible value within the confidence interval is already above the 60% threshold, it is likely that the Ateco2022 stratum has a high enough frequency of the Ateco2025 code to be assigned and the stratum is marked as 'ok'. However, there is still a possibility of error, which needs to be considered. When the highest possible value within the confidence interval is below the 60% threshold, it is highly unlikely that the Ateco2022 stratum has a sufficient frequency of the Aeco25 code. Therefore, any prediction or assignment based on this data would be unreliable and the stratum is marked as 'ko'.

select units within strata defined by these propensity scores to assign the new codes, ensuring that the distribution of auxiliary variables within each stratum is similar to the overall population. This method, known as propensity score matching, helps to mitigate potential biases in the allocation process and improve the representativeness of the scaled-up estimates.

Other methods that have been considered are: i) logistic regression to estimate propensity scores, ii) decision trees to model the relationships between auxiliary variables and the activity code, and finally iii) machine learning for a more sophisticated approach to classification and prediction.

Of course, the choice of method depends on the nature of the data, the sample size with respect to the availability of each source on its units and the objectives of the analysis. It is important to carefully assess the quality of the auxiliary variables and their relationship to the phenomenon of interest. In addition, validation of the estimates obtained is essential to verify the reliability of the results.

To make use of all the wealth of information offered by auxiliary sources, it is certain that it will be necessary to revise the entire process of assigning economic activity in the SBR, not only for Legal Units but also for Enterprises and Local units. We will use administrative sources differently than in the past, employing automated string coding using deterministic and machine learning systems and combining these results with data from the other administrative sources and the statistical information.

In the past and up to now, we used the sources that provided information directly through codes (mainly VAT, ISA, published business accounts, Central bank and IVASS and, in a residual way, Chambers of Commerce and Tax Register in the absence of anything else or to confirm in case of discordances). Now we are broadening the analysis to the other sources that provide the information by textual description of the activity.

We are testing new methods of text analysis to code in an automatic way the strings that are present in the administrative sources that record descriptions of economic activity carried out by legal units (in future, we will receive also descriptions of economic activity in the Chambers of Commerce master file):

- in the notes on balance sheets;
- in the Synthetic Index of Reliability (ISA, Revenue Agency).

The contents of the balance sheets are known and need no further elucidation, and the notes to the accounts are the additional information and explanations that accompany the financial statements. They provide more details and clarity about the items, amounts, and transactions reported in the balance sheet, income statement, statement of changes in equity, and cash flow statement.

The second source (ISA) is a special survey managed by the Tax Authority, basically designed for fiscal purposes addressed to SME. The data available from these detailed sector studies provide Istat with more appropriate information in order to classify the economic activity of enterprises as information is thus available on inputs, process and outputs, expressed with textual description. Using currently available information, therefore, for a significant number of units the correct code of the new Ateco2025 could be chosen in cases where a code of Ateco2022 corresponds to several new codes in the table of correspondence between old and new classification.

### 4.2.4 Mapping and operational correspondence table

Once the new classification is defined both in terms of structure (codes and titles) and explicative notes, correspondence tables represent the most relevant theoretical framework to support recoding activities. In absence of any kind of information describing the economic activity at individual level, whether from statistical surveys or administrative sources, some methods can be developed to assign new codes on a group basis.  The method based on the use of a unique code can only be applied to the one-to-one, or many-to-one changes from the old to the new classification i.e. using simple type of correspondences a classification at the lowest aggregation level is directly recoded to the revised classification. In those cases, the corresponding new code is the only eligible one. In order to solve in an automatic way the cases of one-to-many splits at macro level, an operational correspondence table between the old code and the new codes was developed. The logic underlying the operational table is to associate each Ateco2022 code (the old code) with a single, most representative, Ateco2025 code (new code) from among those proposed by the theoretical correspondence table. The choice of the Ateco2025 code to be uniquely associated to the starting Ateco2022, is the result of an automatic matching algorithm that compares headings and inclusion notes (strings) of the two classifications Ateco2022 and Ateco2025. This algorithm assigns a similarity measure to each of the M possible pairs associated to the Ateco2022 code, resulting from the comparison of the words into which the set of strings (notes and titles) is broken down. Similarity measures are calculated using absolute and/or relative frequencies obtained as the result of the intersection of words between Ateco2022 and Ateco2025. Of course, the more words written in the same way between the two classifications, the more likely it is to find matches with high similarity. In order to obtain accurate results, some attention is paid to the text pre-processing phase prior to the comparison of strings. For examples, irrelevant words (e.g. articles and prepositions) are eliminated and a stemming process is performed to treat words with the same root (singular, plural); synonyms are also used. The search and comparison constraint is the vector of theoretical Ateco2025 codes correspondent to each Ateco2022 code. The operational correspondence table chooses the most representative Ateco2025 code that is, among the possible M, the one that inherits most of the content in terms of economic activities described present in the corresponding Ateco2022 code. A measure of similarity is associated to each chosen new code indicating its quality. This representative new code then can be used directly in an automatic recoding.

