1 Clusters spatiaux en données fonctionnelles catégorielles

1.1 Introduction

Les **données fonctionnelles catégorielles** représentent l'évolution d'un individu ou d'un objet au cours du temps sous forme de catégories (états) successives, plutôt que de valeurs numériques continues. Autrement dit, chaque individu suit une trajectoire composée d'états (par exemple : différentes phases d'une maladie, niveaux de pollution, etc.) enregistrés à différents moments. On peut **résumer cette trajectoire par un score** ou un **petit ensemble d'indicateurs** qui reflètent, entre autres, la **fréquence d'apparition de certains états**, **le temps passé dans ces états** et **l'ordre dans lequel l'individu passe d'un état à un autre**.

Exemple (santé) : Imaginons le parcours d'un patient à travers différents stades de maladie (états « léger », « modéré », « sévère », « rémission »). Sa trajectoire catégorielle pourrait être quelque chose comme :

$$\text{léger} \to \text{modéré} \to \text{sévère} \to \text{modéré} \to \text{rémission},$$

et l'on peut en extraire un score qui résume ce parcours (par exemple, le temps passé en stade sévère, le nombre de transitions, etc.).

Exemple (environnement) : Considérons un capteur environnemental qui classe chaque journée selon la qualité de l'air (« bonne », « moyenne », « mauvaise »). La suite de ces catégories au fil du temps forme une trajectoire (par exemple :

$$\text{bonne} \to \text{moyenne} \to \text{bonne} \to \text{mauvaise} \to \ldots),$$

que l'on peut également résumer par des indicateurs (tels que la fréquence des jours « mauvais » ou des séquences typiques (motifs ou schémas récurrents des transitions)).

1.2 Cadre des données fonctionnelles catégorielles

Considérons un ensemble fini d'états catégoriels $E = \{e_1, e_2, \ldots, e_K\}$ et un domaine temporel T (par exemple $[0, T]$ pour un temps continu, ou $1, \ldots, T$ pour un temps discret). Une donnée fonctionnelle catégorielle est la trajectoire d'un individu au cours du temps, représentée par une fonction $X_i : T \to S$. Autrement dit, chaque individu i est associé à une suite d'états $X_i(t)$ évoluant dans le temps : à tout instant $t \in T$, $X_i(t)$ indique l'état $e_k \in E$ dans lequel se trouve l'individu i. On obtient ainsi pour chaque individu une trajectoire constituée de segments constants et de sauts d'un état à un autre au fil du temps (suite chronologique d'états). Ces trajectoires catégorielles, vues comme des réalisations d'un processus stochastique à valeurs dans S, forment le cadre des données fonctionnelles catégorielles.

Encodage optimal des trajectoires catégorielles (FMCA)

Les données fonctionnelles catégorielles ne se prêtent pas directement aux méthodes classiques d'analyse fonctionnelle, car les états e_k sont qualitatifs et non numériques.

Afin de pouvoir appliquer des techniques de réduction de dimension ou de comparaison statistique, on doit transformer ces trajectoires en représentations numériques, c'est-à-dire les encoder dans un espace vectoriel.

Une approche efficace est l'encodage optimal par Analyse des Correspondances Multiples Fonctionnelle (FMCA). L'idée de la FMCA est d'étendre l'analyse des correspondances multiples (ACM) classique à un ensemble (potentiellement infini) de variables indexées par le temps. Concrètement, on cherche à trouver une représentation numérique des états évoluant dans le temps qui maximise l'inertie (la variance) expliquée par les premières composantes principales.

Pour chaque état $e \in E$, on définit une fonction d'encodage optimale $a_e(t)$ qui attribue une valeur numérique à cet état à chaque instant t. Ensuite, la trajectoire catégorielle d'un individu, notée $X_i(t)$ initialement une suite d'états dans E mesurés à des instants discrets ou sur un intervalle continu est transformée en une fonction numérique continue par expansion sur une base fonctionnelle (par exemple, une base de Fourier ou de B-splines). Autrement dit, on approxime $X_i(t)$ par une combinaison linéaire de fonctions de base, ce qui permet d'obtenir pour chaque individu i un vecteur de scores

$$z_i = \big(z_{i1}, z_{i2}, \ldots, z_{id}\big) \in \mathbb{R}^d,$$

où la dimension d est faible (typiquement $d = 2$ ou 3).

