ÉCOLE SUPÉRIEURE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION

STATISTIQUE SPATIALE ET MODÉLISATION SPATIO-TEMPORELLE

Exploration de la Dynamique Spatiale et
Spatio-Temporelle des Joueurs de NFL

Présenté par :
Marzouk Zayd

Enseignante :
Jelassi Selma

Année Universitaire 2024-2025

Contents

1 Introduction **5**
- 1.1 Contexte du projet . . . 5
- 1.2 Problématique . . . 5
- 1.3 Objectifs du projet . . . 5
- 1.4 Méthodologie . . . 5
- 1.5 Plan du rapport . . . 6

2 Cadre théorique et concepts fondamentaux **6**
- 2.1 Introduction . . . 6
- 2.2 Autocorrélation spatiale . . . 7
  - 2.2.1 Indice de Moran . . . 7
  - 2.2.2 Indice de Geary . . . 7
- 2.3 Matrices de pondération spatiale . . . 7
- 2.4 Indicateurs locaux d'autocorrélation spatiale (LISA) . . . 8
- 2.5 Modèles de régression spatiale . . . 9
  - 2.5.1 Modèle linéaire classique (OLS) . . . 9
  - 2.5.2 Modèle SAR (Spatial Autoregressive Model) . . . 9
  - 2.5.3 Modèle d'erreur spatiale (SEM) . . . 10
  - 2.5.4 Modèle SLX . . . 10
  - 2.5.5 Modèle GWR . . . 11
- 2.6 Le Krigeage . . . 11
- 2.7 Modélisation spatio-temporelle . . . 12
  - 2.7.1 Régression linéaire spatio-temporelle . . . 12
  - 2.7.2 Regression Random forest . . . 12
  - 2.7.3 Modèle STAR . . . 13
  - 2.7.4 Modele STSLM . . . 13

3 Étude pratique et modélisation **13**
- 3.1 Description des données . . . 13
- 3.2 Prétraitement des données . . . 13
- 3.3 Application des méthodes spatiales . . . 14
  - 3.3.1 Distribution spatiale des joueurs . . . 14
  - 3.3.2 Construction de matrice de contiguïté W . . . 15
  - 3.3.3 Analyse d'autocorrélation spatiale . . . 16
  - 3.3.4 Carte LISA . . . 17
  - 3.3.5 Modèle de régression linéaire OLS . . . 18
  - 3.3.6 Modèle SAR . . . 19
  - 3.3.7 Modèle SEM . . . 20
  - 3.3.8 Modèle SLX . . . 21
  - 3.3.9 Modèle GWR . . . 21
  - 3.3.10 Krigeage . . . 22

3.4 Modélisation spatio-temporelle 26
3.4.1 Modèle STAR 26
3.4.2 Modèle Modele STSLM 27
3.5 Interprétation des résultats 28
3.5.1 Comparaison des modèles spatiales 28
3.5.2 Comparaison des modèles spatio-temporelles 28

4 Conclusion **29**

List of Figures

1 Visualisation des positions des joueurs 14
2 Visualisation de matrice de poids spatiaux 15
3 Carte LISA 16
4 Visualisation de matrice de poids spatiaux après modification 17
5 Carte LISA 1 18
6 Carte LISA 2 18
7 Carte de Krigeage de la variable dis (distance) 22
8 Carte de Krigeage apres modification : ordinaire 23
9 Carte de Krigeage apres modification : power 23
10 Carte de Krigeage apres modification : gaussien 24
11 Carte de Krigeage apres modification : spherical 24
12 Carte de Krigeage apres modification : exponential 25

List of Tables

1 Résultats de la régression STSLM (Spatial Temporal Spatial Lag Model) 27
2 Effets direct, indirect et total des variables dans le modèle STSLM 27
3 Résumé comparatif des modèles spatiaux 28
4 Comparaison entre les modèles STAR et STSLM 28

1 Introduction

1.1 Contexte du projet

Avec l'essor des technologies de suivi GPS et de capteurs embarqués, le domaine du sport professionnel connaît une révolution dans la manière d'analyser les performances des athlètes. Dans la NFL (National Football League), la base de données Next Gen Stats (NGS) fournit un suivi spatio-temporel extrêmement détaillé de chaque joueur pendant les actions de jeu. Chaque déplacement, chaque changement de direction et chaque accélération est enregistré avec une grande précision, ouvrant la voie à des analyses poussées du comportement des joueurs sur le terrain.

