Paper for the 2023 Wiebaden Group,
Session 4: Industrial Classification Systems: Treatments for the upcoming NACE or ISIC revision and other industry classification issues

Classify firms activities according to machine learning method: the case of France

Marie Leclair, Insee,

Draft paper, 09/2023

Abstract

Insee manages both a statistical business register (called Sirus) and an administrative register (called Sirene). As the registration authority, it receives all the forms filled in by businesses. These forms contain a short, literal description of the activity of the firm at the time of its birth and, in theory, each time the firm changes or adds a new activity.

Until recently, an automatic label coding system called Sicore was used to assign an activity code to the enterprise. However, this system was based on a training file of coding examples. If the label did not match an encryption example, no code suggestion was returned. This led to a critical situation when a new website was created to complete the formalities: the new administrative way of collecting information slightly modified the way companies declared their activity and in numerous cases the old coding system did not propose any code. Instead of adding new coding examples to the old coding system, which would have been too time-consuming, Insee decided to move to a machine learning model using fastText and trained on the labels coded in the past. This new model proved to be quite efficient, even on the new labels.

The forthcoming change in Nace is more challenging. An adaptation of the new coding system will be needed, as well as a new learning base. However, the use of the label already collected may not be sufficient to code the activity of all the companies in our register in the new Nace. We will complete our strategy by using our annual structural survey, which will collect information on activities in both classifications, the new and the old. This will ensure that we are quite sure of the new code of the larger enterprises (all of which are interviewed in our survey) and will also provide a probabilistic transition table for all non-bijective codes. For some specific codes it may be necessary to mobilise other data sources. Finally, the new code will be disseminated to all businesses, as it is an item of information in our administrative business register. We expect that there will be challenges from companies that will lead to improvements in our coding.

Introduction

The activity code of the enterprise is a major variable of any statistical business register. It is indeed used in almost all the sampling and is a major criterium in all the business statistics. Therefore, it is needed to be known for all the enterprises and not only for some of them. In France, this statistical code is mainly deduced from the description of its activity given by the enterprise itself when it filled a more general administrative form. The first part of the article describes how this this activity code is codified.

The forthcoming NACE revision requires statistical business registers to be updated first, in order to be able, among other things, to draw samples in the new nomenclature. This is a huge and difficult task, since it requires not only recoding the entire stock of enterprises in the register, but also adapting all the tools needed to code the flow of new enterprises. The second part of the article presents Insee current plans for carrying out this work, at a time when it has not yet been implemented. The aim here is to share experience (or rather projects) on a task that awaits all NSI.

1/ State of play: how to attribute an activity code to each enterprise

The French statistical business register is called Sirus and is used for statistical purposes. In order to collect information on legal units, it relies on an inter-administrative business register called Sirene (*Système informatique pour le répertoire des entreprises et des établissements*), also managed by Insee (since 1973).

Under French commercial law, every business must be registered in SIRENE and report any change in its circumstances (for example, a change of address or the creation of an establishment in a new location). Since 1997, the use of the SIRENE identifier has been compulsory for all administrative declarations and formalities for companies, which not only guarantees the coverage of the register, but also that the information is up to date and accurate. The scope of SIRENE is even wider than that of business statistics, i.e. all enterprises involved in the commercial production of goods or services: since 1983, both public enterprises and administrations have been obliged to register in SIRENE. Associations are also registered if they are employers, whether they apply for subsidies or pay taxes.

This SIRENE register is therefore an essential resource for creating a register for statistical purposes. Most countries have an administrative register, but it is rarely managed by the national statistical institute, as is the case in France. This peculiarity makes it possible to take account of statistical needs: the possibility of improving the quality of addresses, ensuring that the correct activity code is used, etc.

Even if the purpose of the activity code is primarily statistical, it is defined in the administrative business register and this activity code is published in an API, as are all the variables of the Sirene register. Enterprises are therefore very careful about the code assigned to them, as this public information can be used by other administrative or private actors (for example, to determine subsidies, social contributions, insurance premiums, etc.). This "administrative" use and its practical consequences lead many businesses to contest the coding of their main activity by INSEE.

