Nordic Approaches to AI-Supported Economic Activity Classification after NACE Rev. 2.1

Birgit Nielsen, Statistics Denmark, BGN@dst.dk

Sarah Valentina Møllller, Statistics Denmark, SVM@dst.dk

Tommi Veistämö, Statistics Finland, Tommi.Veistamo@stat.fi

Mari Rantanen, Statistics Finland, Mari.Rantanen@stat.fi

Anders Renstrand, Statistics Sweden, Anders.Renstrand@scb.se

Anders Adolfsson, Statistics Sweden, Anders.Adolfsson@scb.se

Abstract

Following the transition from NACE Rev. 2 to NACE Rev. 2.1, Denmark, Sweden and Finland are intensifying the use of new methods, e.g. artificial intelligence (AI) and machine learning (ML). The aim is to identify methods, which support automated classification, prioritization of manual reviews, and risk-based quality assurance, and thereby ensuring accurate and timely economic activity classification of enterprises in their National Statistical Business Registers (NSBRs).

This paper highlights the experiences and lessons learned across the borders with the aim to foster methodological alignment and knowledge sharing. The paper presents ways in which the various National Statistical Institutes (NSIs) apply these methods in their work, including the recent recoding of their NSBRs from NACE Rev. 2 to NACE Rev. 2.1 during which some new methods were tested, e.g. web scraping. Since then, more advanced methods will be tested on the maintenance of dual coding, classification of new units and quality assurance of the activity codes of existing units. In addition, the paper will outline how new data sources can be integrated into this work, including the use of unstructured text data from company websites and companies' financial reports.

Furthermore, Eurostat is creating a keyword dictionary (index entries) for NACE Rev. 2.1, presenting a structured text dataset for AI and ML purposes. This is particularly relevant considering a future in which updates to NACE are likely to occur more frequently, thereby increasing the demands on maintaining dual coding to safeguard time series data.

Keywords: NACE Rev. 2.1, classification, economic activity, business statistics, machine learning, quality

1. From NACE Rev. 2 to NACE Rev. 2.1

1.1 Data sources

Denmark, Sweden and Finland used a wide range of data sources to support the recoding of their NSBR units from NACE Rev. 2 to NACE Rev. 2.1. Overall, the sources can be grouped into register data, survey or self-service information, business-provided descriptions, annual reports, website text, correspondence tables and expert/manual coding. These are illustrated in figure 1 below and explained in the text below.

Figure 1: Data sources used in the recoding across the Nordic countries

	Denmark	Sweden	Finland
Register data	Core source	Core source	Core source
Surveys / self-service	Used	Core source	Used
Activity descriptions	Core source	Core source	Core source
Annual reports	Core source	Core source	Limited / not systematic
Website text	Core source	Limited / not systematic	Limited / not systematic
Correspondence tables	Core source	Core source	Core source
Core source	Core source	Core source	Core source

Core source: Central and systematic in practice
Used: Regularly used, but not the primary source
Limited / not systematic: Used occasionally or not in a systematic way

In Denmark, relevant statistical data from specific domains, e.g. retail statistics, industrial statistics and the Large Case Unit (LCU), were used to recode selected populations. For other units, data from administrative registers (external) were used, including registers on tattoo establishments, authorised technical activities, embassies and utility companies etc.. Denmark also conducted a representative survey of legal units affected by split codes, and the results were used to calculate a conversion matrix for the recoding process. Additional sources included company names, secondary names, purpose descriptions from annual accounts and company website text.

In Sweden, an e-service on the governmental portal verksamt.se allowed legal units to recode themselves based on the correspondence table between SNI 2007 and SNI 2025. For most units, Statistics Sweden used a multi-step algorithm, including LLM-based methods, mainly relying on registered activity descriptions and annual reports. Manual recoding by staff and experts also used business descriptions and annual report information, and the results were used to train the ML algorithm. Existing information from the business register, particularly business entity type, was also used in rule-based coding where relevant.

