Journées de méthodologie statistique de l'Insee 2022

Enrichissement de données de caisses à partir d'informations nutritionnelles

Une approche par appariement flou sur données de grande dimension

Lino Galiana *(*), Milena* Suarez-Castillo *(**), Lionel* Wilner *(*)*

(*) Insee, Direction des Etudes et Synthèses Economiques
(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

lino.galiana@insee.fr ; milena.suarez-castillo@insee.fr ; lionel.wilner@insee.fr

Mots-clés. Appariements flous, ElasticSearch, analyse textuelle
Domaines concernés 5.1 ; 5.2 ; 11.4 ; 11.5

Résumé

Depuis quelques années, les données de caisse des enseignes de la grande distribution participent au calcul des indices de prix à la consommation. Ces données de caisse offrent également une information très riche sur la composition de la consommation alimentaire sur le territoire français et peuvent ainsi permettre d'évaluer l'hétérogénéité spatiale des paniers de consommation. A la suite d'un enrichissement des produits de caractéristiques nutritionnelles (poids, unités, calories...), il devient possible de produire une cartographie de la qualité nutritionnelle de la consommation alimentaire dans les grandes enseignes afin d'étudier les inégalités dans la qualité nutritionnelle des produits consommées pour ensuite les relier à des caractéristiques socio-démographiques.

L'apport méthodologique de cette étude est l'enrichissement de données de caisse de type *big-data* avec plusieurs sources d'information grâce à des méthodes avancées

d'appariement floues. Les données de caisse utilisées proviennent de la startup RelevanC et couvrent l'ensemble du champ du groupe Casino (principalement Casino, Franprix, Monoprix et Leader Price) sur les périodes 2017-2018. Ces données sont enrichies à partir des bases Open Food Facts (référentiel crowdsourcée de produits alimentaires) et Ciqual (table nutritionnelle de l'ANSES). L'identifiant produit, dit code EAN (équivalent à un code barre), ne permettant pas un appariement systématique entre les sources, des méthodes d'appariement flou, exploitant la proximité entre des libellés textuels, sont mobilisées.

Afin de rendre cohérentes les différentes sources, une première étape, dite de normalisation du texte, permet d'harmoniser les libellés afin de réduire la dimension des entités signifiantes. Cette étape s'appuie sur des techniques classiques de normalisation du texte en traitement naturel du langage adaptées à notre contexte (casse, accentuation, retrait de mots de liaisons, exclusion de termes évoquant le poids, le nombre d'unités. . .).

Le travail d'appariement repose sur l'utilisation du moteur de recherche ElasticSearch dans lequel sont indexées les bases nutritionnelles. Cette méthode permet, de manière très rapide et flexible, de traiter un champ textuel (le nom d'un produit dans les données de caisse) comme une requête à trouver dans un corpus de documents (les produits dans les sources servant l'enrichissement). En l'occurrence, après avoir indexé les sources OpenFood et Ciqual, cela permet de considérer chaque produit dans les données de caisse comme une requête à trouver dans le corpus (OpenFood ou Ciqual). Les données Ciqual, qui représentent des aliments caractéristiques, ont été préalablement enrichies de dictionnaires de marques ou de variétés (par exemple les crus viticoles pour enrichir le corpus sur le vin) disponibles sur wikipedia. Des requêtes "floues" (sensible à la proximité des n-grammes de caractères) des libellés des produits RelevanC dans ce moteur de recherche permettent dès lors d'attribuer au produit disponible dans les données de caisse la caractéristique désirée (kcal, glucide, lipide...) du produit le plus similaire. La performance et la flexibilité d'ElasticSearch préfigurent des réutilisations possibles de cette approche dans de multiples problèmes d'appariements flous.

Afin de réduire le champ des possibles en fournissant une indication au moteur de recherche sur les produits à privilégier, nous utilisons un réseau de neurone, entraîné pour un projet interne "Paul-EAN", afin que la recherche soit faite en priorité dans des produits partageant la même COICOP (nomenclature européenne de produits).

De plus, pour tenir compte des limites des méthodes d'appariement floue, notamment la difficulté à reconnaître, au-delà de la proximité syntaxique des termes plus ou moins synonymes, ce travail mobilise des techniques de *word embedding* (plongement de mots) qui permettent des représentations vectorielles des libellés, à partir d'une technique nommée réseau de neurones siamois, afin de sélectionner des plus proches voisins selon une mesure de distance multidimensionnelle.