Enterprises are obliged to declare their birth to the SIRENE database (otherwise they do not have a SIRENE ID, a unique and universal identifier required by the administration to carry out any administrative procedure). In this first declaration, companies must describe their main activity. This description is then used by Insee to assign an activity code to the company. How this assignment, from label to code, is done is presented in part 1.1. In part 1.2 we present the way in which this code is updated after the initial declaration of the enterprises.

1.1/ the textual declaration by the firm is codified thanks to a machine learning algorithm, helped by humans

The description of the activity by the enterprise in its original form is used to assign an activity code to each enterprise. This task is partly automated and partly carried out by a human being when the machine is not able to distinguish between two codes with sufficient certainty.

Until recently, an automatic label coding system called Sicore was used to assign an activity code to the company. However, this system was based on a training file of coding examples. If the label did not match an encryption example, no code suggestion was returned. This led to a critical situation when a new formalities website was created: the new administrative way of collecting information slightly changed the way companies declared their activity and, in many cases, the old coding system did not propose any code at all. The automation rate dropped from 60% to 30%. Instead of adding new coding examples to the old coding system, which would have been too time-consuming, Insee decided to move to a machine learning model using fastText.

To do this, it can draw on a database of 10 million observations containing the company's textual description, tagged both by Sicore, the previous automated system, and by hand. In addition to the textual description, other variables are available, such as the type of activity, the surface area (m2) and other information on the administrative forms fulfilled by the company.
The textual description and the values of the auxiliary variables are concatenated. The result is then pre-processed for natural language processing (lower case conversion, removal of punctuation, removal of numbers, removal of one-letter words, removal of stop words, stemming, etc.).
This database was used to train the machine learning model.

The new model proved to be highly effective, even on new labels. Nearly 80% of the labels are coded correctly, and most prediction errors are close to the nomenclature (Figure 1). The correct classification is in the top 5 predictions 94% of the time (Figure 2).

Although the new model suggests an activity code regardless of the literal description (which was not the case with the previous tool, sicore), if the two most likely codes have a probability that is not very different, we send the case to an administrative agent who decides which code to keep, or who may ask the company itself for more details on the company's activity. Initially, we were quite cautious and sent a significant proportion of codes for human review. We knew that the company's enquiries about the activity code might cost more to process later than the human check. Our fears proved to be unfounded.

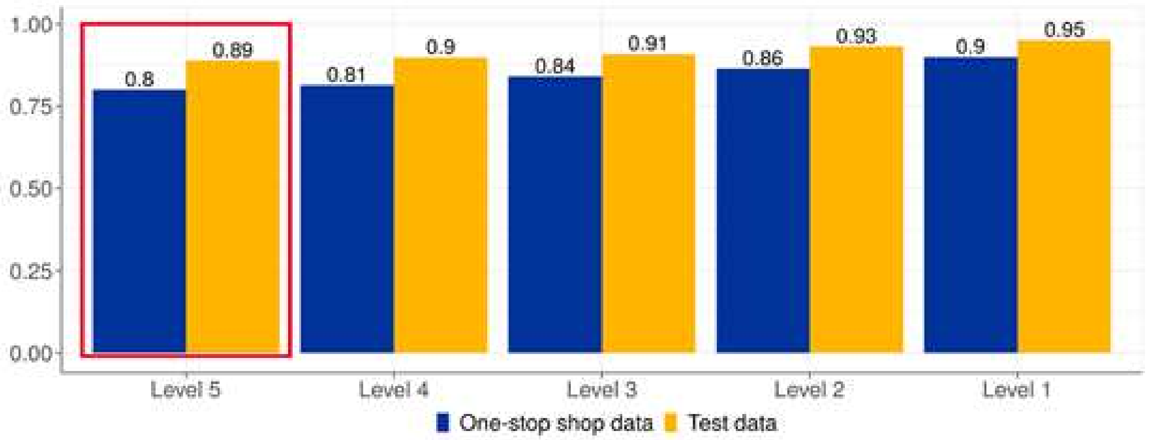

Figure 1: Accuracy for various level of the NACE nomenclature.

