Domaine : traitement de données textuelles

L'apport de l'analyse textuelle à la statistique d'entreprise : l'exploitation de dix années de visites d'entreprises par les Direccte

Nicolas CAVALLO[1]

Cette étude montre l'enjeu pour la statistique d'entreprises de l'exploitation par des méthodes de statistique textuelle de l'information économique qualitative recueillie auprès des entreprises par les services en région du ministère de l'économie.

Afin d'analyser les caractéristiques économiques des entreprises qui déterminent leur développement, la statistique d'entreprise mobilise traditionnellement une information quantitative, soit par la réalisation d'enquêtes, soit par la collecte de données administratives. Les limites intrinsèques de cette information sont connues : la description d'un phénomène est réduite par le choix d'un nombre limité de modalités prédéfinies.

L'information qualitative recueillie dans le cadre d'entretiens avec les dirigeants d'entreprises permet d'appréhender certaines questions moins ou mal éclairées par la statistique, mais aussi des questions traditionnelles telle l'analyse de la conjoncture ou encore la connaissance des filières industrielles. L'information textuelle est en effet « ouverte » donc potentiellement très riche. Toutefois, les données textuelles soulèvent une difficulté majeure, qui limite fortement leur exploitation : un coût de collecte élevé rendant difficile la constitution de bases de données de taille assez conséquentes pour en inférer des résultats robustes par des traitements statistiques de masse.

La base de données ISIS, qui rassemble l'information issue des 60 000 visites d'entreprises réalisées en dix ans par les Directions régionales des entreprises, de la concurrence, de la consommation, du travail et de l'emploi (Direccte), les antennes régionales de la Direction Générale des Entreprises (DGE), permet de dépasser cette difficulté. Par sa volumétrie - plus de 30 millions de mots ! - et son enrichissement continu, cette base constitue une matière idéale pour mettre en œuvres les méthodes de la statistique textuelle avec, à la clé, des applications importantes aussi bien pour la statistique d'entreprise elle-même que pour les décideurs économiques grâce à un éclairage original porté sur les entreprises.

Deux catégories de méthodes du Data Mining ont été mises en œuvre avec les données sur les entreprises de la base ISIS afin d'éclairer des domaines stratégiques de l'analyse économique : la conjoncture et la connaissance des filières.

Une première analyse exploite le vocabulaire et les champs lexicaux de 18 000 projets d'entreprises décrits dans la base ISIS. Elle permet d'apprécier la situation conjoncturelle et son évolution telle qu'elle est perçue par les entreprises. Plus précisément, une dizaine d'indicateurs conjoncturels ont été construits sur la période 2008-2013, correspondant à autant de thématiques clés, comme l'activité, l'investissement, le développement à l'international, etc. Ces thématiques clés ont été identifiées par l'utilisation de méthodes d'analyse factorielle et de classification appliquées aux données textuelles relatives à la description des projets d'entreprises. Les indicateurs reflètent la fréquence d'apparition de ces thématiques clés. Grâce à l'enrichissement quotidien d'ISIS, cette première analyse constitue la base d'un nouvel outil d'analyse conjoncturelle dont pourrait se doter la DGE et les Direccte.

Une seconde analyse, réalisée à partir de l'information - des données textuelles également – issue de plus de 25 000 visites d'entreprises, permet de déterminer à quelle filière appartient une entreprise à partir des données textuelles disponibles. L'exercice a été mené dans le cas de la filière automobile. Il permet de mesurer, par le calcul d'un score, la probabilité qu'une entreprise appartienne à cette filière. Le calcul de ce score est établi grâce à un *modèle d'apprentissage supervisé* exploitant le pouvoir de caractérisation des entreprises plus fort et surtout plus ouvert des variables textuelles. Ce modèle apporte une connaissance appréciable des filières, placées au cœur de la politique industrielle actuelle, la statistique d'entreprise, sectorielle, n'offrant qu'un éclairage partiel sur ce sujet.

[1] Insee, nicolas.cavallo@insee.fr. L'auteur a réalisé cette étude au sein de la sous-direction de la prospective, des études et de l'évaluation économiques (P3E) de la DGE.