The provision of a tool to automatically convert even multiple-match cases is a necessity for the Italian SBR, which 95% consists of very small units, whose updating depends exclusively on the use of administrative sources. Unfortunately, at least in the first year of dual coding, administrative sources can be used partially and with caution since they too must adopt and implement the new classification and consequently the statistical procedures and the estimation methodologies used by SBR based on them have to be updated. In fact, all administrative bodies adopt the same Istat ATECO classification at the same time and the strategy of its implementation is part of a common implementation plan coordinated by the ATECO Committee which effort is to use harmonized and coordinated tools. It is the case of the Register maintained by the Italian Chamber of Commerce (CCIAA) whose implementation plan is to recode the entire register ‘ex officio’ thanks to the use of operational correspondence tables. This tool thus becomes a single and shared tool between the two bodies, Istat and CCIAA.

### 4.2.5 Automatic classification tools

To support classification users, both internal and external, it was decided to update the CIRCE classification tool, available on the Istat website, which was an R package developed by Istat for the automatic coding of texts. It was designed to automatically assign a code to a given text, but the system only works by providing results in the old Ateco2022 classification.[7]

[7] CIRCE is one of the systems based on weighting algorithms. It manages applications of i) automatic coding: i.e., coding of entire files (batch mode); ii) interactive coding: which, with the help of the graphical interface, allows for interactive analysis of the coding of individual cases; iii) web coding: i.e., web service for coding single strings. In the latter case, a web service dedicated to ATECO coding is currently available through the page.
Regardless of the type of coding, the comparison between the text to be coded and the entries contained in its dictionary is preceded by the text standardization phase defined as parsing. This phase is completely controlled by the user, who is responsible for adapting it to the specific application context. The aim is to eliminate grammatical or syntactic differences so that two descriptions with identical semantic content can be considered identical. A set of 14 different parsing functions are available, including: character mapping, removal of words or strings deemed irrelevant, removal of prefixes and suffixes, and synonym treatment. Following the standardization of the texts is the matching phase. The standardized text is compared

Updating the system each time to a new classification, especially with a view to its future continuous updating, involves a great deal of work on the part of the classification experts. This work must be continued in any case, but with the occasion came the idea of moving towards other methods of automatic coding not only because they are more innovative (e.g. machine learning models can learn more automatically) but also to overcome technical issues related to the use of R and to the confidentiality issues.

Although the development of a sophisticated new classification system for economic activities was planned, it was realised early on, when carefully analysing the new structure of the Ateco2025 classification and the available Circe database, that there were some gaps in the available data that could have hampered the effectiveness of a traditional machine learning approach. To overcome these challenges, we chose to implement a hybrid solution that combines the strengths of deterministic rule-based methods and machine learning methods.

The new system will feature two main components:

- **An online application** that uses a deterministic search model to provide quick and accurate results for simple queries. This tool will also play a crucial role in training the machine learning model by collecting user interactions and feedback.
- **An offline application** powered by a machine learning algorithm, capable of handling more complex and nuanced classification tasks. This model will be trained using data from the online application, as well as additional data from administrative sources (above all the Chambers of Commerce).

To ensure seamless integration and flexibility, both applications will be implemented as RESTful API services. This will allow us to easily switch between the two, based on the complexity of the query and the performance of the models. For those instances where neither application can provide a definitive answer, we will have a process in place to route the query to a human expert for manual classification. Furthermore, to expand the reach of our system and enhance the training data for our machine learning model, we plan to make the online application accessible to the Public Administration via an API in the future.

