THE PRELIMINARY RECODING PROCEDURE FROM ATECO 2022 TO ATECO 2025

Francesca Alonzi, Annarita Mancini, Caterina Viviano

Abstract. On 1 January 2025, the revised Italian classification of economic activities, Ateco 2025, entered into force. To implement it for both statistical and administrative purposes, an automated procedure recodes enterprises according to the new scheme. This process relies on a mapping and operational correspondence table that automatically resolves one-to-many cases between Ateco 2022 and Ateco 2025. This research work is intended to describe the process of development of the above-mentioned tool including: i) the application of an automatic matching algorithm that compares the text strings headings and inclusion notes of the two classifications Ateco 2022 and Ateco 2025, ii) the analysis of the SEA Survey of Economic Activities results and iii) the involvement of classification experts. Results have shown that, in the absence of any kind of information describing the economic activity at individual level, the developed tool is very useful to make a preliminary large scale recoding of registers and archives of enterprises maintained by various bodies.

1. Introduction

The Italian classification of economic activities is known as Ateco. Deriving from the European NACE, it is adopted by Istat for statistical purposes especially within the business statistics, i.e. for the compilation and dissemination of official statistical data on enterprises. However, it is also used for non-statistical purposes by several administrative bodies such as Chambers of Commerce and tax authorities; the Ateco classification is also widely used by the Government and local agencies, trade associations and other organisations.

The latest version of the Ateco classification, namely Ateco 2025, represents the national version of the European NACE Rev. 2.1 nomenclature. Economic classifications like NACE and its national counterparts, such as Ateco in Italy, play a fundamental role in organising statistical information. In order to depict the real world of enterprises in a reliable way, such classifications must adapt over time to structural economic changes and emerging technologies. The transition from the previous version, Ateco 2022, to the newest Ateco 2025, reflecting the updated NACE version, has introduced new challenges in ensuring the continuity and comparability of economic data.

Into force since January 1, 2025, Ateco 2025 started to be implemented by Istat and other organisations since April 1, 2025. Known as ‘the implementation date’, April 1, 2025 is therefore not a regulatory constraint for other administrations but a choice deliberately agreed upon by Istat, Chambers of Commerce and tax authorities to undertake in a coordinated way the numerous tasks necessary to guarantee the complete transition to the new classification within their own registers and informative systems. In order to assist such a complex operation, Istat jointly with the Italian Chambers of Commerce have developed an operational correspondence table to massively and automatically reclassify enterprises in statistical and administrative registers according to the new Ateco code.

As stated by Alonzi *et al.* (2024), from a theoretical perspective, correspondence tables provide the most relevant framework for supporting recoding activities. However, in the absence of detailed information on economic activities at the individual level — whether from statistical surveys or administrative sources — several methods can be developed to assign new codes on a group basis. The method based on a single code can only be applied to one-to-one (1:1) or many-to-one (N:1) changes from the old to the new classification; that is, using simple correspondences, a classification at the lowest aggregation level is directly recoded to the revised classification. In these cases, the corresponding new code is the only eligible option. In all other cases, further methods are needed. To this aim, the new tool developed by Istat was specifically designed to answer the following research question: *Which methods can be used to recode enterprises involved in one-to-many links from the former classification to the new one without detailed information on the specific activities of each enterprise?* To the best of our knowledge, the developed tool is one of a kind and offers a range of potential applications. Moreover, unlike a standard technical report, it is provided in a user-friendly format that can be integrated into any program or information system with minimal adjustments.

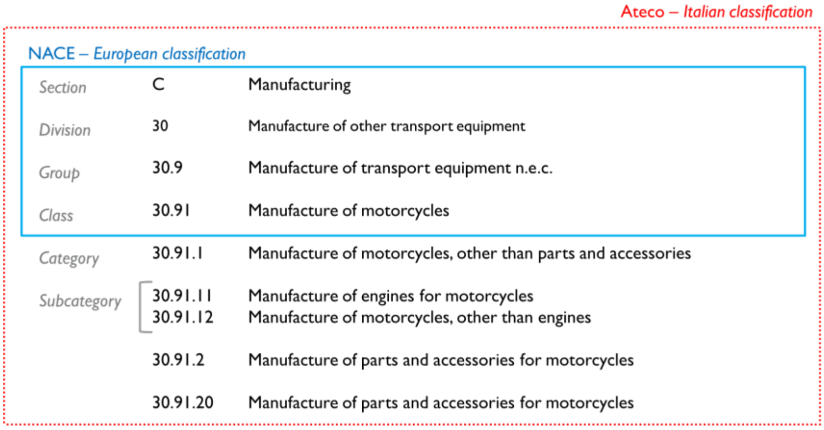
The content of this work is divided into five sections. Apart from the first section that is devoted to the introduction, the second one is intended to provide a background on the Ateco classification. All data and methods used in this study are presented in the third section while the major results are described in the fourth section. Finally, the last section contains the main conclusive remarks on the issues dealt with in this study presenting and critically evaluating the output.

