

# Classification of business activities by machine learning: The case of France

Wiesbaden group, 2023  
Session 4: Industrial Classification  
Systems



- 1 STATE OF PLAY : HOW WE ASSIGN AN ACTIVITY CODE**
- 2 INSEE'S PLANS FOR NACE REVISION**

# 01

## STATE OF PLAY : HOW WE ASSIGN AN ACTIVITY CODE

---

## THE FRENCH STATISTICAL REGISTER, CALLED SIRUS, IS BASED ON AN ADMINISTRATIVE REGISTER CALLED SIRENE

- Both registers are administered by Insee
- The activity code is an administrative variable (given for statistical purposes) but it is publicly available in the administrative register
- As it is known, it is widely used for other purposes (taxes, social contribution, subsidies, insurance...).

## DIFFERENT PROCESSES

- Thanks to a literal description of the activity of the enterprise during an administrative procedure
  - The most important one : when an enterprise is set up, a form is filled in and a large amount of information is sent to Insee. This includes a literal description of the activity by the firm
  - New forms can be transmitted if the enterprise changes its activity
- Claims : if the enterprises disagree with the activity code by Insee
- Surveys : structural statistical surveys (the so-called ESA, EAP) in which enterprises are asked about the breakdown of their activity



## FROM A LITERAL DESCRIPTION TO A CODE

- Until to last November, we used an automatic label coding system, called Sicore
  - Based on a training file of encoding examples
  - Drawbacks : if the label did not match an encryption example, no code suggestion was returned. It was then coded manually by a human being
- Since last November, we have implemented a new model based on machine learning (FastText)
  - The training sample : 10 million observations coded by Sicore or manually
    - Use of the literal description+auxiliary variables
    - Need for preprocessing : lower case conversion, removal of punctuation, removal of numbers, removal of one-letter words, removal of stop words, stemming...
  - Very accurate even with new literal descriptions that have never been coded before
  - A 100 % result even with a low accuracy rate
  - However we have decided to maintain a manual check if the accuracy rate is not good

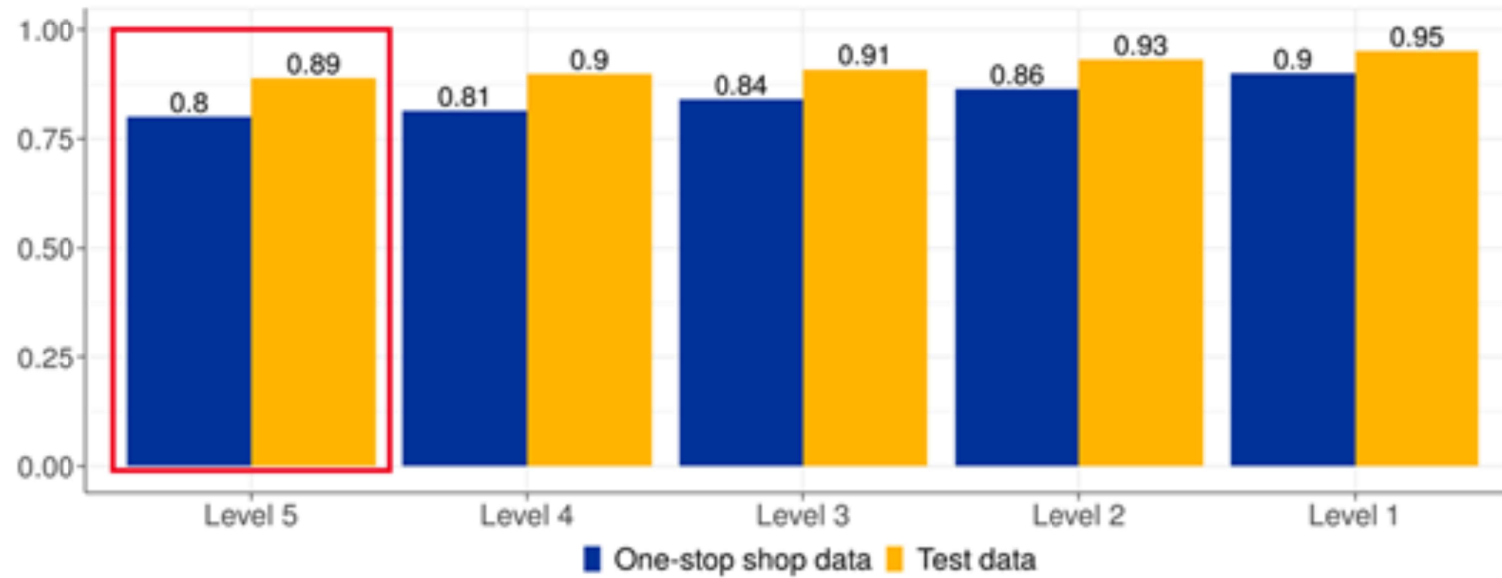


Figure 1: Accuracy for various level of the NACE nomenclature.