In summary, our new classification system is designed to improve accuracy by combining rule-based and machine learning approaches, enhance efficiency by streamlining the classification process for various query types, scale effectively to handle growing data volumes and changing user needs, and offer flexibility by adapting to changes in data and user requirements.

This innovative approach will provide a more robust and adaptable classification system, ultimately leading to more accurate and efficient economic data analysis.

Among the most delicate and challenging activities were, for the classification experts, the creation of the training set necessary to train the model, and for their IT colleagues, the phase of identifying the necessary data pre-processing rules, prior to the use of the strings (both those of the official classification, i.e. titles, explanatory notes and index, and those typed in by users).

As far as the first is concerned, setting up the training set is not an easy task and it implies several efforts by classification experts. In order to reduce the burden of this activity, we decided to take advantage of clarification requests presented by external users in the last years (emails and entries registered in the official automated coding system CIRCE) to create the training set.

As regards the pre-processing phase, it is essential for extracting ‘tokens’ (word that serves as a building block in natural language processing and text analysis) from the text to power both our online text search and offline

with the descriptions, also standardized, of the reference dictionary. If the result of the comparison is a direct match, then the software assigns a unique code. Otherwise, a weight-based algorithm is used to identify the best partial match, thus providing an indirect match. Being a product developed internally at the Institute, it offers the opportunity for modifications and/or additions of new features, both related to the set of parsing functions and the matching algorithm.

classification. Tokens help break down text into smaller, manageable parts, making it easier for computer to understand and process. This is essential for tasks like searching for specific phrases or classifying text into categories. In the online application, tokenization assists in text search within 'tags', while in the offline process the tag is constructed and used in the classifier. The functionality of tags can be expanded by associating them with synonyms.

Pre-processing is structured in the following steps.

- Removing non-informative text with regular expressions: e.g., special characters or symbols.
- Breaking down the text into tokens and removing stop words (common and non-informative words).
- Assigning weights to tokens using inverse document frequency: tokens (unigrams) are weighted based on the inverse of their frequency within the corpus, thereby giving more weight to rare and distinctive terms.
- N-gram analysis and weight calculation: exploring all combinations of tokens (bi-grams and tri-grams), rather than single tokens, to identify more complex relationships between all possible combinations for improved selectivity. Similar to single tokens, it is necessary to recalculate the weight of token combinations (inverse term frequency ITF). This allows for the identification of a greater complexity of textual patterns present in the data. However, it is important to consider that increasing the size of n-grams leads to an increase in computational cost and modeling complexity. One should balance the benefit of capturing more detailed patterns with the ability to process and manage this data efficiently.
- Reducing words to their base form (extracting root words): i) Lemmatization: considers context and part of speech. ii) Stemming: operates on words individually, ignoring context.
- Composition of selective tags.
- Synonym acquisition and gathering.

## 5. An integrated approach to the use of the different tools and methods

After conducting the analysis of the information from internal and external sources (including profiling) as well as the identification of techniques and methods to classify all SBR units in Ateco2025 codes, the overall strategy at micro level has been designed giving priority to different kind of information. The identification of complex and simple correspondences represents the starting point of the whole strategy; for the complex ones we will use a multi-strategy approach, using all the available administrative and statistical sources.