2. Background

2.1. The structure and contents of the Ateco classification

Ateco is a hierarchical classification system aligned with the European NACE at the first four levels, and extended to additional levels (fifth and sixth) to provide a more detailed representation of the national context. Structurally, it consists of codes and headings (categories in the broad sense). Each Ateco code represents an economic activity that is shortly described by its heading. In this way, Ateco codes, together with their headings, provide a standardised definition of all economic activities carried out by enterprises, enabling the grouping of enterprises with similar characteristics for statistical, administrative, fiscal, and other purposes. Figure 1 illustrates an example of Ateco codes as well as of the relationship between NACE and Ateco classifications.

Figure 1 – *The structure of the Ateco classification: an example.*



Ateco – *Italian classification*

NACE – *European classification*

Section	C	Manufacturing
Division	30	Manufacture of other transport equipment
Group	30.9	Manufacture of transport equipment n.e.c.
Class	30.91	Manufacture of motorcycles
Category	30.91.1	Manufacture of motorcycles, other than parts and accessories
Subcategory	30.91.11 30.91.12	Manufacture of engines for motorcycles Manufacture of motorcycles, other than engines
	30.91.2	Manufacture of parts and accessories for motorcycles
	30.91.20	Manufacture of parts and accessories for motorcycles

Nevertheless, while a short title may be sufficient to describe simple and common economic activities (e.g., growing of rice), it generally does not meet the clarity standards required for a statistical classification. For this reason, the structure (codes and headings) of the Ateco classification are enriched by definitional descriptions (explanatory notes) that provide supporting information about the classification codes and headings. They are statements which clearly define the category or they may assist users in determining the boundaries of the category. Explanatory notes may explain the content by giving examples of inclusions and exclusions or provide rules or guidelines for how to use that category. Explanatory notes are optional but are usually included in the classification when further definition of categories is required.

2.2. The theoretical correspondence table between Ateco 2022 and Ateco 2025

A correspondence provides a link between different versions of a classification or between different classifications. A correspondence details how a category in one classification relates, or links, to the new/other classification (Hancock, 2013).

To facilitate the full implementation of the new classification within statistical, administrative or other types of processes, correspondence relationships between the new Ateco 2025 classification and its previous version (Ateco 2022) have also been made available. The correspondence relationships are presented in the form of a table, known as correspondence, conversion or mapping table, in which each Ateco 2025 code is linked with at least one Ateco 2022 code (Alonzi and Viviano, 2025). This is therefore a theoretical mapping; for instance, when an old category splits into several categories in the new classification, all possible links are provided. Thus, the correspondence table is designed to support the recoding of enterprises (and other statistical units) when transitioning between the previous and the new classification.

A total of 6,678 correspondence relationships were identified across the various hierarchical levels of the Ateco classification. Each correspondence relationship defines a link between a pair of Ateco codes: one Ateco 2025 code and one Ateco 2022 code. Table 1 presents the theoretical correspondence links between Ateco 2022 and Ateco 2025 across hierarchical levels, along with the number of codes in each classification at each level.

Table 1 – *Theoretical correspondence links between Ateco 2022 and Ateco 2025*

Hierarchical level	No. of Ateco 2022 codes	No. of Ateco 2025 codes	Correspondence links
1 Section	21	22	80
2 Division	88	87	251
3 Group	272	287	566
4 Class	615	651	1,171
5 Category	920	920	1,883
6 Subcategory	1,241	1,290	2,727
Total	3,157	3,257	6,678

The correspondence links are of different types:

- 1:1 (one-to-one): one Ateco 2022 corresponds to only one Ateco 2025, and vice versa;
- N:1 (many-to-one): many Ateco 2022 codes merging into a single Ateco 2025 code;
- 1:M (one-to-many): one Ateco 2022 code splitting into several Ateco 2025 codes;

- N:M (many-to-many): many Ateco 2022 codes linked to many Ateco 2025 within complex structural changes.