1.2 Problématique

Malgré la richesse de cette base de données, l'analyse fine des trajectoires et des interactions entre les joueurs reste complexe. Comment exploiter efficacement cette base spatio-temporelle pour extraire des connaissances sur les comportements individuels ou collectifs ? Quels outils statistiques peuvent être utilisés pour modéliser ces dynamiques ? Plus spécifiquement, comment les techniques de statistique spatiale et de modélisation spatio-temporelle permettent-elles de mieux comprendre les stratégies, performances ou anomalies de jeu ?

1.3 Objectifs du projet

Ce projet vise à :

- Exploiterla base NGS pour analyser les déplacements des joueurs en fonction du temps et de leur position sur le terrain
- Mettre en œuvre des méthodes de statistique spatiale (intensité, distribution des trajectoires, densité de mouvement)
- Construire des modèles spatio-temporels pour comprendre l'évolution des actions dans le temps (trajectoires,interactions)
- Extraire des indicateurs de performance et des insights pertinents à partir de ces analyses.

1.4 Méthodologie

La méthodologie adoptée s'articule autour des étapes suivantes :

1. Prétraitement des données (nettoyage, synchronisation)
2. Exploration spatiale (cartographie des trajectoires, densité de présence)
3. Analyse spatio-temporelle (reconstruction des vitesses, détection d'événements)
4. Modélisation statistique (modèles de points spatiaux, séries spatio-temporelles)

5. Interprétation et visualisation des résultats

1.5 Plan du rapport

Chapitre 1 : Cadre théorique et concepts fondamentaux

- Introduction à la statistique spatiale
- Autocorrélation spatiale
- Modèles de régression spatiale
- Le Krigeage
- Modélisation spatio-temporelle

Chapitre 2 : Étude pratique et modélisation

- Description des données
- Prétraitement des données
- Application des méthodes spatiales
- Modélisation spatio-temporelle
- Interprétation des résultats

Chapitre 3 : Conclusion générale

- Bilan du travail réalisé
- Apports de la modélisation spatio-temporelle dans le sport

2 Cadre théorique et concepts fondamentaux

2.1 Introduction

La statistique spatiale est une discipline qui étudie des données associées à des positions géographiques. Contrairement à la statistique classique, elle tient compte de la dépendance entre observations proches dans l'espace. Elle permet de décrire, modéliser et prévoir des phénomènes spatiaux à l'aide d'outils comme l'autocorrélation spatiale, la régression spatiale ou le krigeage. Dans ce projet, elle est appliquée à l'analyse des déplacements de joueurs sur un terrain de football américain, afin de mieux comprendre leur comportement en fonction de leur position.

2.2 Autocorrélation spatiale

L'autocorrélation spatiale mesure la relation entre des valeurs d'une variable et leur localisation dans l'espace. Elle permet de savoir si des observations proches ont tendance à se ressembler ou à différer. Des indices comme celui de Moran ou de Geary sont utilisés pour quantifier cette dépendance. Une autocorrélation positive indique une organisation spatiale (valeurs similaires regroupées), tandis qu'une autocorrélation négative traduit une dispersion.