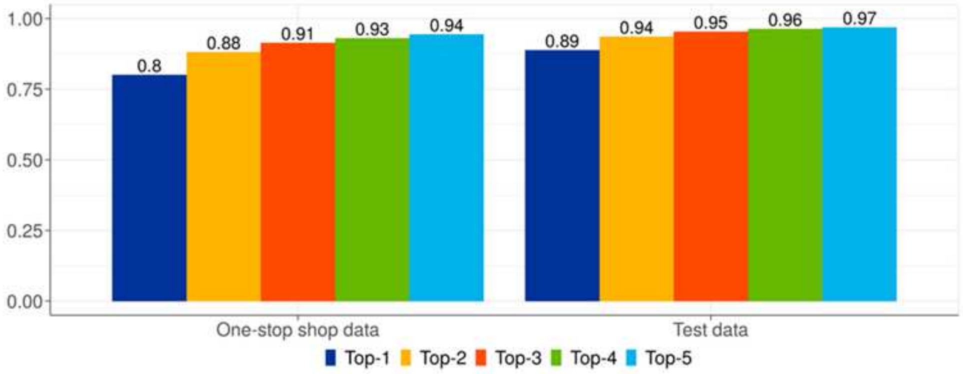

Figure 2: Top-k accuracy per sample.

1.2/ the results may be updated by new declaration of the firms, claims or answer to statistical surveys

Claims: Once the code is assigned, automatically or manually, it is made public through an API and in an administrative document provided by Insee and requested for most administrative procedures. Once the company knows the code, it may not like it for various reasons: because it considers that this code does not reflect its real activity, or because sometimes, even if it shouldn't, this code may have an impact on its subsidies, obligations... In these cases, Insee may receive a request from this company to obtain another code. The company has to describe the distribution of the company's turnover between the different activities and Insee, thanks to these elements, decides the final code, which can be a new code or the same as before. It is important to note that even if the company provides some additional information in the application process, it remains a purely declarative process, which means that Insee never verifies the truth of the company's statements or compares its statements with other administrative documents. In this respect, it is very similar to the response to a statistical survey.

Change in the activity: The activity of a firm is not permanent and a firm whose activity changes must declare it by means of a new administrative form, which is sent to Insee in order to update the activity code. The procedure is then very similar to that described for the creation of a company.

Surveys: Finally, some enterprises are asked about their activity and the breakdown of this activity between different products in Insee surveys: these structural business statistics surveys (ESA, "Enquête Structurelle Annuelle" and EAP, "Enquête Annuelle de Production", for industrial enterprises) can lead to a new activity code. The new code is calculated from the enterprise's reply to the questionnaire. In this case, the administrative business register is updated and the enterprise is informed of its new activity code (which may lead to new entitlements...).

2/ NACE revision: which plan to codify in the new classification all the enterprises?

The final NACE classification and its correspondence table have recently been known and sent by Eurostat. Therefore, what is presented in this article is an Insee plan before further investigation of the implications of the new classification. It should be adapted after a more detailed study of the relevance and practicability of the following leads.

2.1/ The French classification of activity, NAF

The aim is to assign a new code to all active units included in our administrative register, Sirene (legal units and branches/establishments), but also to other statistical units included in our statistical business register, Sirus (mainly enterprises, groups).

The new code should be consistent with NACE, but also with our new national classification, NAF, Nomenclature d'activité française, which is more detailed (and of course consistent with the new NACE1). The establishment of this new NAF was carried out by the CNIS, Conseil National de l'Information et de la Statistique, a body that brings together both statisticians and users of statistics (representatives of businesses, researchers, trade unions, professional agencies, consumer or family associations...), following an Internet consultation. The result of the consultation was analysed and a decision was taken to accept or reject the proposal, depending on its consistency with NACE, the economic weight of the new subclasses proposed, the possibility of easily identifying the activities... In the end, a new NAF was submitted to a new consultation, which is still ongoing. For the time being, there are 750 subclasses in this new NAF (compared to 651 classes in NACE)

The new NAF will be finalised and adopted before the end of 2023.

2.1/ For the flow: adapt the machine learning process

In order to continue to assign an activity code to the new companies registered in Sirene, we have to adapt our tools to the new classification. This means adapting the machine learning model, but also training all the Insee administrative agents who have to manually code the activity and explain the new classification to them.

To adapt the machine learning, we need to get a new training sample. As I mentioned before, the one we built before was based on the whole dataset of textual description and final codes, that is, a database of 10 million observations. This is clearly not our goal for the new training sample we have to build, because we have limited resources to recodify the activity in the new classification, and because it is not useful to have such a large training sample.

