Scanner data in the French CPI: classifying products using NLP

Keywords: CPI, Scanner data, GTIN description, Classification, Natural language processing

1. INTRODUCTION

Scanner data are data collected by retailers when consumers pay for goods in store. For each sale, the GTIN/EAN, price, quantity as well as a short description of each product bought is recorded. The resulting data, aggregated by outlet and day of sale, are then sent daily to INSEE.

Since January 2020 [1], the Consumer Price Index (CPI) has been calculated using scanner data for processed food, maintenance, personal and home care products from supermarkets and hypermarkets in metropolitan France.

1.1. Scanner data: an additional source of information that does not change the methodology of the French CPI

The introduction of scanner data does not imply any change to the core concepts of the CPI but merely involves the use of a new data source. In particular, the CPI with scanner data remains an annually-chained fixed-basket Laspeyres-type index. The prices of a fixed basket of goods representative of household consumption are monitored on a monthly basis with the aim of measuring price movements at a constant level of quality and structure of consumption. The basket is updated annually to ensure that it is representative of household consumption, and, if products disappear during the year, they are replaced, and a quality adjustment is made.

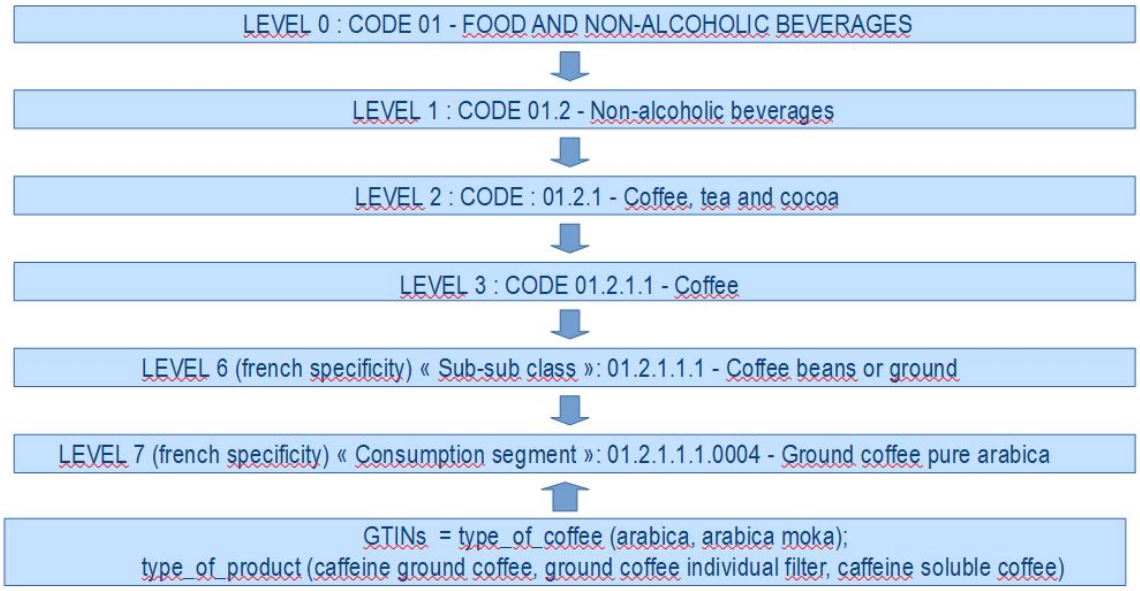
1.2. The necessity of an article repository and classification rules

To build the basket of fixed products in a year, we need to classify scanner data in order to reflect household consumption. For that, Insee buys an article repository from market research companies. This repository describes precisely consumer products (e.g. type of product, brand, packaging, volume, etc.). The classification rules are defined by experts in the relevant consumption sector and are based on these characteristics.

On the other hand, in a fixed basket of goods, when a product disappears, it is replaced to avoid a process of attrition resulting in an increasingly less representative basket over time. The replacement product is selected to be as similar as possible to the product that has disappeared, although a quality adjustment is made if a difference in quality remains. In the case of scanner data, the replacement product is chosen randomly from the products belonging to the same consumption segment and in the same outlet. A quality adjustment is systematically made by overlap method. In other words, the prices of the two products, that of the product and of its replacement, are compared (two months before the disappearance of the product to avoid taking into account, as an indication of lower quality, the fact that products at the end of their life generally see their price fall).

The classification of scanner data and replacements are done automatically via an application using Big Data technologies.

Example: Groud coffee arabica pure



2. METHODS

Classification rules are fundamental to establishing the basket. A precise study of each family of products must be carried out in order to establish these rules to be representative of household consumption.