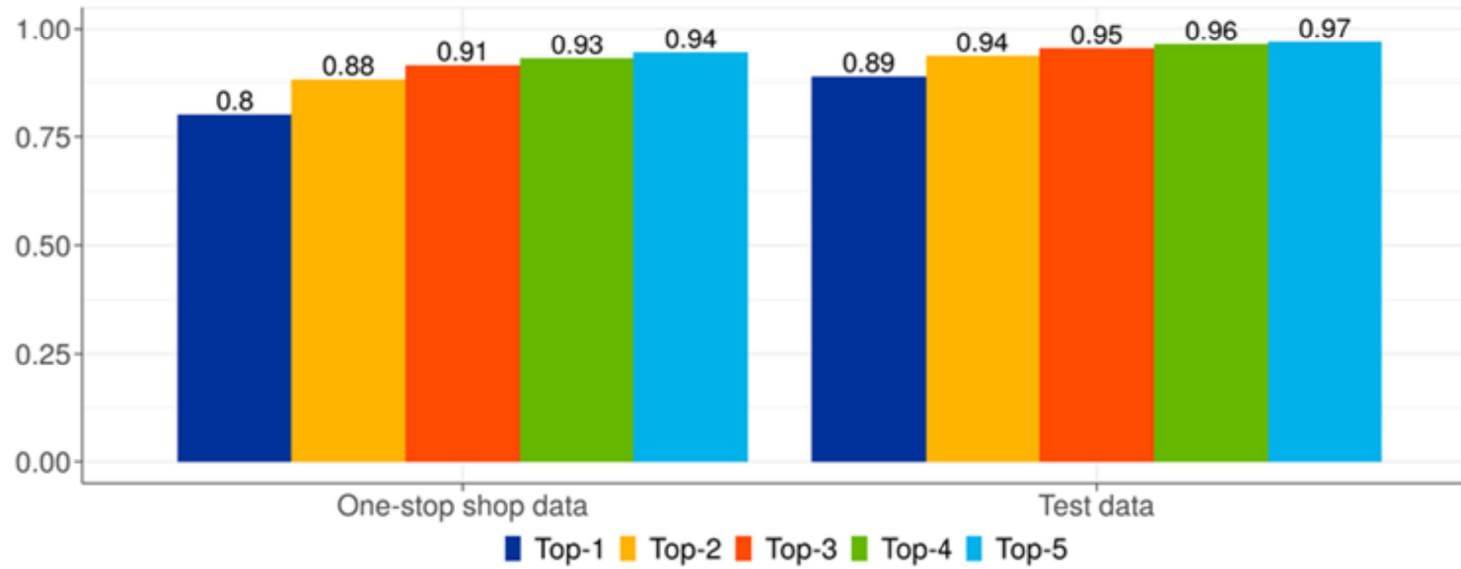


Figure 2: Top- $k$  accuracy per sample.



# 02

## INSEE'S PLANS FOR NACE REVISION

## TWO MAIN ISSUES

- The flow : How to codify the activity of new enterprises according to the new Nace/Naf
- The stock : how to recodify the activity of more than 10 million active legal units of our administrative business register


- **We plan to adapt our machine learning model**
  - Which training sample ? What size ?
  - Do we have enough information in the literal description to identify some new classes (for instance, intermediation services ?)
- **Do we need to codify the new enterprises in both classifications ?**
  - Not too costly if the codification is automatic ;
  - Not possible because of the human cost of human verification for both classifications
  - Consequently, not the same quality for both codes



## DIFFERENT STRATEGIES

- The correspondence table for the subclasses that are unambiguous
  - 4.5 million of legal unit with an ambiguous code
  - (plus 3 million of renters of furnished accommodation, mainly households)
- Structural business survey (ESA/EAP) for the sampled enterprises (mainly the largest):
  - Sometimes the activity is broken down to a more detailed level that allows coding in both classifications
  - We plan to add some new product codes and some new questions
    - For example on intermediation services

## DIFFERENT STRATEGIES

- Use our new trained FastText model on past literal descriptions
    - Will it be useful ?
      - The literal description is not available for all the UL (not kept)
      - The literal description may be old and no longer up to date
      - The literal description may be not detailed enough to assign a new code.
  
  - Are there any databases or registers that can help to recodify the activity ?
  - Ad hoc survey for sensitive activities
  - If no information is available, use a probabilistic model based on the results of structural business survey
  - Disclose the new codes and wait for claims
- 

- We need to recodify the activity code in both the statistical and the administrative BRs
  - Where to start ?
- Some special problems with an administrative BR
  - Fortunately or unfortunately, the activity code is widely used
    - An important communication need
  - Make sure that all the administrations are ready to use the new classification
  - More consequences for the enterprise itself if you assign a inaccurate code
    - Communicate the information about the new code before the official date of the classification change
    - Develop a tool to deal with a massive wave of claims
      - Automatic enough
        - (some issues about authentication)
      - Can you allow the enterprise to choose its new code ?



## Retrouvez-nous sur

[insee.fr](http://insee.fr)



**Marie Leclair**

Head of the repertories, infrastructure and structural statistics  
department

Insee, Business Statistics Directorate

[marie.leclair@insee.fr](mailto:marie.leclair@insee.fr)



Mesurer pour comprendre