In Finland, the recoding used known old NACE codes, correspondence tables and, for some units, new NACE codes acquired from surveys or chosen by AI-based methods. Multi-establishment legal units were surveyed using suggested TOL 2025 classes derived from the correspondence table. Finland also used external and specialised sources, including data from the Natural Resources Institute Finland for agriculture and Accommodation Statistics for accommodation activities. For legal units, activity descriptions from the Finnish Patent and Registration Office's Virre service were an important source. These descriptions are free-text statements provided by the businesses themselves.

1.2 Methods

The countries used several different methods to recode their NSBR units from NACE Rev. 2 to NACE Rev. 2.1. Across Denmark, Sweden and Finland, the recoding was not based on one single method, but on combinations of rule-based coding, machine learning, large language models, survey input, self-recoding and expert validation.

In Denmark, the recoding was mainly based on a deterministic, dictionary-based approach. The existing DB07 code was used as the starting point, and only DB25 codes included in the relevant correspondence table were considered. Textual information from annual accounts, enterprise names, secondary names and website text was then analysed using regular expressions and NLP methods. The process followed a sequential decision chain, where different information sources were used in a predefined hierarchy. For units that could not be resolved directly, survey results and manual recoding were compared in a conversion matrix, which supported a probability-based assignment of the final DB25 code.

In Sweden, self-recoding through an e-service was used where possible, but this was not considered realistic for all enterprises. Statistics Sweden therefore developed a multi-step recoding algorithm for the majority of cases. The process began with rule-based coding for clear cases, followed by an unsupervised text-based workflow combining LLM support and human oversight to match business descriptions to activity definitions. Non-digital financial reports were also introduced as an additional text source. For unresolved cases, a Random Forest model was used to impute codes based on previously coded units and explanatory variables from the Business Register and employee education data. The final implementation combined algorithmic coding, enterprise self-recoding and manual coding by experts.

In Finland, two main AI-based methods were tested: an R-based supervised learning model and a large language model. The supervised model used NLP methods to process activity descriptions and a gradient boosting model to identify similarities between descriptions and NACE codes. The model was designed to be low-cost, controlled by the SBR, and able to assess whether individual units could be classified with sufficient accuracy. The LLM approach used Azure OpenAI to compare natural-language activity descriptions with the NACE classification. Unlike the supervised model, it did not require prior examples of similar classified cases and could therefore be applied where observations were limited. The LLM combined several techniques, including GPT, TF-IDF, embeddings and voting mechanisms, and both the LLM and R-based model achieved approximately 72 per cent accuracy.

2. Maintenance of dual coding

2.1 Data sources

In Denmark and Sweden, dual coding is maintained automatically for the years 2024-2028/2029.

In Finland, dual coding is maintained for selected units for the years 2024–2030. This maintenance is based on several data sources, including NSI surveys, data from the Tax Administration, and other available corporate information. Additional inputs may come from cooperation with LCU colleagues on enterprise groups, as well as media monitoring of corporate news.

2.2 Methods

In all three countries, dual coding is a largely automated process, but with different scopes, unit levels and degrees of expert supervision.

In Denmark and Sweden, dual coding is carried out automatically in the NSBR at micro level. When a unit changes activity code through the online government portal for business registration, the NSBR assigns a corresponding DB07/SNI 2007 code using a one-to-one conversion matrix between DB25/SNI 2025 and DB07/SNI 2007. This matrix is based on the most frequently observed relationship identified during the recoding process. In Denmark, LCUs and FGPs cannot change activity codes through self-registration and must instead contact Statistics Denmark directly to ensure quality and consistency.

Denmark has also carried out historical dual coding back to 2018 at micro level within each enterprise. All unique DB07 codes across local units in the same enterprise are translated into DB25 using the conversion matrix, and the translated codes are then reassigned to the relevant local units. This ensures consistency across units within the same enterprise.

Sweden has also carried out historical dual coding back in time at micro level. However, for selected units, particularly LCUs, FGPs and statistically significant enterprises, the conversion back to SNI 2007 is supervised by staff and experts to safeguard data quality and statistical consistency.