Ces scores représentent les coordonnées de l'individu i sur les axes principaux optimaux et résument l'essentiel de sa trajectoire catégorielle. Ils intègrent notamment des informations sur :

- la fréquence d'apparition de chaque état,
- la durée passée dans chacun des états,
- et l'ordre dans lequel les transitions entre états se produisent au fil du temps.

Ainsi, grâce à cette projection, chaque trajectoire catégorielle est finalement représentée par un point (ou un vecteur numérique) dans $\mathbb{R}^d$. Cette représentation réduite facilite ensuite la comparaison quantitative entre les trajectoires des individus ainsi que l'application de méthodes statistiques usuelles (clustering, régression, etc.) sur ces données transformées.

1.3 Cadre spatial des trajectoires

Dans de nombreuses applications, les individus suivis spatialement disposent, en plus de leur trajectoire catégorielle, d'une localisation géographique. Autrement dit, à chaque individu i est associée une position spatiale s_i, qui peut correspondre, selon le contexte, à des coordonnées GPS (notamment pour des capteurs environnementaux, stations météorologiques ou objets mobiles), ou au centroïde de l'unité spatiale de rattachement (notamment pour des individus humains, afin de garantir l'anonymat). On dispose alors d'informations à la fois fonctionnelles (la trajectoire temporelle $X_i(t)$) et spatiales (la position s_i). Le cadre spatial consiste à tenir compte de ces positions afin d'étudier comment les comportements dynamiques (trajectoires catégorielles) peuvent varier d'une zone géographique à une autre. On cherche en particulier à identifier des sous-régions du domaine spatial dans lesquelles les trajectoires présentent des comportements atypiques par rapport au reste du territoire. Grâce à l'encodage optimal introduit précédemment, on peut aborder cette question en comparant les vecteurs de scores z_i des individus d'une région donnée à ceux des individus situés ailleurs.

1.4 Définition d'un cluster spatial en données fonctionnelles catégorielles

Un **cluster spatial** est une zone géographique où les trajectoires catégorielles des individus **sont très différentes** de celles observées ailleurs. Plus précisément, cela signifie que si l'on considère le score (ou les indicateurs) qui résume la trajectoire de chaque individu, alors **la moyenne des scores** dans cette zone est **sensiblement différente de la moyenne observée ailleurs**. Autrement dit, les individus de cette zone partagent, en moyenne, un comportement (un profil) qui diffère de celui du reste de la population.

Ce « décalage » peut se traduire par une valeur moyenne plus **élevée** ou plus **basse** que celle du reste de la région (on parle alors de cluster à haute ou à basse valeur).

En quoi les trajectoires du cluster diffèrent-elles ?

Si une zone forme un cluster, c'est que le profil moyen des trajectoires des individus qui y vivent est significativement différent de celui des individus dans le reste du territoire. Par exemple, on pourrait constater dans cette zone que les personnes :

- **Passent plus (ou moins) de temps dans certains états** qu'ailleurs (par ex. une durée moyenne plus longue dans un stade sévère de maladie, ou plus courte dans cet état, comparé au reste du territoire).
- **Connaissent plus fréquemment (ou plus rarement) certains états** spécifiques (par ex. une fréquence anormalement élevée de journées de mauvaise qualité de l'air dans un secteur donné par rapport aux autres secteurs).
- **Suivent un enchaînement d'états différent** (l'ordre des transitions entre états est particulier, par ex. les patients du cluster passent directement de « léger » à « sévère » sans étape « modéré », ce qui serait inhabituel en dehors de cette zone).

