1 Définir un cluster spatial dans le cas des données fonctionnelles qualitatives

1.1 Contexte et encodage des trajectoires catégorielles

Dans le cas des données fonctionnelles quantitatives, un cluster spatial est défini comme une zone géographique dans laquelle la courbe moyenne des fonctions observées diffère de manière significative de la courbe moyenne calculée pour l'ensemble des individus situés en dehors de cette zone. Pour les données fonctionnelles catégorielles, chaque individu est représenté par une séquence d'états (par exemple, des statuts ou catégories observés en continu) plutôt que par une fonction réelle continue. Afin d'analyser ces trajectoires de manière comparable, on applique une méthode d'encodage optimal (comme celle présentée dans l'article cfda). Cette méthode transforme chaque trajectoire catégorielle en un ou plusieurs scores, notés z_i, qui résument la dynamique temporelle de la séquence.

1.2 Définition d'un décalage fonctionnel appliqué aux scores :

Pour étendre le concept de cluster spatial aux données fonctionnelles catégorielles, on remplace la moyenne des fonctions réelles par la moyenne des scores issus de l'encodage optimal. Autrement dit, on considère qu'un sous-ensemble spatial w forme un cluster si la moyenne des scores z_i des individus situés dans w est décalée par rapport à celle des individus situés en dehors de w. Formulé de manière formelle :

$$\mathbb{E}[z_i \mid s_i \in w] = \mathbb{E}[z_i \mid s_i \notin w] + \Delta, \quad \text{avec } \Delta \neq 0.$$

Ici, z_i est le score (ou le vecteur de scores) qui résume la trajectoire catégorielle de l'individu i. Le décalage Δ peut refléter une différence en terme de **fréquence** (ou durée) d'états et de **séquence** (ordre) des transitions entre états (similaire à une différence de magnitude ou de forme dans le cas réel). Cette différence peut être testée par un test non paramétrique (par exemple, une version fonctionnelle du test de Wilcoxon–Mann–Whitney).

Remarque : Les individus du cluster partagent un **comportement temporel spécifique**.

Exemple concret :

- Une région où les patients passent plus de temps dans un état critique après hospitalisation.

1.3 Données disponibles

- **Données fonctionnelles catégorielles** : chaque individu i possède une trajectoire temporelle

  $$X_i(t), \quad t \in [0, T], \quad \text{avec } X_i(t) \in S = \{s_1, s_2, \ldots, s_K\}$$

  Cela signifie que l'état de l'individu i évolue dans le temps, en prenant des valeurs dans un ensemble fini d'états catégoriels S.

- **Localisation spatiale** : chaque individu i est associé à une position géographique

$$s_i \in \mathbb{R}^2$$

qui peut correspondre à des coordonnées GPS où à une unité spatiale (région, commune, etc.).

1.4 1.3 Transformation des trajectoires

Grâce à l'**encodage optimal** (méthode CFDA ou FMCA), chaque trajectoire catégorielle $X_i(t)$ est projetée dans un espace vectoriel de dimension q :

$$X_i(t) \longmapsto \mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$$

- $\mathbf{z}_i$ est un vecteur de *scores numériques* représentant la trajectoire fonctionnelle de l'individu i.
- Chaque z_{ij} correspond à la contribution de la $j^{\text{ème}}$ composante principale.
- Cette projection permet de résumer les trajectoires fonctionnelles dans un espace numérique adapté à l'analyse statistique (clustering, régression, etc.).

2 Proposition de test : scan spatial basé sur Wilcoxon

Dans le cas des données fonctionnelles catégorielles, chaque trajectoire $X(t) \in S$ est représenté via des fonctions indicatrices 1_t^x, permettant de construire un vecteur $\tilde{X}(t) \in \{0,1\}^K$. Grace à un encodage optimal $a^x(t)$, les trajectoires sont projetées dans l'espace $\mathcal{L}^2(T, \mathbb{R})$, permettant ainsi d'analyser les données dans un cadre similaire à l'espace $\mathcal{L}^2$ des données fonctionnelles réelles.

2.1 1. Objectif

Comparer les scores z_i des individus dans une zone candidate w avec celles hors de cette zone w^c.

2. Hypothèses (d'après Smida et al., 2022)

- H_0 : $P_w = P_{w^c}$ (même distribution fonctionnelle dans et hors de w) (absence de cluster)
- $H_1^{(w)}$: P_w et P_{w^c} diffèrent par un **décalage fonctionnel** $\Delta \neq 0$ (présence de cluster)

3. Statistique de test

On considère les trajectoires fonctionnelles projetées en scores X_i pour $s_i \in w$ et X_j pour $s_j \in w^c$. Le test est défini par :

$$T_w = \frac{1}{\sqrt{|w||w^c|n}} \sum_{i, s_i \in w} \sum_{j, s_j \in w^c} \frac{X_j - X_i}{\|X_j - X_i\|} \quad \text{avec} \quad n = |w| + |w^c|$$

Puis, la statistique du test global est :

$$\text{NPFSS} = \max_{w \in \mathcal{W}} \|T_w\|$$

où $\mathcal{W}$ est l'ensemble des zones de balayage (candidats à être un cluster spatial).

4. Ce test est :

- Compatible avec les scores CFDA/FMCA (projection dans un espace vectoriel)
- Non paramétrique : pas besoin de supposer une loi normale

2.2 5. Rappel : Test de Wilcoxon–Mann–Whitney

Test de Wilcoxon–Mann–Whitney

Test non-paramétrique permettant de comparer deux groupes **indépendants** sur leur position centrale (médiane), sans faire d'hypothèse sur la forme des distributions.
Hypothèses :

- H_0 : les deux groupes suivent la même loi (aucun décalage),
- H_1 : les distributions diffèrent par un décalage de localisation.

Application ici :
On note P_w la loi des scores fonctionnels *à l'intérieur* d'une zone w, et P_{w^c} celle *en dehors*.
On considère les trajectoires catégorielles fonctionnelles $X_i(t)$ projetées en scores z_i.
On applique le test aux deux échantillons indépendants :

$$\{z_i : s_i \in w\} \quad \text{vs.} \quad \{z_j : s_j \notin w\}$$

où s_i représente la position spatiale de l'individu i. .