In Finland, each establishment has been genuinely dual coded for one statistical year, either 2024 or 2025. From year 2025 onwards, until 2030, all units will remain dual coded but based only on TOL 2025 code which will then be translated back to TOL 2008 using the conversion matrix. Some units, mainly the largest multi-establishment legal units, are surveyed annually. The activity codes for legal units, enterprises and enterprise groups are derived automatically from establishment level data on activity, persons employed and average value-added per person in each TOL 2025 class.

Finnish dual coding is maintained through parallel recording of the new TOL 2025 code and the old TOL 2008 code. Conversion from TOL 2025 to TOL 2008 is automated and, as in Denmark and Sweden, it selects the most common old class in cases of splits. The same process is used when activity changes are identified through sources such as surveys, Tax Administration data, LCU cooperation or media monitoring.

3. Classification of new units

3.1 Data sources

For the classification of new units, Denmark and Sweden primarily rely on self-registration through online governmental portals. When a new unit is registered, the enterprise selects its own activity code according to NACE Rev. 2.1.

Finland applies a similar approach, combining self-reported information with automated conversion. In addition, Finland uses the textual description of the unit's activity provided by the registering enterprise through Virre, the Finnish Patent and Registration Office's data service.

3.2 Methods

For the classification of new units, Denmark and Sweden apply broadly similar methods. New units register through national online portals, where they select an activity code according to the new classification: DB25 in Denmark through virk.dk and SNI 2025 in Sweden through verksamt.se. Once the new code is received by the NSBR, the corresponding old NACE Rev. 2 code is assigned automatically at micro level using a one-to-one conversion matrix. In both countries, the matrix is based on the most frequently observed relationship between old and new codes identified during the recoding process.

Finland also applies automatic conversion where a TOL 2025 code is received from the Tax Administration. The TOL 2025 code is then converted into a TOL 2008 code through an automated process. However, for some new units, no TOL 2025 code is available because the unit is not required to register with the Tax Administration or has not yet done so. In these cases, Finland may instead rely on the free-text activity description provided by the unit through Virre, the Finnish Patent and Registration Office's data service.

To classify such units, Finland plans to use both a supervised learning model and a large language model. The supervised model requires high-quality training data of a large volume. The LLM approach does not require training data in the same way and can classify units based on their textual activity descriptions, but its performance depends heavily on the quality and specificity of those descriptions.

4. Quality assurance of existing units

4.1 Data sources

Statistics Denmark is participating in Eurostat's grant project on OBEC (Online Business Enterprise Characteristics), under which new data sources are to be used to train a machine learning model for the classification of economic activity. These data sources include statements of purpose in annual accounts and text extracted from enterprise websites. Statistics Denmark enters into a data partnership agreement with the national top-level domain registry (nTLD), under which Statistics Denmark is granted access to all registered .dk domains.

In Finland, quality assurance of existing units is not carried out systematically. It is mainly done manually by experts and based on case-by-case assessment.

4.2 Methods

For the quality assurance of existing units, Denmark and Sweden plan to test different classification methods based on new data sources, including large language models and machine learning models. In Denmark, this work is carried out as part of the OBEC project, which runs until the end of May 2027.

Finland currently applies a more targeted and manual approach. Medium-sized and large units that have an impact on aggregated statistics are quality-checked by the relevant statistical domains and the LCU. If a possible incorrect activity code is identified, the case is reported to the SBR, where experts investigate and update the code if needed.

5. Lessons learned and the next steps

5.1 Data sources

Eurostat are developing Index Entries, which in practice function as a structured keyword dictionary linked to individual NACE codes. These index entries contain examples of activities, products and services associated with a given code and can therefore serve as a valuable input for training AI and machine learning models. By linking words, concepts, and descriptions to the correct NACE codes, the index entries can be used to build training data that helps models recognise patterns in business purpose statements, website texts, and other unstructured sources. They can thus support supervised model training, the development of classification rules, and the validation and quality assurance of automated NACE coding.