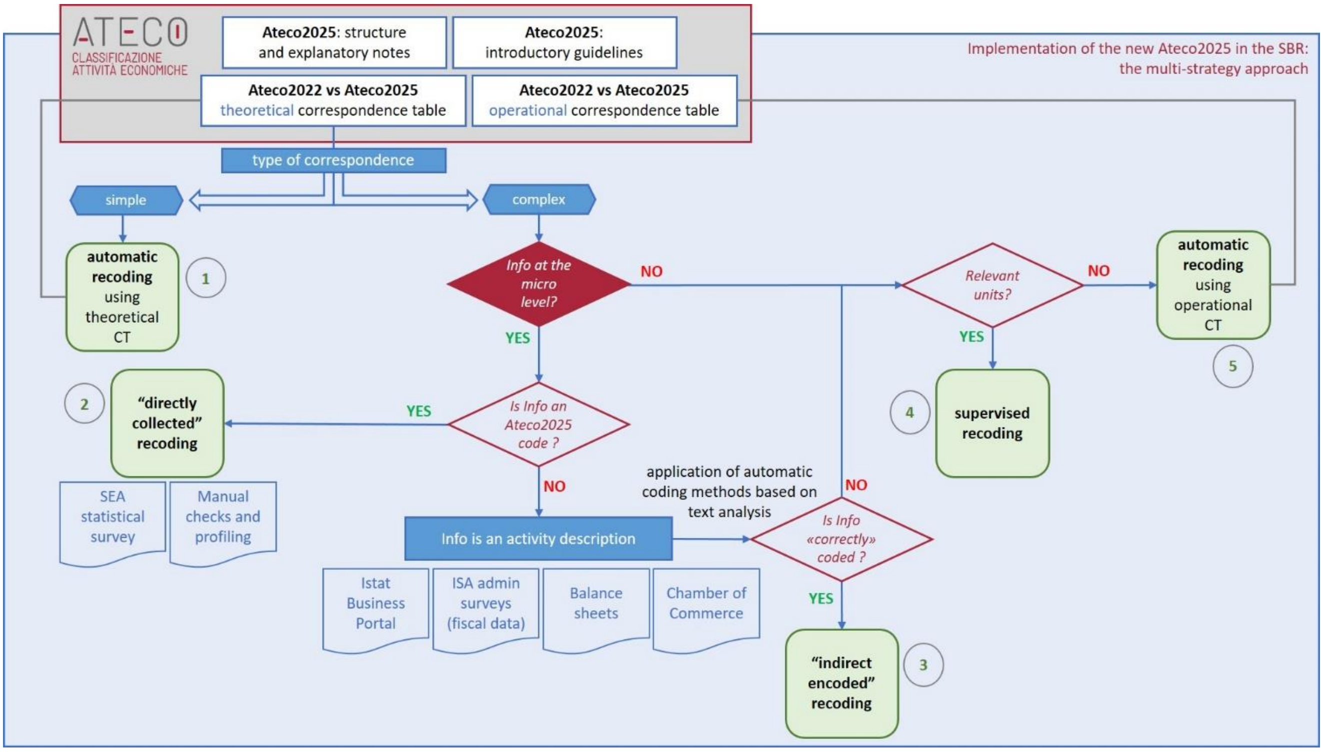
First of all, **simple automatic recoding** methods [1] are used when the unit is classified according to an Ateco2022 code belonging to simple correspondences between Ateco2022 and Ateco2025; the theoretical correspondence table is applied in these cases.

When the correspondence is complex, the availability of information on economic activities at the micro level is investigated. When existing, it can be provided as an Ateco2025 code or as a description.

A **“directly collected” recoding** approach [2] is applied to respondents to the SEA statistical survey as well as units involved in profiling activities because in all these cases the information on economic activities is directly provided and thus it is easily codifiable according to Ateco2025. In all other cases, when the information on economic activities is not already coded, automatic coding methods based on textual analysis are applied. If they success in providing an Ateco2025 code it means that the unit is reclassified according to **“indirect encoded” recoding** approach [3]. In all residual cases and in absence of information at the micro level, **operational recoding tables** (or operational correspondence table) [5] are used if the units are not relevant; otherwise, a **supervised recoding** activity [4] will be undertaken by SBR staff.

The overall strategy is presented in the picture below.

*Picture 4 – The multi-strategy approach to recoding the Italian SBR at a glance*


ATECO
CLASSIFICAZIONE ATTIVITÀ ECONOMICHE
Ateco2025: structure and explanatory notes
Ateco2025: introductory guidelines
Ateco2022 vs Ateco2025 theoretical correspondence table
Ateco2022 vs Ateco2025 operational correspondence table
Implementation of the new Ateco2025 in the SBR: the multi-strategy approach
type of correspondence
simple
complex
automatic recoding using theoretical CT
1
Info at the micro level?
NO
YES
Relevant units?
NO
YES
automatic recoding using operational CT
5
2
"directly collected" recoding
YES
Is Info an Ateco2025 code ?
4
supervised recoding
SEA statistical survey
Manual checks and profiling
NO
application of automatic coding methods based on text analysis
NO
Info is an activity description
Is Info «correctly» coded ?
YES
Istat Business Portal
ISA admin surveys (fiscal data)
Balance sheets
Chamber of Commerce
"indirect encoded" recoding
3