2.2.1 Indice de Moran

L'indice global de Moran mesure l'autocorrélation spatiale. Il est défini par :

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

L'indice I de Moran varie généralement entre -1 et 1 :

- $I > 0$: autocorrélation spatiale positive
- $I < 0$: autocorrélation spatiale négative
- $I \approx 0$: absence d'autocorrélation spatiale

2.2.2 Indice de Geary

Le coefficient C de Geary (Geary, 1954) est une mesure alternative d'autocorrélation spatiale qui, contrairement à l'indice de Moran, se focalise sur les différences entre les valeurs des unités voisines :

$$C = \frac{(n-1)}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Le coefficient C de Geary varie généralement autour de 1 :

- $C < 1$: autocorrélation spatiale positive
- $C > 1$: autocorrélation spatiale négative
- $C \approx 1$: absence d'autocorrélation spatiale

2.3 Matrices de pondération spatiale

La prise en compte des relations spatiales repose sur la définition d'une matrice de pondération spatiale, notée W, qui formalise les relations de voisinage entre les unités spatiales. Cette matrice est composée d'éléments w représentant l'intensité de la relation entre les unités i et j. Plusieurs approches permettent de définir cette matrice :

- **Matrice de contiguïté** :
  $w_{ij} = 1$ si les unités i et j partagent une frontière, 0 sinon.

- **Matrice de distance** :
  w_{ij} est une fonction décroissante de la distance entre i et j (par exemple : $w_{ij} = \frac{1}{d_{ij}}$).
- **Matrice des k plus proches voisins** :
  $w_{ij} = 1$ si j est l'un des k plus proches voisins de i, 0 sinon.

Dans la pratique, la matrice W est souvent standardisée par ligne, c'est-à-dire que la somme des poids pour chaque ligne est égale à 1 :

$$\sum_j w_{ij} = 1$$

Cela facilite l'interprétation des coefficients et assure une échelle cohérente entre les unités spatiales.

2.4 Indicateurs locaux d'autocorrélation spatiale (LISA)

Les indicateurs globaux comme l'indice de Moran ou le coefficient de Geary donnent une mesure globale de l'autocorrélation spatiale sur l'ensemble de la zone étudiée. Cependant, ils ne permettent pas de détecter des structures locales ou des anomalies spatiales spécifiques. Pour répondre à cette limite, Anselin (1995) a introduit les indicateurs locaux d'association spatiale, appelés LISA (Local Indicators of Spatial Association). Ces indicateurs décomposent les indices globaux pour détecter les contributions locales à l'autocorrélation spatiale globale. Le LISA local basé sur l'indice de Moran est défini pour chaque unité i par la formule :

$$I_i = z_i \times \sum_j w_{ij} z_j$$

où :

- $z_i = \dfrac{x_i - \bar{x}}{\sigma}$ est la valeur standardisée de la variable pour l'unité i,
- $\bar{x}$ est la moyenne des x_i,
- σ est l'écart-type,
- w_{ij} est le poids spatial entre les unités i et j.

Les LISA permettent d'identifier quatre types de configurations spatiales :

- **High-High (HH)** : unité à valeur élevée entourée d'unités à valeur élevée (cluster positif)
- **Low-Low (LL)** : unité à valeur faible entourée d'unités à valeur faible (cluster négatif)
- **High-Low (HL)** : unité à valeur élevée entourée d'unités à valeur faible (outlier)
- **Low-High (LH)** : unité à valeur faible entourée d'unités à valeur élevée (outlier)

Les deux premiers (**HH** et **LL**) révèlent des *clusters spatiaux homogènes*, tandis que les deux derniers (**HL** et **LH**) indiquent des *anomalies spatiales locales*.

2.5 Modèles de régression spatiale

2.5.1 Modèle linéaire classique (OLS)

Le modèle de régression linéaire classique est formulé comme suit :

$$y = X\beta + \epsilon$$

où :

- y est le vecteur de la variable dépendante,
- X est la matrice des variables explicatives,
- β est le vecteur des coefficients à estimer,
- ε est le terme d'erreur, supposé indépendant et identiquement distribué (i.i.d.).