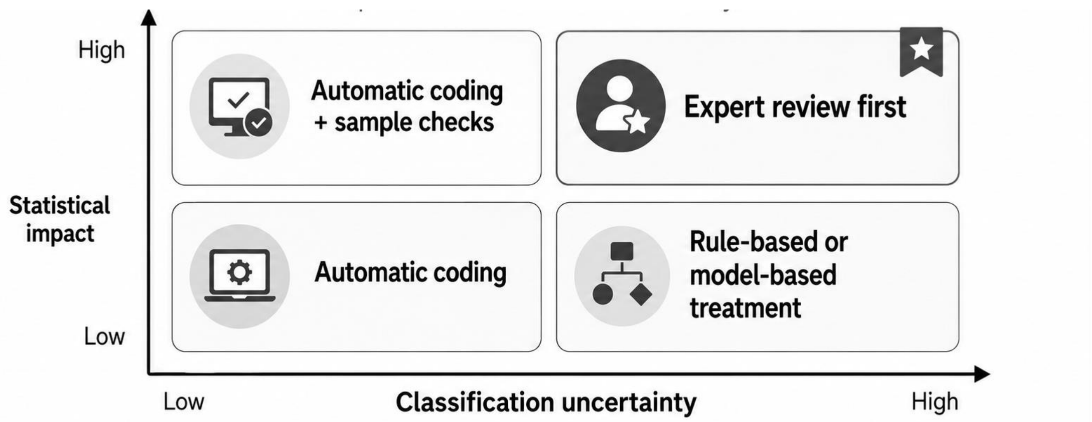
The experiences from Denmark, Sweden and Finland show that the combination of different data sources is highly valuable, particularly because some variables are missing for certain units, while being available for others. In addition, the quality and quantity of available structured and unstructured data and text vary considerably across units, e.g. by size, activity, organisational setup, etc.

The experiences also show that although unstructured text data is inherently difficult to process, it often contains valuable information required to classify a unit. At the same time, it may in some cases be insufficient as a standalone data source.

5.2 Methods

The experiences from Denmark, Sweden and Finland show that ML and AI can be valuable tools in the recoding of SBR units to NACE Rev. 2.1, but they should be seen as supplements to, rather than replacements for, controlled classification processes. The classification process and whether to use automated methods should be based on a risk and quality assessment as illustrated in the figure below.

Figure 2: Risk-based quality assessment of methods



A key learning point is that no single method is sufficient across all parts of NACE. Classification challenges vary depending on the activity, the available data, the size of the population and the complexity of the classification rules. Issues such as small populations, vague or missing activity descriptions, multi-establishment enterprises, vertical and horizontal integration, and FGPs limit the extent to which automated models can be applied consistently. Thus, the different methods all have their advantages and disadvantages as illustrated in the figure below.

Figure 3: Comparison of methods and properties

Method	Requires training data	Transparency	Scalability	Suitable for small populations	Requires expert review	Cost / Price	Reproducibility
Correspondence table	Low	High	High	High	High	High	High
Dictionary (lookup)	Low	High	Medium	High	High	High	High
Supervised machine learning	High	Low	High	Low	Medium	Medium	High
Large language models (LLM)	Low	Low	High	Medium	High	Low	Medium
Manual expert coding	Low	Medium	Low	High	High	Medium	Medium

High   Medium   Low

For almost all methods, the quality of the input data is crucial. ML models perform best where businesses have reliable, registered activity descriptions, or where information is available from e-services, surveys, websites or other structured and unstructured sources. Manual coding remains important, but should be used strategically: for example, to train models, develop keywords for split codes and validate the results for statistically important units.

The countries' experiences also show the value of combining several methods. ML and AI may be useful for some parts of NACE, while rule-based approaches, probability weights, legal form and expert assessment may be more appropriate in other cases. For many splits, especially where the relevant units are difficult to identify, classification resembles a "needle-in-a-haystack" problem and requires targeted methods and often expert review.

LLMs offer new possibilities, particularly where training data is limited, and may be relatively easy and inexpensive to apply. However, they also introduce new risks. Unlike traditional coding procedures, their behaviour is not always fully predictable or controllable. They may produce unexpected results or use information outside the intended sources unless carefully constrained. In addition, it is difficult to foresee how the cost of using these tools will evolve in the future.