

Titre : modélisation par carte de Kohonen des effets reels de la production sur l'environnement et la santé

Laboratoires partenaires impliqués : IBISC (UEVE), Paris 2 **durée totale du stage** 6 mois (M2) ou 3 mois (M1) **date de début et de fin du stage** 15/02/2022 au 1/09/2022

MOTS CLÉS

machine learning, neural networks, feature extraction, représentation non linéaire, données discrètes, pollution industrielle

Contexte et objectifs

Plusieurs millions de morts prématurées par an sont imputées à la pollution de l'air. En France, ce nombre est estimé à 40 000 par Santé Publique France. Une large part de cette pollution est imputable à la production et aux industries [1]. Pourtant, des décisions de transformation des processus de production ou même des décisions comptables peuvent réduire significativement ces externalités négatives [2].

Dans le cadre d'un projet de recherche multidisciplinaire devant étudier l'apparition des clusters de cancer, nous souhaitons effectuer un travail préliminaire qui répond à plusieurs questions : Quels sont les effets d'une décision de réduction de l'activité économique sur les niveaux de pollution ? Quels sont les effets d'une diminution de la rentabilité sur les niveaux de pollution ?

Et plus largement, la diminution de l'utilisation de différents types de polluants a-t-elle le même effet sur la rentabilité de l'entreprise quel que soit le milieu (air, sol, eau) ?

L'objectif de ce stage est de trouver une feature map de la pollution produite par plusieurs milliers d'entreprises (du fait de leur activité industrielle) et ce que ces variables « expliquent » de la cartographie de la pollution à l'échelle nationale.

Méthodes

Pour cela nous aurons recours aux techniques d'analyse exploratoire de données non linéaires dérivées de l'analyse des correspondances multiples de Benzecri [3]. L'analyse des données est multidimensionnelle et descriptive et permet la visualisation des données et des individus dans une même représentation. Sans limiter l'analyse exploratoire à une simple observation de graphiques, son intérêt principal réside néanmoins dans la recherche de formes (clusters) sur les graphiques bivariés [5].

Programme de travail

étape 1 : Produire une base de données (BDD) à partir de plusieurs bases de données préexistantes. Cette BDD relationnelle est déjà fonctionnelle.

étape 2 : Faisabilité. Réalisation d'un modèle neuronal de type quantification vectorielle capable de réaliser une classification des observations, admettant une représentation analogue à une ACP.

Étape 3 : comparer ce modèle avec plusieurs modèles factoriels qui comportent des variables organisationnelles, des variables comptables ainsi que des variables sur les émissions de pollution.

Références

- [1] Citepa, juin 2022. Inventaire des émissions de polluants atmosphériques et de gaz à effet de serre en France – Format Secten, Rapport n°2071sec / 2022.
- [2] Chen et al. (2018)
- [3] Benzecri J.-P. (1977) - Analyse discriminante et analyse factorielle, Les Cahiers de l'Analyse des Données, II, n °4, p 369-406.
- [4] Lebart, L. Piron, M. ,Morineau, A. (2006) Statistique exploratoire multidimensionnelle : visualisation et inférence en fouilles de données, Sciences sup, Dunod.
- [5] Tukey J.W. (1977) Exploratory Data Analysis. Reading, MA., Addison-Wesley Publishing Company, 688p

Profil et compétences recherchées

Capacité à comprendre et à développer des algorithmes d'apprentissage adaptatif et à traiter données médicales, les indexer et les exploiter dans un système opérationnel pour réaliser la mission décrite ci-dessus. Compétences en programmation : Python ou R.

La pratique du français n'est pas obligatoire.

Vous êtes étudiant en M1/M2 en analyse de données ou en école d'ingénieur et vous souhaitez vous impliquer dans un projet de recherche qui vous challenge. Les productions réalisées durant le stage pourront donner lieu à un article de recherche / une communication pour un congrès scientifique.

Qualités professionnelles recherchées

autonomie, sens du relationnel pour interagir avec les équipes de recherche, motivation pour les nouvelles technologies, créativité pour mettre en place une solution innovante.

Encadrement et conditions scientifiques

Le projet est pluridisciplinaire, à l'interface de l'apprentissage automatique, de l'informatique et de l'économie. L'étudiant sera encadré par Vincent Vigneron du laboratoire IBISC (Univ d'Évry, Université Paris-Saclay), spécialistes en machine learning, traitement du signal et des images.

Contact:

Vincent Vigneron vincent.vigneron@univ-evry.fr

Téléphone : +33 6 63 568 760