1.5 Définition (mathématique) d'un cluster spatial en données fonctionnelles catégorielles

On définit un **cluster spatial fonctionnel catégoriel** comme une zone géographique w (un sous-ensemble de la région d'étude) pour laquelle **les trajectoires des individus locaux diffèrent significativement de celles des autres individus**. Formellement, si l'on note $\mathbb{E}[z_i \mid s_i \in w]$ la moyenne des scores des individus dont la position s_i est à l'intérieur de w, et $\mathbb{E}[z_i \mid s_i \notin w]$ la moyenne des scores en dehors de w, le fait que w constitue un cluster spatial se traduit par l'existence d'un écart non nul entre ces deux moyennes :

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

En d'autres termes, la moyenne des vecteurs z_i à l'intérieur de w s'écarte de la moyenne globale par un vecteur Δ non nul. Ce décalage Δ peut refléter une différence significative dans l'un ou plusieurs des aspects caractérisant les trajectoires :

- une **fréquence** de visite de certains états plus élevée ou plus faible pour les individus de w,
- une **durée de séjour** dans ces états qui diffère (par exemple, les individus de w restent plus longtemps dans un état donné ou, au contraire, changent d'état plus rapidement),
- un **enchaînement d'états spécifique** (par exemple, des transitions qui suivent un ordre ou un motif inhabituel chez les individus de w).

De telles différences dans les trajectoires moyennes signifient que la zone w présente un comportement atypique. La définition ci-dessus généralise ainsi les notions de cluster de magnitude et de cluster de forme rencontrées dans le cas de données fonctionnelles réelles (numériques) : un cluster de magnitude correspond classiquement à un décalage global de la fonction moyenne (par exemple un niveau global plus élevé ou plus faible sur toute la courbe), tandis qu'un cluster de forme traduit une altération de la forme de la fonction (par exemple un profil temporel différent, avec des variations locales spécifiques). Dans le contexte catégoriel, la différence Δ observée pour un cluster spatial peut simultanément impliquer des écarts de "magnitude" (tels qu'une fréquence d'occurrence d'un état nettement plus grande) et de "forme" (un enchaînement temporel des états distinctif), offrant une caractérisation complète des anomalies de trajectoires dans la région w.

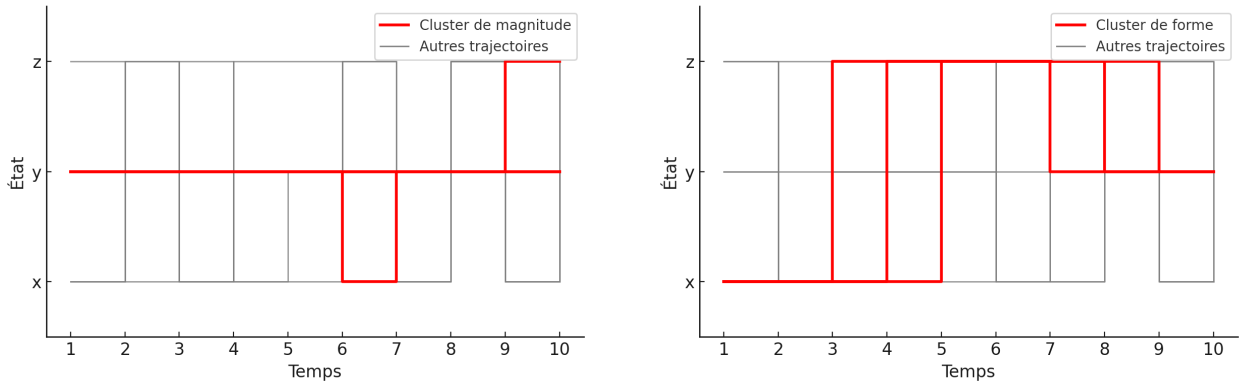

FIGURE 1 – Trajectoires simulées illustrant un cluster de magnitude (gauche) et de forme (droite)

Les courbes discrètes ci-dessus représentent l'évolution d'un état x, y ou z en fonction du temps (axe horizontal de $t = 1$ à 10 ; axe vertical indiquant l'état). Dans cet exemple, on observe un **cluster de magnitude** (**en rouge**) composé de plusieurs trajectoires catégorielles qui demeurent longtemps dans le même état (ici l'état y domine sur presque tout l'intervalle temporel). En revanche, les autres trajectoires (**en gris**) changent fréquemment d'état au fil du temps. Autrement dit, les membres du cluster rouge effectuent très peu de transitions et se caractérisent par un temps passé très élevé dans l'état y, alors que les trajectoires hors-cluster présentent une évolution plus erratique (instable) entre les trois états.