One-to-one and many-to-one cases can be seen as simple matches: in such cases, an automatic recoding of Ateco 2022 codes can be applied straightforward. However, the most frequent types of relationships (representing more than 70 percent of all correspondences) are of many-to-many type, which, along with one-to-many relationships, is considered complex for reclassification purposes. In these cases, the recoding is possible only if specific information is available, i.e. information describing the economic activity carried out by the statistical unit (information at the micro level). Especially in the early phase of recoding, this step is impractical for direct operational use when ambiguous correspondences exist, no specific information is available or when there is the need to have a tool to recode a large number of units in a short time.

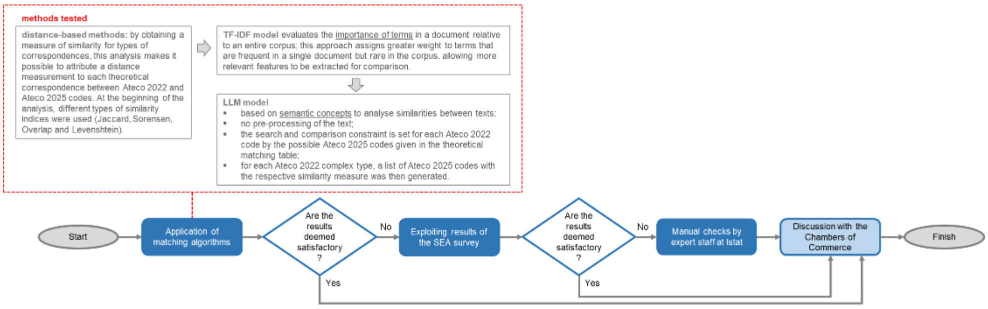
3. Developing the operational correspondence table: methods

The theoretical correspondence table can be easily applied to 1:1 links or to complex links when information at the micro level, i.e. referring to each single unit, is available. In order to solve in an automatic way the cases of complex matches transforming them in simple matches (1:1), Istat developed an operational correspondence table, which assigns a single predominant Ateco 2025 code to each Ateco 2022 code, ensuring consistency for automated reclassification across large statistical and administrative registries. The logic underlying the operational table is to associate each Ateco 2022 code with a single, most representative, Ateco 2025 code among those proposed by the theoretical correspondence table. The choice of the Ateco 2025 code to be uniquely associated to the starting Ateco 2022 is the result of an integrated methodology combining in cascade the following three components (Alonzi *et al.*, 2025).

1. Using an automatic matching algorithm that compares headings and inclusion notes (text strings) of the two classifications, Ateco 2022 and Ateco 2025. This algorithm assigns a similarity measure to each of the M possible Ateco 2025 linked to each Ateco 2022 code, resulting from the comparison of the words into which the set of strings (headings and explanatory notes) is broken down.
2. Using the results of the Survey of Economic Activities (SEA) — also known as Ateco survey — to support automatic unsatisfactory choices based on similarity scores resulted from the above algorithm.
3. Applying manual checks when results are still unsatisfactory.

Figure 2 illustrates the main steps of the methodology.

Figure 2 – *The methodology: main steps.*


methods tested
distance-based methods: by obtaining a measure of similarity for types of correspondences, this analysis makes it possible to attribute a distance measurement to each theoretical correspondence between Ateco 2022 and Ateco 2025 codes. At the beginning of the analysis, different types of similarity indices were used (Jaccard, Sorensen, Overlap and Levenshtein).
TF-IDF model evaluates the importance of terms in a document relative to an entire corpus; this approach assigns greater weight to terms that are frequent in a single document but rare in the corpus, allowing more relevant features to be extracted for comparison.
LLM model
▪ based on semantic concepts to analyse similarities between texts;
▪ no pre-processing of the text;
▪ the search and comparison constraint is set for each Ateco 2022 code by the possible Ateco 2025 codes given in the theoretical matching table;
▪ for each Ateco 2022 complex type, a list of Ateco 2025 codes with the respective similarity measure was then generated.
Start
Application of matching algorithms
Are the results deemed satisfactory ?
No
Yes
Exploiting results of the SEA survey
Are the results deemed satisfactory ?
No
Yes
Manual checks by expert staff at Istat
Discussion with the Chambers of Commerce
Finish