## 6. Conclusions

The implementation of the new classification of economic activities in the Business Register has been a complex endeavour, especially due to the diverse expertise required from statisticians, classification experts, methodologists and IT professionals. Managing the activities across these different professional backgrounds proved challenging at times, as each group had its own approach and understanding of the objectives the project. Despite our best efforts, several challenges emerged during the project's implementation phase. One significant hurdle was the sheer diversity of expertise and roles involved in the project. Integrating the perspectives and methodologies of statisticians, classification experts, methodologists, and IT professionals required a delicate balance and frequent communication to ensure alignment with project goals. Also external stakeholders had a central role in the whole process, especially within the development of the Ateco2025 classification phase.

While our commitment to transparency was unwavering – tracking all operations, disseminating each subsequent version, and documenting every change request – it inadvertently led to a heavier burden on the classification experts. This transparency, while beneficial for clarity, risked causing confusion among users due to the management of frequent updates and changes. The transparency we aimed for, while commendable, created an overwhelming workload for the subject matter experts, who had to continuously update and adapt to changes. This not only strained resources but also risked creating confusion among end-users due to the frequent modifications which were however necessary in order to guarantee a perfect alignment with the NACE Rev. 2.1 that was still work-in-progress.

Furthermore, the Italian classification's complexity, compared to the European one, led to additional operational challenges, particularly in managing code changes and ensuring coherence. The difference between the Italian classification (6-digit codes) and the European classification (4-digit codes) added another layer of complexity, additional operations due to the greater number of codes and especially those in complex correspondences between the old and new classification, which required meticulous management to maintain consistency with the European classification.

As we conclude the first steps for the implementation of the new classification of economic activities in the Business Register, it is clear that this project has been both challenging and rewarding. Despite the complexities and hurdles we faced along the way, we have successfully achieved our primary objectives and delivered a comprehensive and transparent classification system, almost surely of a better quality than it was in the past.

The main steps of the multi-strategy approach to implementation have been carefully considered; they are listed below.

1. Defining the national classification (structure, explanatory notes, correspondence tables and also introductory guidelines).
2. Assessing the impact of the changes introduced by the new classification on the BR in order to identify complex and simple correspondences between Ateco2022 and Ateco2025.
3. Analysis of the information from internal and external sources (including profiling).
4. Mapping of available information and identification of techniques and methods to classify all SBR units in the new Ateco2025 codes distinguishing between cases where information on the economic activity at unit level is available from other cases. When such information is not available, the use of operational recording tables has been preferred.
5. Implementing the overall strategy at micro level giving priority to different kind of information.

Looking ahead, now that the entire planning phase has been completed and the objectives and actions to be taken are clear, we hope to be able to cover all units of the SBR with adequate quality. We are also confident that the insights and the knowledge gained so far will serve as a valuable foundation for meeting future challenges with greater confidence and expertise.

## References

**ALONZI, F.,** (2021), *The new organization set up in Istat to maintain, update and revise the classification of economic activities.* Meeting of the EU Standards Working Group.

**ALONZI, F., CONSALVI, M., GENTILI, B., VIVIANO, B.,** (2024), *Planning a new ad hoc survey to recode units in the Italian SBR,* Eurostat NACE implementation webinar (29-30 April 2024).

**ALONZI, F., VIVIANO, C.** (2022), *The new updating process of the ATECO classification.* IAOS Conference, Kraków, 26-28 April 2022. https://www.iaos2022.pl/

**AMBROSELLI, S.,** (2010), *A New Methodology For Determining The Main Economic Activity Code In The Italian Business Register*. Presentation in the "Study visit: Business register and census of economic enterprises".

**AMBROSELLI, S.,** (2011), *Metodologia per l'attribuzione del codice Ateco 2007 - Registro Asia* Istat Working Papers 5/2011. Roma: Istat. https://www.istat.it/it/archivio/31343

**AMBROSELLI, S., Vicari P.,** (2007), *New economic classification and new instruments for Business Register classification: an opportunity to improve the quality in the Business Register*. International Roundtable on Business Survey Frames – Wiesbaden Group on Business Registers.