Pour le **cluster de forme**, Chaque membre du cluster rouge commence à l'état x, passe ensuite à l'état z, pour finalement aboutir à l'état y, soit une séquence commune $x \to z \to y$ sur l'intervalle $t = 1 \ldots 10$. Ce profil temporel particulier s'oppose avec celui des autres trajectoires en gris, qui suivent des transitions d'états variées (par exemple, certaines passent de x à y directement, ou alternent plus fréquemment entre les trois états). En somme, le cluster de forme se caractérise par une structure ou forme temporelle commune des trajectoires catégorielles, plutôt que par un état dominant. Les membres du cluster partagent le même ordre de passages par les états (la même "forme" de courbe catégorielle), ce qui correspond bien à la notion de cluster de forme et les distingue clairement des autres trajectoires.

2 Détection de clusters géographiques par scan spatial non paramétrique (méthode de Jung & Cho, 2015)

2.1 Contexte et objectif du scan spatial non paramétrique

On considère N observations localisées géographiquement (par exemple des individus ou des sites), pour lesquelles on dispose d'un score numérique z_i obtenu par la méthode CFDA. L'objectif est **d'identifier une zone géographique où les scores z_i sont significativement plus élevés ou plus faibles que dans le reste de la région**. Pour cela, on utilise un scan spatial non paramétrique inspiré de **Jung et Cho** (2015), qui ne fait aucune hypothèse de distribution des z_i et s'appuie sur le test de somme des rangs de **Wilcoxon-Mann-Whitney** pour comparer les scores à l'intérieur vs. à l'extérieur de multiples fenêtres spatiales candidates. Le cluster le plus probable (**MLC**) sera défini comme la fenêtre spatiale présentant la plus forte différence (d'après le test de Wilcoxon), c'est-à-dire associée à la plus petite **p-valeur** de test. La significativité globale de ce cluster sera ensuite évaluée par une procédure de permutation afin de tenir compte des tests multiples effectués lors du scan.

2.2 Fenêtres spatiales candidates et hypothèses du test

On doit d'abord définir l'ensemble des fenêtres spatiales à examiner par le scan. Typiquement, on considère un très grand nombre de fenêtres de différentes tailles et positions (centres) couvrant la zone d'étude. Par exemple, on peut utiliser des fenêtres circulaires centrées successivement sur chaque point de donnée (ou sur chaque centroid de région), en faisant varier le rayon du cercle de 0 jusqu'à une taille maximale (souvent le rayon pour lequel la fenêtre couvre au plus 50% des observations, afin d'éviter de considérer des fenêtres trop grandes englobant presque toute la population). Chaque fenêtre w (par exemple un cercle) est un cluster potentiel contenant n_w observations (avec scores $z_i : i \in w$) et excluant les $N - n_w$ autres observations en dehors ($z_i : i \notin w$).

Pour chaque fenêtre candidate w, on souhaite tester :

- **Hypothèse nulle** H_0 : les scores à l'intérieur de w suivent la même distribution que ceux à l'extérieur, i.e. aucun cluster, les différences observées sont dues au hasard.
- **Hypothèse alternative** $H_1^{(w)}$: la distribution des scores à l'intérieur de w est décalée par rapport à celle à l'extérieur. En particulier, on peut détecter soit un décalage positif ($\Delta > 0$) indiquant que les valeurs à l'intérieur de w tendent à être plus élevées qu'à l'extérieur (cluster de scores élevés), soit un décalage négatif ($\Delta < 0$) pour un cluster de scores faibles. Le test est non paramétrique et ne requiert pas que les z_i suivent une loi particulière (pas d'hypothèse de normalité, etc.), il s'appuie uniquement sur le rang des observations.

2.3 Statistique de test de Wilcoxon pour une fenêtre w donnée

Pour comparer les scores à l'intérieur vs. à l'extérieur de w, on utilise la **statistique de rang de Wilcoxon-Mann-Whitney**. La procédure est la suivante :

- **Classement (rang) des données** : On ordonne l'ensemble des N valeurs z_i par ordre croissant et on attribue à chaque observation i son rang R_i (entre 1 et N, en utilisant des rangs moyens en cas d'ex æquo).