3.1. *The automatic matching algorithm to compare headings and inclusion notes of Ateco 2022 and Ateco 2025*

The innovative aspect of this procedure lies in its use of automatic matching methods based on textual analysis[1], designed to simplify a complex task that would be excessively burdensome if performed manually. The initial aim of the algorithm was to associate each Ateco 2022 code with all Ateco 2025 codes exhibiting similar informational content, based on an analysis of the headings and explanatory notes of both classifications. Each identified pair was assigned a similarity score, guiding the selection toward the pair with the highest score while respecting the constraints of the theoretical correspondence table[2] (Rajaraman and Ullman, 2011).

Several methods were tested to determine the most appropriate similarity measure between the strings of the two classifications. The processing phase was carried out using the Python programming language. At first, a pre-processing of the initial data was necessary, which involved two main steps: i) removal of irrelevant words (stop words): terms not meaningful for content analysis; ii) conversion of words to their lexical root (stemming), to reduce them to their base forms.

Each Ateco 2025 code deemed similar to the source code was associated with a similarity indicator (or score). Different types of similarity indices were tested, for example Jaccard, Sorensen, Overlap and Levenshtein (Cohen *et al.*, 2003; Levenshtein, 1966). However, these distance-based methods presented some critical issues. It was found that the words used in the descriptions of Ateco 2025 did not always match

[1] This step of the study received valuable support from Istat's Directorate for Methodology and Statistical Process Design.

[2] At first the analysis undertaken by Istat focused on 5-digit codes because, for statistical purposes, these are the codes subject to reclassification. But then the analyses were extended to all hierarchical levels of the Ateco respecting the same hierarchical level of the starting code.

those in Ateco 2022, making the previously mentioned similarity indexes ineffective. Furthermore, there have been cases in which the highest level of similarity was associated with a pair not present in the theoretical correspondence table.

To address this problem, a TF-IDF (Term Frequency-Inverse Document Frequency) method was then applied, which assesses the importance of terms in a document relative to the entire corpus. This approach assigns greater weight to terms that are frequent in a single document but rare across the corpus, allowing for the extraction of more relevant features for comparison.

Nevertheless, the TF-IDF method also has some limitations: it does not consider the semantic meaning of words, making it less effective with synonyms or different phrases that express the same concept. Moreover, it assumes independence between terms and focuses solely on frequency, making it difficult to capture the context or relationships between words (Manning *et al*., 2008).

In a later phase, the activity shifted towards models based on semantic concepts to analyse text similarities. Among the most promising approaches, a deep learning algorithm — a Large Language Model (LLM) — was used (Wolf *et al.*, 2020; Devlin *et al.*, 2019). Its main function was to assign a similarity score between texts capable of better capturing semantics, thereby improving upon previous distance-based methods. In this method, no text pre-processing was carried out, and the scope of search and comparison was defined, for each Ateco 2022 code, by the possible Ateco 2025 codes listed in the theoretical correspondence table. The similarity score ranges from 0 to 1: the higher the value, the greater the likelihood that the match is reliable and accurate. For each complex Ateco 2022 code, a list of corresponding Ateco 2025 codes was generated along with their respective similarity scores.

Although the latter method was considered the most reliable, it still presented some critical issues, as it was not able to map all complex correspondences with a high similarity score. For this reason, the selection of the predominant or most representative Ateco 2025 code was finally based on a combination of the results from this automated tool, with the results of the Ateco survey and the manual checks by Business Register and classification experts.

3.2. Using the results of the SEA survey and the contribution of classification experts

Started in April and run until July 31, 2024, Istat has conducted a survey in order to detect the economic activity carried out by enterprises and to use the information collected to reclassify the units registered in the statistical business register (SBR) according to the new Ateco 2025 classification.

To address cases where automatic choices based on similarity scores proved unsatisfactory due to low values, the results of the SEA survey, where available, were taken into account, guiding the selection toward the most frequent empirical responses. In addition, critical codes and ambiguous matches were carefully analysed by expert staff, who manually reviewed the algorithm’s automatic assignments. In particular, the explanatory notes of the two classifications were checked and compared to determine the predominant Ateco 2025.

