Introduction à la modélisation des données fonctionnelles

Dr Alassane AW, Enseignant-Chercheur, ENSAE-Dakar.
Email: alassane.aw.ansd@gmail.com

Résumé

L'objectif de cette communication est de donner une introduction à la modélisation des données fonctionnelles. Ces dernières sont des courbes lisses de la forme

$$\{X_i(t), \quad t \in \mathcal{T}, \quad i = 1, \ldots, n\},$$

pour lesquelles les nombres réels $X_i(t)$ existent pour tout $t \in \mathcal{T} = [a, b]$, mais ne peuvent être observés que sur un nombre fini de points $t_j, j = 1, \ldots, p$. Les données de ce type sont rencontrées dans presque tous les domaines d'activités. Par exemple:
-En Économie: on peut s'intéresser au taux de croissance du PIB $X_i(t)$ d'un pays i au temps t. Ce taux existe à chaque instant t, mais ne peut être observé que pour un nombre fini de temps $t_j, j = 1, \ldots, p$. Le Sénégal fait partie des rares pays africains qui calculent le PIB tous les trimestres.
- En Santé: on peut s'intéresser à la prévalence $X_i(t)$ d'une certaine maladie M dans une région i au temps t. En effet, cette prévalence $X_i(t)$ existe à chaque instant t mais ne peut être observée que pour un nombre fini de temps $t_j, j = 1, \ldots, p$. La prévalence du diabète peut être calculée annuellement dans la région de Ziguinchor.
-En Sociologie: ont peut s'intéresser au taux de divorce $X_i(t)$ dans un département i au temps t. On peut encore noter que ce taux $X_i(t)$ existe à chaque instant t mais n'est observé que pour un nombre fini de temps $t_j, j = 1, \ldots, p$. Le taux de divorce peut être calculé annuellement dans le département de Ziguinchor.
Après un rappel sur les fondements mathématiques nécessaires à la modélisation des données fonctionnelles, nous étudierons comment les décrire d'un point de vue statistique.
Une partie importante de la communication sera consacrée à l'étude du modèle linéaire fonctionnel donné par la relation suivante

$$y_i = \int_{\mathcal{T}} X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \ldots, n,$$

où $\beta(\cdot)$ est une fonction inconnue dans $L^2(\mathcal{T})$, les ϵ_i sont les termes d'erreur du modèle supposés indépendants et identiquement distribués de moyenne nulle et de variance constante égale a σ^2. On fait l'hypothèse que les ϵ_i sont indépendants des $X_i(t)$. Le modèle linéaire fonctionnel est largement étudié dans la littérature. Il existe au moins trois méthodes qui permettent d'estimer le paramètre fonctionnel $\beta(\cdot)$: i) estimation à l'aide d'une base de fonctions fixe comme la base des B-splines, la base de Fourier ou la base des ondelettes, ii) estimation à l'aide de la base des fonctions propres de l'opérateur de covariance de la fonction aléatoire $X = \{X(t), t \in \mathcal{T}\}$ et iii) estimation à l'aide de la base des fonctions PLS.
Une autre partie importante de la communication portera sur les modèles autorégrssifs spatiaux fonctionnels qui sont des extensions du modèle linéaire fonctionnel de nature spatiale. Nous ferons l'état des lieux des modèles autorégressifs spatiaux fonctionnels développés jusqu'ici et dégagerons des perspectives de recherche qui permettront de faire avancer considérablement les connaissances sur la littérature des modèles autorégressifs spatiaux fonctionnels.