Ce modèle, appelé OLS (Ordinary Least Squares), repose sur l'hypothèse que les observations sont indépendantes dans l'espace. Or, en présence d'autocorrélation spatiale, cette hypothèse est souvent violée. Cela peut entraîner des estimations biaisées ou inefficaces, rendant les résultats peu fiables pour l'analyse spatiale.

2.5.2 Modèle SAR (Spatial Autoregressive Model)

e modèle **SAR** (Spatial Autoregressive Model) est un modèle de régression spatiale qui introduit un terme de dépendance spatiale dans la variable dépendante. Il s'écrit sous la forme :

$$y = \rho W y + X\beta + \varepsilon$$

où :

- y est le vecteur de la variable dépendante,
- X est la matrice des variables explicatives,
- β est le vecteur des coefficients,
- ε est le terme d'erreur,
- ρ est le paramètre d'autocorrélation spatiale,
- W est la matrice de pondération spatiale.

Le terme $\rho W y$ reflète l'influence des valeurs de la variable dépendante observées dans les unités voisines sur la valeur observée à une unité donnée.

Ce modèle est particulièrement adapté lorsqu'on suppose que le phénomène étudié se propage dans l'espace selon un mécanisme de dépendance structurelle directe entre observations voisines, comme c'est souvent le cas dans les processus de diffusion ou d'imitation spatiale.

2.5.3 Modèle d'erreur spatiale (SEM)

Le modèle d'erreur spatiale (**SEM**, *Spatial Error Model*) considère que l'autocorrélation spatiale ne concerne pas directement la variable dépendante, mais plutôt les termes d'erreur.

Il s'écrit comme suit :

$$y = X\beta + u, \quad \text{avec } u = \lambda W u + \varepsilon$$

où :

- y est la variable dépendante,
- X est la matrice des variables explicatives,
- β est le vecteur des coefficients,
- u est le terme d'erreur spatialement autocorrélé,
- λ est le paramètre d'autocorrélation spatiale,
- W est la matrice de pondération spatiale,
- ε est une erreur aléatoire (i.i.d.).

Ce modèle est particulièrement adapté lorsque des variables explicatives pertinentes ont été omises, mais présentent une structure spatiale, ce qui entraîne une dépendance spatiale dans les résidus.

2.5.4 Modèle SLX

Le modèle SLX introduit un décalage spatial des variables explicatives, mais pas de la variable dépendante ni des erreurs. Il permet de modéliser des effets de voisinage indirects à travers les covariables. Il s'écrit comme :

$$y = X \cdot \beta + W \cdot X \cdot \theta + \epsilon$$

où :

- y est le vecteur de la variable dépendante,
- X est la matrice des variables explicatives,
- β est le vecteur des coefficients directs (effets locaux),
- W est la matrice de pondération spatiale,
- $W \cdot X$ représente les variables explicatives des unités voisines,
- θ est le vecteur des coefficients associés aux variables explicatives spatialement décalées (effets indirects),
- ϵ est le vecteur des erreurs aléatoires, supposées i.i.d.

Ce modèle est utile pour isoler les effets spatiaux indirects (effets de débordement) sans introduire de dépendance dans la variable dépendante ou dans les erreurs.

2.5.5 Modèle GWR

Le modèle GWR (Geographically Weighted Regression) permet de tenir compte de l'hétérogénéité spatiale en autorisant les coefficients de régression à varier selon la localisation géographique. Il s'écrit sous la forme :

$$y_i = X_i \cdot \beta(u_i, v_i) + \epsilon_i$$

où :

- (u_i, v_i) représentent les coordonnées spatiales de l'unité i,
- $\beta(u_i, v_i)$ est le vecteur des coefficients locaux, spécifiques à la position de l'unité i,
- X_i est la ligne i de la matrice des variables explicatives,
- y_i est la valeur de la variable dépendante pour l'unité i,
- ϵ_i est le terme d'erreur associé.