**AUTORI VARI,** (2015), *Atti del 9° Censimento generale dell'Industria e dei Servizi e Censimento delle Istituzioni non profit. Fascicolo 2: Il censimento delle imprese*. https://www.istat.it/it/archivio/179737

**AUTORI VARI,** (2020), *La classificazione delle attività economiche Ateco 2007: il processo di manutenzione e aggiornamento.* Istat Working Paper N. 1/2020 ISBN: 978-88-458-2008-3.

**AUTORI VARI,** (2021), *L'aggiornamento periodico dell'ATECO per finalità statistiche e amministrative* - 14^ Conferenza Nazionale di Statistica - Spazio Innovazioni e Progetti.

Commission Delegated Regulation (EU) 2023/137 of 10 October 2022 amending Regulation (EC) No 1893/2006 of the European Parliament and of the Council establishing the statistical classification of economic activities NACE Revision 2 (Text with EEA relevance).

**CONSALVI M., GENTILI B., SPERANZA A.,** (2019), *The new role of the SBR within the Italian Business Portal*, Meeting of the Group of Experts on Business Registers.

**CONSALVI M., FAZIO N.,** (2012), *The Business Portal*, 23rd Meeting of the Wiesbaden Group on Business Registers - International Roundtable on Business Survey Frames, Washington.

**EUROPEAN COMMISSION,** (2023), *Handbook on implementation of NACE Rev. 2.1 in Business Registers* (Version 1.2 - April 2023).

**EUROPEAN COMMISSION,** (2024), *NACE Rev. 2.1 backcasting & double reporting manual* (draft version 6).

**ISTITUTO NAZIONALE DI STATISTICA - ISTAT,** (2022), Classificazione delle attività economiche - Ateco 2007. URL: https://www.istat.it/it/archivio/17888

Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 establishing the statistical classification of economic activities NACE Revision 2 and amending Council Regulation (EEC) No 3037/90 as well as certain EC Regulations on specific statistical domains Text with EEA relevance.

**ZELLER., M.,** (2021), *Ways of implementation of a new activity classification.* Meeting of the Standards Working Group.

## Annex – The questionnaire and its cognitive testing

To develop the questionnaire of the special survey held to determine the main activity of SBR units, the recommendations of the "Handbook on the implementation of the NACE Rev. 2.1 in SBRs' have been followed. As previously mentioned, the special survey is based on the CAWI methodology, therefore, the questionnaire is electronic, with significant advantages in terms of lower costs and the possibility of collecting information on unexpected activities without additional effort.

The top-down method or the bottom-up method could be adopted for designing the survey questionnaire. The top-down method follows a hierarchical principle starting with the identification of the enterprise macro-sector of activity, usually through closed questions, and then progressively identifying the activity with the highest share of value added. The main shortcoming of the top-down method is that the incorrect choice of the macro-sector at the beginning implies misclassification error of enterprise economic activity. The bottom-up method starts with the simplest request to enterprises to provide a detailed description of their activity usually through open questions. Unfortunately, this method can be very demanding for both the respondent and the NSI, especially when open-ended questions are used and respondents do not answer clearly. After a thorough assessment, the Task Team have decided to adopt the top-down method for the development of the questionnaire limiting the drawbacks of this method with some solutions.

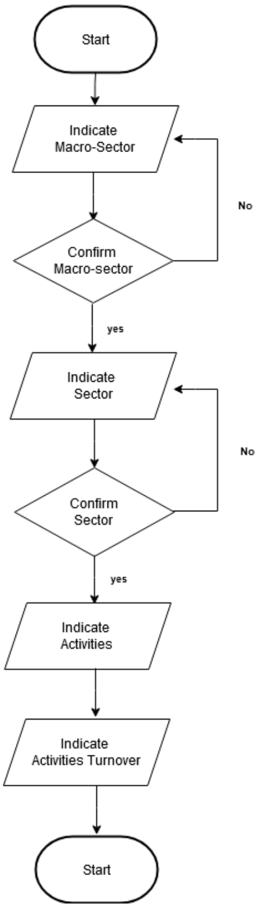
As recommended by the Handbook, the draft version of the questionnaire was tested by a cognitive testing on a small group of enterprises.

The structure of the final questionnaire is quite simple (Figure 2.4), divided into two main sections and an additional section to collect contact information and comments from respondents. In the first section, the respondent is invited to indicate the main macro-sector of activity of the enterprise. The modalities of the macro-sector list are often economic activities at section level of NACE Rev. 2.1 and in few cases a higher (aggregation of sections) or lower (divisions) level of generality. Successively, the definition of the macro-sector chosen will be displayed and then the respondent can confirm his choice and proceed or indicate another macro-sector. The provision of the definition of the macro-sector selected and the possibility for the respondent to change his choice were thought to reduce misclassification error.