- **Somme des rangs dans** w : On calcule la statistique

$$W_w = \sum_{i \in w} R_i,$$

c'est-à-dire la somme des rangs de toutes les observations situées dans la fenêtre w. Intuitivement, si w contient principalement des valeurs z_i élevées par rapport au reste (rangs élevés), W_w sera grande, et inversement W_w sera petite si w ne contient que des valeurs faibles. Sous l'hypothèse nulle d'absence de cluster, les rangs à l'intérieur et hors w sont mélangés aléatoirement.

- **Distribution de** W_w **sous** H_0 : Sous H_0, la statistique W_w suit une loi dont on peut déterminer l'espérance et la variance en fonction de n_w (taille de w) et N. En effet, chaque ensemble de n_w observations aurait en moyenne la moitié des plus grands rangs. Mathématiquement :

$$\mathbb{E}(W_w) = n_w \frac{N+1}{2} \quad \text{et}$$

$$Var(W_w) = n_w(N - n_w)\frac{N+1}{12}$$

d'après la théorie du test de Wilcoxon.

- **Statistique normalisée** T_w : Pour évaluer l'écart de W_w par rapport à H_0, on la transforme en un score normalisé :

$$T_w = \frac{W_w - \mathbb{E}(W_w)}{\sqrt{Var(W_w)}}.$$

Sous l'hypothèse nulle, T_w est approximativement distribuée selon $\mathcal{N}(0,1)$ pour des tailles d'échantillon suffisantes (règle usuelle : $n_w \geq 10$ et $N - n_z \geq 10$). Si n_z ou $N - n_w$ est petit, on peut au besoin calculer la p-valeur exacte de W_z en utilisant sa distribution hypergéométrique sans passer par l'approximation normale.

- **Calcul de la p-valeur (test de Wilcoxon)** : On réalise un test de Wilcoxon-Mann-Whitney pour comparer les niveaux de z_i à l'intérieur vs. extérieur. Étant donné la statistique observée T_w :
  - Pour rechercher un cluster de scores élevés (alternative H_1 : valeurs de w plus grandes), on effectue un test unilatéral à droite : la p-valeur associée est $p_w = P_{H_0}(T_w \geq t_{\text{obs}}) \approx 1 - \Phi(t_{\text{obs}})$, où Φ est la fonction de répartition de $\mathcal{N}(0,1)$. Une valeur T_w très grande (beaucoup plus grande que 0) donnera une p-valeur très petite, signifiant que w est un cluster de scores anormalement hauts.
  - Inversement, pour détecter un cluster de scores faibles, on utilise un test unilatéral à gauche : $p_w = P_{H_0}(T_w \leq t_{\text{obs}}) = \Phi(t_{\text{obs}})$. Un T_w très négatif implique une p-valeur gauche faible, révélant un cluster de valeurs exceptionnellement basses.

Dans la pratique, on examine les deux alternatives : si T_w est positif on peut interpréter w comme un cluster potentiel de haute valeurs (p-valeur droite), et s'il est négatif comme un cluster de faibles valeurs (p-valeur gauche). On retient dans les deux cas la p-valeur unilatérale appropriée la plus petite. (Pour un test purement bilatéral de différence, on pourrait prendre $2\min(\Phi(t_{\text{obs}}), 1-\Phi(t_{\text{obs}}))$, mais ici l'intérêt est d'identifier spécifiquement des clusters haut ou bas).

2.4 Statistique de scan spatial et identification du cluster le plus probable

On applique le test ci-dessus à toutes les fenêtres w de l'ensemble des candidats. Chaque fenêtre fournit une p-valeur p_w mesurant à quel point les données à l'intérieur de w diffèrent de l'extérieur. **La statistique de scan spatial proposé par Jung & Cho (2015)** est défini comme la plus petite de ces p-valeurs :

$$\Lambda_{\min} = \min_{w \in W} p_w,$$

où $\mathcal{W}$ est l'ensemble de toutes les fenêtres considérées. Le cluster le plus probable est alors la fenêtre w^* correspondant à cette p-valeur minimale $\Lambda_{\min}$:

$$w^* = \arg\min_{w \in W} p_w.$$

Par construction, w^* est la zone qui présente la plus forte évidence statistique de différentiel de valeurs par rapport au reste du territoire. Autrement dit, w^* maximise en valeur absolue l'écart de rangs T_w (que ce soit dans le sens positif ou négatif) et fournit l'indication la plus significative d'un cluster.