Using a correspondence table, we can match the textual description with the new code if the correspondence table is unambiguous (a former NAF code is entirely directed in a single new NAF code). We can also use some of the projects described later in part 2.2 to recodify all the past stock of units. Finally, for the remaining textual descriptions without a new attributed code, we will sample them and ask our administrative agents to codify them according to the new classification.

We plan to use both tools, the current tool, which codes the current code according to Naf rev2, and the new tool (for Naf rev 2.1): in fact, in the statistical business register it is necessary to have both codes for a certain period of time, because not all statistics will change to the new classification at the same time. In particular, short-term statistics will have to be produced in the current classification until 2028. We will not be able to guarantee the same level of quality for both codes as this would be too costly. As far as the coding of the activity of new enterprises is concerned, we plan to use a single automatic tool without human verification.

2.2/ for the stock: different strategies

In order to codify in the new nomenclature the whole stock of enterprises that belong to our business register, we are planning different strategies depending on the class of the nomenclature, the size of the enterprise, the type of statistical units... The objectives are twofold: to give each enterprise the more accurate new code; to find a solution that doesn't require intensive human work. We will then try to develop automated methods.

a\ First of all, for a large part of the nomenclature (even if not so large), the correspondence table is unambiguous: a former NAF code is entirely directed in a single new NAF code. In these cases, recoding is quite simple. We have to refine our calculation (because the national nomenclature is not completely stable), but there remain 4.5 million legal units in our register whose recodification is not clear and for which another solution has to be found.

b\ For the largest companies, we can use surveys. In fact, these enterprises are surveyed every year in order to compile our structural statistics and they are asked to provide a breakdown of their activity according to different products (through the already mentioned surveys, ESA and EAP). These products are more detailed than the current NACE/NAF and in some cases the level of detail may be sufficient for the activity to be reclassified in the future NACE/NAF. But more often this is not the case: we plan to add some product codes where possible (which is not always possible for technical reasons) and also to add questions where the new boundary between different activities is difficult to capture by a product code. For example, we will add a question on intermediation services. The question is rather rough at the moment and we will see if it is well understood and identifiable by enterprises. In some cases, if the number of enterprises is not too important, we will

only collect the information by calling the respondent and asking some qualitative questions (without a formal survey).
The new questions will be included in the 2024 survey to be carried out in 2025. It will also be a test for the 2025 survey, which will provide the results in both classifications, the new and the old, and will be used by National Accounts to produce a correspondence table for turnover. The 2024 survey may also be used to assign a new code using a probabilistic method (see below).

c In some cases, we plan to use our machine learning method (which we will use for the flow of new registrations) to recodify the activity code of the firms we have in our inventory from the textual declaration we have received. However, this strategy has many shortcomings: firstly, we haven't kept the textual declarations for all the companies (we don't have them for those who declared their activity a long time ago without having changed it since); secondly, the detail in the description of the activity may not be enough to distinguish between two codes in the new classification. Normally, when the administration receives a formality and codifies the activity (treatment of the flow), it has the possibility of returning the formality to the enterprise and asking for additional information. This possibility does not exist for the stock of enterprises. Finally, the description of the activity is quite old. Since the last formality, the activity code may have been updated following a claim or a survey. The textual description is no longer accurate. And the information on the new code (survey or claims) cannot be recoded using the machine learning method or any other automated method (for claims, for example, the code can be assigned after many exchanges of mail; the analysis of this long exchange is not feasible, at least not automatically).

d In some cases, it might also be possible to identify data sources that could help in the recodification of the enterprise stock. We have to admit that we haven't identified any such data sources at the moment. Some fiscal sources may help to identify some intermediation activities, but they are not completely conclusive. However, we have planned to meet the different specialists of the different sectors in order to find out if such a source exists, for example, if an activity has to be administratively registered in a specific register. However, this is unlikely to be the case, as users usually ask for new sub-classes to be added to the NAF when they don't have any other way of identifying these new sub-classes.

e/ Ad hoc surveys may be carried out for some subjects which may be sensitive because the correspondence table leads to different codes very far down the classification, or because there is specific interest in the area. However, our capacity to deal with them and the desire to control the burden on businesses will limit the number of such surveys. The areas in which these surveys might be carried out have not yet been determined.