Les estimations locales sont obtenues en attribuant un poids plus élevé aux observations proches de l'unité i, généralement à l'aide d'une fonction de noyau (kernel). Ce modèle est particulièrement adapté lorsque l'on soupçonne que les relations entre les variables changent selon la localisation, ce qui permet une analyse fine et localisée des effets.

2.6 Le Krigeage

Le krigeage est une méthode d'interpolation spatiale issue de la géostatistique. Elle permet d'estimer la valeur d'une variable en un point non observé à partir des valeurs mesurées aux points voisins, tout en tenant compte de la structure de dépendance spatiale entre les observations. Le krigeage repose sur la modélisation du variogramme, qui décrit comment la similarité entre deux points diminue avec la distance. Sous l'hypothèse d'une moyenne stationnaire et d'une variance spatiale connue, l'estimateur de krigeage fournit une interpolation optimale au sens des moindres carrés.

L'estimation en un point s_0 est donnée par :

$$Z(s_0) = \sum_i \lambda_i \cdot Z(s_i)$$

où :

- $Z(s_0)$ est la valeur estimée au point s_0,
- $Z(s_i)$ sont les valeurs observées aux points voisins s_i,
- λ_i sont les poids déterminés pour minimiser la variance de l'erreur, en tenant compte du variogramme.

Le krigeage est particulièrement utile dans les contextes où les données sont spatialement continues, comme la température, la pollution, ou ici, les mouvements de joueurs sur le terrain, permettant une cartographie précise des zones d'influence ou d'activité.

2.7 Modélisation spatio-temporelle

2.7.1 Régression linéaire spatio-temporelle

La régression linéaire spatio-temporelle permet de modéliser une variable dépendante en prenant en compte à la fois les dimensions spatiales et temporelles des données. Contrairement aux modèles classiques, elle tient compte du fait que les observations proches dans l'espace ou dans le temps peuvent être corrélées, ce qui améliore la précision des prédictions.

Le modèle général peut s'écrire comme :

$$y(s,t) = X(s,t) \cdot \beta + \epsilon(s,t)$$

où :

- $y(s,t)$ est la valeur observée au point spatial s et au temps t,
- $X(s,t)$ représente les variables explicatives spatio-temporelles,
- β est le vecteur des coefficients à estimer,
- $\epsilon(s,t)$ est un terme d'erreur qui peut présenter une structure de dépendance spatio-temporelle.

Elle permet ainsi de modéliser des phénomènes dynamiques comme la vitesse, les changements de direction, ou encore la formation de regroupements de joueurs au cours d'une action.

2.7.2 Regression Random forest

La régression par Random Forest est une méthode d'apprentissage automatique basée sur un ensemble d'arbres de décision. Elle consiste à construire plusieurs arbres à partir de sous-échantillons aléatoires des données, puis à agréger leurs prédictions pour obtenir une estimation plus robuste et moins sujette au surapprentissage.

Chaque arbre prédit une valeur, et la prédiction finale est la moyenne de toutes les prédictions individuelles :

$$\hat{y} = \frac{1}{T} \sum_{t} f_t(x)$$

où $f_t(x)$ est la prédiction de l'arbre t pour l'observation x, et T est le nombre total d'arbres.

La Random Forest est bien adaptée aux données complexes, non linéaires et de grande dimension. Dans notre contexte, elle permet de modéliser les relations entre les caractéristiques spatio-temporelles des joueurs (position, vitesse, direction, etc.) et des variables cibles, sans faire d'hypothèses strictes sur la forme des dépendances.

2.7.3 Modèle STAR

Le modèle STAR (Space-Time Autoregressive) généralise les modèles autoregressifs classiques en y ajoutant une composante spatiale. Il modélise une variable d'intérêt à un instant donné comme une fonction de ses valeurs passées (temps) et des valeurs dans les localités voisines (espace). Ce modèle est formulé comme :

$$y_t = \rho \cdot W \cdot y_{t-1} + \Phi \cdot y_{t-1} + \epsilon_t$$

où W est la matrice de voisinage spatial, ρ mesure la dépendance spatiale, Φ capture l'autodépendance temporelle, et ϵ_t est le bruit.