However, not all items are described precisely in the item repository, especially products sold specifically in the overseas departments due to to their proximity to other countries. A NLP-based (Natural Language Processing) approach can be a solution for these specific GTINS.

2.1. How to define classification rules?

The article repository is divided into 900 product families. About 500 families contain products tracked in scanner data (processed food, maintenance, personal and home care products). Around 350,000 products are tracked in the “scanner data” basket.

To establish the classification rules, the previous turnover of the articles (GTIN/EAN) of each family are studied by experts in the relevant consumption sector. Indeed, scanner data provide objective information about the weight of each consumption segment in the item (which has meant revisiting the importance of certain consumption segments: for example, cotton swabs are no longer monitored because of their very low weight), but also about the weight of each outlet and product.

Over the course of 2020, almost 600 consumption segments were monitored in scanner data included in 108 COICOP[1] (Classification of Individual Consumption by Purpose) items. In terms of expenditure in 2020, the “scanner data” basket thus compiled represents 9% of the CPI basket.

1 The Classification of Individual Consumption by Purpose (COICOP) is one of the "functional" classifications of the National accounts system (SCN). It is used to classify transactions made between producers and the institutional sector of households. It is called functional because it identifies objects or objectives for which these transactions are made. It allows to know the expenses which households dedicate to food, to health, to education etc.

2.2. Products of overseas departments: the necessity to use text mining

Currently, scanner data only covers products sold in mainland France. In the medium term, we would like to integrate data from overseas departments.

In the overseas departments, some products sold are from Asia or the United States (study of numbers at the start of GTINS, 300 to 379 for France for example) due to the geographical proximity or some are made in local factories (eggs, milk...) and identified in different GTIN. Thus, it is estimated that only 75% of the products sold are present in the article repository (those also sold in mainland France).

To fully exploit scanner data, we would like to classify these articles in COICOP. Since we already have a sample of the articles classified using the classification rules, we can train a classification algorithm on the descriptions of the GTINS.

Example of specifics products sold in overseas departments

EAN/GTIN	DESCRIPTION_EAN/GTIN
3176571681003	LAIT ENTIER GRANDLAIT BK 1L
6091091000189	NOUILLE CREVETTE APOLLO 85G
1318145273630	QU CAMARON DEV 13/18 280G PNE
5449000000439	COCA COLA PET 1 5L,1.5L
3490950312815	HLE TOURNE EN MODE CREOLE1LPAL
6091043801741	LIQUIDE VAISSELLE 5L CITRON
3301421003012	FARINE MENAGERE MOULIN BLEU 1K
3377450000018	RHUM BLANC TRADITIONNEL 49D 1L
3275880003206	KOKOT X20 OEUFS FRAIS

3. RESULTS

This section presents the work on the classification algorithm that has been developed for the NACE nomenclature (which contains 129 categories). The approach will be extended to the COICOP in the coming months in order to help classifying products sold in overseas department.

3.1. Classification algorithm

We use the supervised module of fastText [2, 3], which is a one-layer neural network that uses (in particular, but not only) n-grams of characters as tokens and encode product description as an average of its tokens' embeddings. The use of n-grams of characters makes the classification algorithm particularly robust to abbreviations and typos.

We train the model on a sample of almost 5 millions products classified in NACE and obtain a precision of above 96% of a test sample of more than a million product. The model is trained in less than 20 minutes on 5 CPUs.

We develop a measure based on the difference of prediction probabilities for the top-2 categories to quantify the confidence in a product classification, and show that on average the error rate is a decreasing function of this measure.

3.2. Labelling campaign using a web application

A supervised approach requires a labelled sample of products so the algorithm can learn to associate each product's description with a category. Unfortunately, not all products

are indexed in the article repository, and articles that are not indexed might be different from indexed articles.

In order to collect a sample of labelled product descriptions for article that are not indexed, we developed a web-based application that encapsulates the trained model and helps humans to manually classify products. Since human resources are not to be wasted, we offer some thoughts on how to select the sample of products that will be sent for labelling and organize data collection.

3.3. Extension to COICOP and products sold overseas

We discuss preliminary results on the algorithm using the COICOP nomenclature.

Products sold only in overseas departments might be different from product sold both there and in mainland France, also requiring a labelling campaign.

REFERENCES

[1] Using scanner data from 2020 on : Impact on the CPI (January 2020)

[2] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, *Enriching Word Vectors with Subword Information* (2016)

[3] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, *Bag of Tricks for Efficient Text Classification* (2016)