f In all other cases we will end up using a probabilistic model. The ESA/EAP surveys referred to in point (b) will also make it possible, for the smaller enterprises, to construct a correspondence table between the two classifications, with the proportion of enterprises that are reclassified from a previous sub-class to a new one.

g\ What has been described above is our plan for recoding the activity of legal units. But we also have to recodify it for all the other statistical units: establishments, groups, enterprises. In most cases the activity code of the establishment will be derived from the code of the legal unit, but in some cases it won't be possible. Our strategy has not yet been finalised, except that for the largest enterprises we will be able to check directly with them, as we have a SIRENE repertory management service dedicated to them. The activity code of groups and enterprises usually depends on the activity code of the legal units through algorithms that can be applied or collected through surveys.

2.3/ Some considerations of the consequences for the administration and firms: be careful about that.

Insee plans to include the new activity code in the statistical business register in mid-2025 (because it is required by Eurostat and also because it is needed to sample surveys such as ESA and EAP according to the new classification). But this first attempt at a new codification may still be rough (correct on average but not in detail) and, in our opinion, can't be published in our administrative business register, Sirene. In fact, the activity code may have an impact on the administrative situation of the company. Even if this is not the case, the company informed of its new activity code may contest it. These claims will be an important part of the change of classification. To deal with them can be a heavy burden: enterprises are used to contact the Insee hotline for these claims (even if there is an online form for the claim) and this hotline, which is also used for other Insee services, as the respondents of the Insee survey, can be overwhelmed; the analysis of the claims by the administrative agents is also time-consuming. Finally, the disclosure of a too rough reclassification code may also damage Insee's image and credibility in terms of codifying the activity accurately.

Therefore, we want to give ourselves more time to recodify our entire administrative business register and while we will include the new activity code in the statistical business register, Sirus, in mid-2025, we will not publish the information in our administrative register, Sirene, before the end of 2025. This additional time will allow us to perfect the tasks described in 2.2. The additional time will also be used to communicate with businesses and administrations and to prepare tools to deal with claims.

To address this issue, we are thinking on :

(a) Development of a new tool for claims that will facilitate the automation of their processing; currently, the processing of these claims is entirely manual, both because Insee has to check the identity of the claimant (the fact that he is authorised to request a change in the activity code) and because the analysis of the data transmitted by the company (mainly a breakdown of turnover by type of activity) is also done manually. For the first part, the verification of the identity of the applicant, Insee can use some tools managed by Insee (called ProConnect) that authenticate the electronic identity of the director of the company. However, it does not work for all companies (the identity of the director is not always known for companies in particular and it is not always the director who wishes to carry out this formality). For the second part, tools such as Fasttext can easily facilitate the automation of the new codification.

(b) Possibly disclose the information on the new activity code before it is entered in the administrative repertory. This will allow claims and corrections to be made before the new activity code affects the administrative life of the company. The interests of automation are the same as those previously discussed.

Communication and publicity around this new activity code is also an important issue.

- It is important to make companies aware of the change in classification and their new code. Unfortunately, Insee doesn't have the e-mails of all the companies. A personalised information can therefore only be done by post and it is very costly (because we have to inform about 10 million legal entities about their new code). Another option is to carry out a broad communication campaign (using as relays the other administrations in contact with businesses and the various portals of websites specialised in business formalities) and let the businesses look for their new activity code on the Insee website.
- It is also important to communicate with the users of the administrative register. They can be administrations, but also private actors. It is quite easy to inform the users of our Sirene API

because we have a community of users, but they are often IT engineers. From an IT point of view, the change of classification is not an important issue because we have decided not to change the format of the code. It is more difficult to sensitise our end users. In some administrations, the administrative business register is used as part of a wider internal information system. The users of these internal systems are not known to Insee and it is they who have sometimes defined rules based on the activity code. For example, some charges paid by companies for accidents at work are determined on the basis of the activity code, as are some subsidies related to the energy crisis. Insee is not aware of all these uses (which it does not promote). On the contrary, a decree has been issued to prohibit rights and obligations based solely on the activity code provided by Insee) and an important part of the change in classification is to inform the various administrations in advance so that they can adapt.