4. Main results

The operational reclassification table between Ateco 2022 and Ateco 2025 contains 3,157 one-to-one relationships (Table 2), corresponding to the number of codes included in the Ateco 2022 classification; all hierarchical levels have been considered. Each Ateco 2022 codes (source codes) is then linked to only one Ateco 2025 chosen among the M possible Ateco 2025; when the theoretical link is simple (one-to-one or many-to-one), the choice has been straightforward. The chosen code has the same hierarchical level as the source code except for some cases for which it was necessary to choose a code with a higher hierarchical level. In only one case — J Information and communication services — it was not possible to assign any code that is representative of all the thirteen possible choices.

Table 2 – *Operational reclassification table: number of links between Ateco 2022 and Ateco 2025 by type of hierarchical level.*

Hierarchical level	Same hierarchical level	Higher hierarchical level	Total
1 Section	20	1[(1)]	21
2 Division	88	0	88
3 Group	268	4	272
4 Class	594	21	615
5 Category	887	33	920
6 Subcategory	1,113	128	1,241
Total	2,970	187	3,157

[(1)] *J Information and communication services*

The choice of a higher hierarchical level arises from the difficulty in choosing the most representative code among Ateco 2025 codes at the same hierarchical level as Ateco 2022, due to the original activities being distributed among new codes. In such cases, it was necessary to select a code from a higher hierarchical level because it includes all the activities covered by the more detailed lower-level codes. It is the case, for example, of Ateco 2025 code 47.1 — Non-specialised retail sale — that is being repeated several times as a representative code also for several lower-hierarchy Ateco 2022 codes (Table 3).

Table 3 – *Example of the operational table when the chosen representative code 47.1 'Non-specialised retail sale' occurs several times.*

Ateco 2022 code	Ateco 2022 heading	No. of theorical matches	Type of correspond.	Representative Ateco 2025 code
47.1	Retail sale in non-specialised stores	2	N-M	47.1
47.8	Retail sale via stalls and markets	6	N-M	47.1
47.9	Retail trade not in stores, stalls or markets	27	N-M	47.1
47.91	Retail sale via mail order houses or via Internet	32	N-M	47.1
47.91.1	Retail sale via Internet	74	N-M	47.1
47.91.10	Retail sale via Internet	96	N-M	47.1
47.91.2	Retail sale via television	72	N-M	47.1
47.91.20	Retail sale via television	94	N-M	47.1
47.91.3	Retail sale via mail order houses, radio, telephone	73	N-M	47.1
47.91.30	Retail sale via mail order houses, radio, telephone	95	N-M	47.1
47.99	Other retail sale not in stores, stalls or markets	54	N-M	47.1
47.99.1	Retail sale through door-to-door sale agents	95	N-M	47.1
47.99.10	Retail sale through door-to-door sale agents	118	N-M	47.1
47.99.2	Retail sale through vending machines	22	N-M	47.1
47.99.20	Retail sale through vending machines	25	N-M	47.1

The advantages of such a table are clear, as its use enables large-scale automation of conversion operations. However, some critical issues also arise that deserve attention. By design, the selection of a single code, although considered the most representative, is suitable in most cases, offering an effective and standardised solution. Nonetheless, it tends to sacrifice heterogeneity, failing to represent the specificities of economic activities considered less frequent or representative, which may have been split from the original Ateco code. It is therefore worth noting that some Ateco 2025 codes are never selected as representative codes, while other codes are chosen many times (Table 4). Moreover, the adoption of a representative code showing a lower hierarchical level than the one required by the recoding operations (e.g. in the SBR the Ateco code is implemented at 5 digits), necessitates the development of supplementary methods for completing the code to the required digits.

Table 4 – *Operational reclassification table: number of Ateco 2025 chosen repeatedly and not represented.*

Hierarchical level	No. of Ateco 2025 codes chosen repeatedly	No. Ateco 2025 codes not represented	Total
1 Section	20	2	22
2 Division	86	1	87
3 Group	260	27	287
4 Class	564	87	651
5 Category	804	116	920
6 Subcategory	979	311	1,290
Total	2,713	544	3,257

As expected, the percentage of Ateco 2025 codes never chosen increases as the granularity of the classification structure increases: it reaches a 24 percent rate of codes never chosen at lowest level (hierarchy 6). Anyway, there is no doubt about the potential offered by the operational reclassification table, which makes it possible to “squeeze” complex correspondences into simple relationships.