Interprétation : Si $\Lambda_{\min}$ est très petit, cela suggère qu'il existe au moins un cluster spatial inhabituel dans les données. Cependant, il faut faire attention car tester de multiples fenêtres entraîne un risque de faux positifs (Type I : rejetter H_0 alors que H_0 est vrai) accru (tests multiples). En effet, même si aucune fenêtre n'est réellement cluster (sous H_0 vrai), en examinant des centaines de fenêtres on peut obtenir par hasard une p-valeur très petite. Pour cette raison, On n'interprète pas $\Lambda_{\min}$ directement comme la probabilité d'observer un cluster sous H_0, mais on va **estimer la significativité par une méthode de permutation** (voir ci-dessous).

2.5 Évaluation de la significativité par permutation

Pour déterminer si le cluster détecté w^* est statistiquement significatif (c'est-à-dire si $\Lambda_{\min}$ observé est vraiment exceptionnellement petit), on utilise une procédure de permutation (Monte Carlo) conforme à celle employée dans le scan spatial de Kulldorff. L'idée est de simuler la distribution nulle de $\Lambda_{\min}$ en l'absence de cluster réel, afin de tenir compte des comparaisons multiples effectuées. Concrètement :

- **Permutation des scores** : On génère un grand nombre B de jeux de données simulés sous H_0 en **permutant aléatoirement** les scores z_i entre les positions géographiques. Chaque permutation attribue au hasard les valeurs aux lieux, brisant toute structure spatiale tout en conservant la distribution des z_i.
- **Recalcule du scan pour chaque permutation** : Pour chaque jeu de données permuté b (avec $b = 1, \ldots, B$), on répète entièrement le procédé de scan spatial : on classe les valeurs, on calcule W_w, T_w et p_w pour toutes les fenêtres w, et on trouve la plus petite p-valeur $\Lambda_{\min}^{(b)}$ de la permutation. Cela donne B réalisations $\{\Lambda_{\min}^{(1)}, \Lambda_{\min}^{(2)}, \ldots, \Lambda_{\min}^{(B)}\}$ sous l'hypothèse nulle.
- **Calcul de la p-valeur globale** : On estime la p-valeur associée au cluster observé w^* en comparant $\Lambda_{\min}$ aux valeurs permutées. Par exemple, la p-valeur empirique du scan peut être calculée comme :

$$p_{globale} = \frac{R}{B+1},$$

où R est le rang de $\Lambda_{\min}$ parmi les $B+1$ valeurs de $\Lambda_{\min}$ calculées (les B issues des jeux simulés sous H_0 et la valeur observée sur les données réelles).

Si P_{globale} est inférieur au seuil α (e.g. 0,05), on conclut que le cluster w^* mis en évidence est significatif et très improbable sous l'hypothèse de hasard spatial. Dans le cas contraire (P_{globale} élevé), on considèrera qu'aucun cluster n'est détecté de façon concluante (même si w^* avait la plus petite p-valeur parmi les fenêtres, elle n'était pas assez exceptionnelle par rapport au bruit aléatoire). Cette procédure de permutation garantit le contrôle du risque de faux positifs du scan spatial au niveau α choisi (on parle de test global de présence d'au moins un cluster). En effet, elle prend en compte le fait qu'on a exploré de très nombreux clusters candidats. Jung et Cho (2015) soulignent que, grâce à ce recalibrage par permutation, leur méthode maintient correctement le niveau de significativité tout en gagnant en puissance par rapport à un modèle paramétrique inadapté. En pratique, on choisit B assez grand (par ex. 999 ou 9999 permutations) afin d'estimer avec précision la p-valeur globale. Les clusters secondaires (autres zones significatives) peuvent également être rapportés : on pourra répéter la recherche en excluant le cluster principal ou utiliser des algorithmes détectant des clusters multiples, en veillant à ne retenir que des clusters indépendants (sans chevauchement géographique) et statistiquement significatifs après correction.

.