In the following question, the respondent has to indicate the main sector of activities according to the macro-sector previously selected. It is possible to display a list of sector of activities corresponding to the NACE Rev. 2.1 divisions (or their aggregation) by electronic questionnaire. For this question, the definition of sectors will be displayed and the respondent can confirm or change the selection made.

Section 2 of the questionnaire is developed to identify the activity carried out by the enterprise with the higher turnover. The first question in this section asks the respondent to identify the turnover generating activities at the category level of the Ateco2025 classification. A related open-ended question allows the respondent to add activities if he/she considers the list displayed to be incomplete. After having indicated all the activities carried out by the enterprise, the respondent has to specify the percentage of turnover for each. The final section gathers information on who has filled the questionnaire and his contacts (e-mail and phone number) and comments on the survey. Even though the questionnaire design and flow are quite simple, the questions need to be personalised hierarchically on the basis of the possible macro-sectors, sectors and activities that the respondent can select.

Picture 5
*The Questionnaire Flowchart*



As mentioned above, the first version of survey questionnaire was tested by means of a cognitive interview and then drastically modified and simplified. The cognitive interview was carried out with regard to a specific sector of activities “Wood Production and Furniture Production”. Thirty-three enterprises classified in this macro-sector were invited to participate but only seven of them accepted the invitation. The interview was a semi-standardised interview conducted via video call, with a planned duration of 45 minutes. Each interview was attended by an interviewer who has guided the conversation, an observer who has noted the verbal and non-verbal reactions of the interviewee during the questionnaire completion, and a survey expert to provide thematic clarifications as needed. The interview started with invitation to the interviewee to share his/her screen and to make all his or her thoughts explicit while answering the questions (think-aloud). The cognitive interview also included in-depth questions (probes) with which to prompt the interviewee on specific aspects of the questionnaire. Each interview was recorded with the agreement of the interviewee. The observer's notes and excerpts from the recordings were coded in a coding scheme for the subsequent interpretative analysis of the interviewee’s difficulties emerged during the questionnaire completion. At the end of each interview, there was a debriefing session in which members of the research team discussed the findings. The main difficulties encountered by interviewees in answering to the draft version of the questionnaire are reported above.

A general annotation concerns the attitude of the interviewees, who often were not concentrated in reading the questionnaire and too hasty in answering the questions. In Section 1, when asked to indicate the macro-sector of activity, interviewees found it difficult to answer because the macro-sectors were too aggregated. In addition, they confused the macro-sector of activity with the final activity that generates their turnover, for example the trade sector. Regards to the definition of the sectors of activity, the interviewees did not consider clear the examples presented in the questionnaire. To help the interviewee, the macro-sectors are more detailed in the final version of the questionnaire and the wording and the definitions of the draft version have been changed considering only essential examples.

Section 2 of the draft version of the questionnaire aimed to identify the main economic activity of the enterprise through a first question, structured as a table, asking interviewee to define the activities carried out by the enterprise and the related turnover. A list of economic activities (at category level of Ateco2025) of macro-sector Wood Production and Furniture Production was shown to the interviewees by the electronic questionnaire. On the basis of the cognitive interview, the interviewees found difficult to answer this question because it was too complex and too long to scroll through. In the final version of the questionnaire this question is split into three consecutive questions asking before the economic sector and then the corresponding activities and the related turnover expressed as percentage of the total. Furthermore, interviewees gave inconsistent answers to the other questions in section 2, which were designed to disambiguate specific and more complex economic activities of the macro-sector “Wood Production and Furniture Production”. These questions were dropped in the final version of the questionnaire.

Section 3 of the draft version of questionnaire was aimed to identify the secondary economic activities carried out by the enterprise. This section did not seem sufficiently clear to the interviewees who did not understand they have to indicate secondary activities carried out in sectors other than “Wood production and processing and furniture manufacturing”. This section was removed from the final version of the questionnaire.

Definitely, the results of the cognitive interview were fundamental in designing the final version of the questionnaire clearer and simpler for respondents.