5. Conclusions

Implementing a new classification of economic activities poses several challenges, particularly when it introduces new categories that statistical and administrative registers lack sufficient information to recode units at the micro level. This research work presented the development of an operational correspondence table between Ateco 2022 and Ateco 2025 to solve in an automatic way and at macro level the cases of one-to-many splits by identifying only one Ateco 2025 (the 'predominant' or 'most representative') among M Ateco 2025 associated to the starting Ateco 2022. Above all, it is the result of an automatic matching algorithm that compares headings and inclusion notes (text strings) of the two classifications, Ateco 2022 and Ateco 2025, and assigns a similarity measure. Such a tool has several advantages. The main advantage is its scalability and ability to transform complex matches into simple 1:1 matches thus enabling rapid and standardised recoding. In fact, it is useful for massive recoding operations of economic units representing a quick and standardised solution when information at the micro level is not available. This is exactly the case of both statistical and administrative registers that have applied this tool in the early phase of recoding.

The operational correspondence table has been used not only by its developers (Istat and the Italian Chambers of Commerce) but also by other administrative bodies, making it a significant step forward in the modernisation of statistical and administrative practices and a best practice in inter-institutional cooperation within the public administration. However, users should be aware of this tool's limitations: possible information loss, especially for underrepresented activities; non-selection of some Ateco 2025 codes, leading to sectoral underestimation (e.g. intermediation service activities); reduced granularity from higher-level coding; and potential statistical error due to the adoption of probabilistic methods and practical choices.

All the above considered, the current implementation of this tool marks a first step in the recoding activities of enterprises. Over time, direct updates from enterprises themselves will refine the classification. Future improvements may involve retraining models, incorporating business feedback, and adapting the operational table to emerging economic realities. Anyway, the Ateco 2025 recoding process demonstrates the potential of combining machine learning with expert insight. The hybrid method presented in this research work offers a practical, scalable solution for updating economic classifications, though continuous refinement will be essential for its long-term success.

Acknowledgements

The authors would like to thank their Istat colleagues Alessandra Fiori, Giulio Massacci and Arianna Speranza for their valuable support in the developing of the operational correspondence table providing technical solutions and undertaking manual checks. We would also like to extend special thanks to all colleagues working for the Italian Chambers of Commerce who have heavily contributed to the development of the first draft of the table and then reviewed its final version before the official dissemination.

References

ALONZI F., CONSALVI M., VIVIANO C. 2024. *A Comprehensive Strategy for Implementing NACE Rev. 2.1 in the Italian Statistical Business Register*. Meeting of the Group of Experts on Business Registers Organised by UNECE
https://unece.org/statistics/events/meeting-group-experts-business-registers-0

ALONZI F., MANCINI A., SPERANZA A., VIVIANO C. 2025. *La tabella operativa di riclassificazione da ATECO 2007 aggiornamento 2022 a ATECO 2025*.
https://www.istat.it/classificazione/ateco-2025/

ALONZI F., VIVIANO C. 2025. *Le relazioni di corrispondenza tra le classificazioni delle attività economiche ATECO 2025 e ATECO 2007 aggiornamento 2022*.
https://www.istat.it/classificazione/documenti-ateco/

COHEN W. W., RAVIKUMAR P., & FIENBERG S. E**.** 2003. *A Comparison of String Distance Metrics for Name-Matching Task* Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)

DEVLIN J., ET AL. 2019 *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* NAACL https://arxiv.org/abs/1810.04805

HANCOCK A. 2013. *Best Practice Guidelines for Developing International Statistical Classifications*
https://unstats.un.org/unsd/classifications/bestpractices/Best_practice_Nov_2013.pdf

LEVENSHTEIN V. I. 1966. *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady

MANNING G., RAGHAVAN P. , SCHÜTZE H, 2008 *Introduction to Information Retrieval*, Cambridge University Press

RAJARAMAN A., ULLMAN J. D. 2011. *Mining of Massive Datasets* Cambridge University Press

WOLF T. ET AL. 2020. *Transformers: State-of-the-Art Natural Language Processing* EMNLP

Francesca ALONZI, Istat - Italian National Institute of Statistics, francesca.alonzi@istat.it
Annarita MANCINI, Istat - Italian National Institute of Statistics, annarita.mancini@istat.it
Caterina VIVIANO, Istat - Italian National Institute of Statistics, caterina.viviano@istat.it