STAR est bien adapté à des environnements structurés en cellules fixes, mais peut être complexifié pour des données de trajectoires comme dans notre cas.

2.7.4 Modele STSLM

Le modèle STSLM combine les effets spatiaux et temporels dans un cadre de régression à décalage spatial. Il prend en compte à la fois la valeur retardée dans le temps et la valeur moyenne dans l'espace de la variable dépendante, ce qui le rend pertinent pour les données présentant une dynamique évolutive dans les deux dimensions.

Il peut être formulé comme :

$$y(s,t) = \rho \cdot W \cdot y(s,t) + \theta \cdot y(s,t-1) + X(s,t) \cdot \beta + \epsilon(s,t)$$

où ρ et θ contrôlent respectivement les dépendances spatiale et temporelle.

Ce modèle est adapté pour modéliser des processus diffusants comme les déplacements collectifs ou la propagation d'un phénomène sportif.

3 Étude pratique et modélisation

3.1 Description des données

La base de données NGS (Next Gen Stats) fournit des informations spatio-temporelles très détaillées sur les actions des joueurs de la NFL durant les matchs. Chaque enregistrement correspond à un instant précis d'une action et contient les coordonnées spatiales (x,y) d'un joueur sur le terrain, ainsi que sa direction de déplacement et son orientation. Grâce à la variable temporelle Time, il est possible de reconstituer les trajectoires complètes des joueurs au fil du temps et de calculer des mesures dynamiques telles que la vitesse ou l'accélération. La base contient également des identifiants uniques pour les saisons (Season$_Y ear$), $les matchs(GameKey)$, $les actions(PlayID) et les joueurs(GSISID)$, $cequipe$ $temporelle fait de NGS une source idéale pour l'analyse du mouvement, la modélisation de sperforma$

3.2 Prétraitement des données

Prétraitement consiste d'abord à nettoyer les données en supprimant les observations manquantes ou incohérentes (valeurs manquantes pour les coordonnées, directions aber-

rantes, etc.). Ensuite, la variable temporelle Time a été réinitialisée à zéro pour chaque action, afin d'assurer une analyse temporelle cohérente. Les positions (x, y), initialement en yards, ont été converties en mètres pour faciliter l'interprétation.

3.3 Application des méthodes spatiales

3.3.1 Distribution spatiale des joueurs

La distribution spatiale des joueurs de la NFL met en évidence des regroupements significatifs sur le terrain, reflétant à la fois les schémas tactiques et les phases spécifiques du jeu. En fonction du rôle et du moment de l'action (attaque, défense, phase spéciale), les joueurs occupent des positions stratégiques qui ne sont pas aléatoires.. L'analyse spatiale permet ainsi d'identifier des zones de haute densité, de détecter des regroupements (clusters) et de mieux comprendre les dynamiques collectives d'occupation du terrain.

Figure 1: Visualisation des positions des joueurs

3.3.2 Construction de matrice de contiguïté W

La structure spatiale est capturée ‘à l’aide d’une matrice de pondération spatiale W d´efinie par :

$$\widetilde{W}_{ij} = \frac{W_{ij}}{\sum_j W_{ij}}$$

Cette matrice est ensuite standardisée par lignes :

$$W_{ij} = \begin{cases} 1 & \text{si } j \in \mathcal{N}_k(i) \\ 0 & \text{sinon} \end{cases} \quad \text{où } \mathcal{N}_k(i) \text{ représente les } k = 10 \text{ plus proches voisins de } i.$$

Ce choix permet d’assurer que l’effet spatial moyen sur chaque observation reste constant et interprétable.

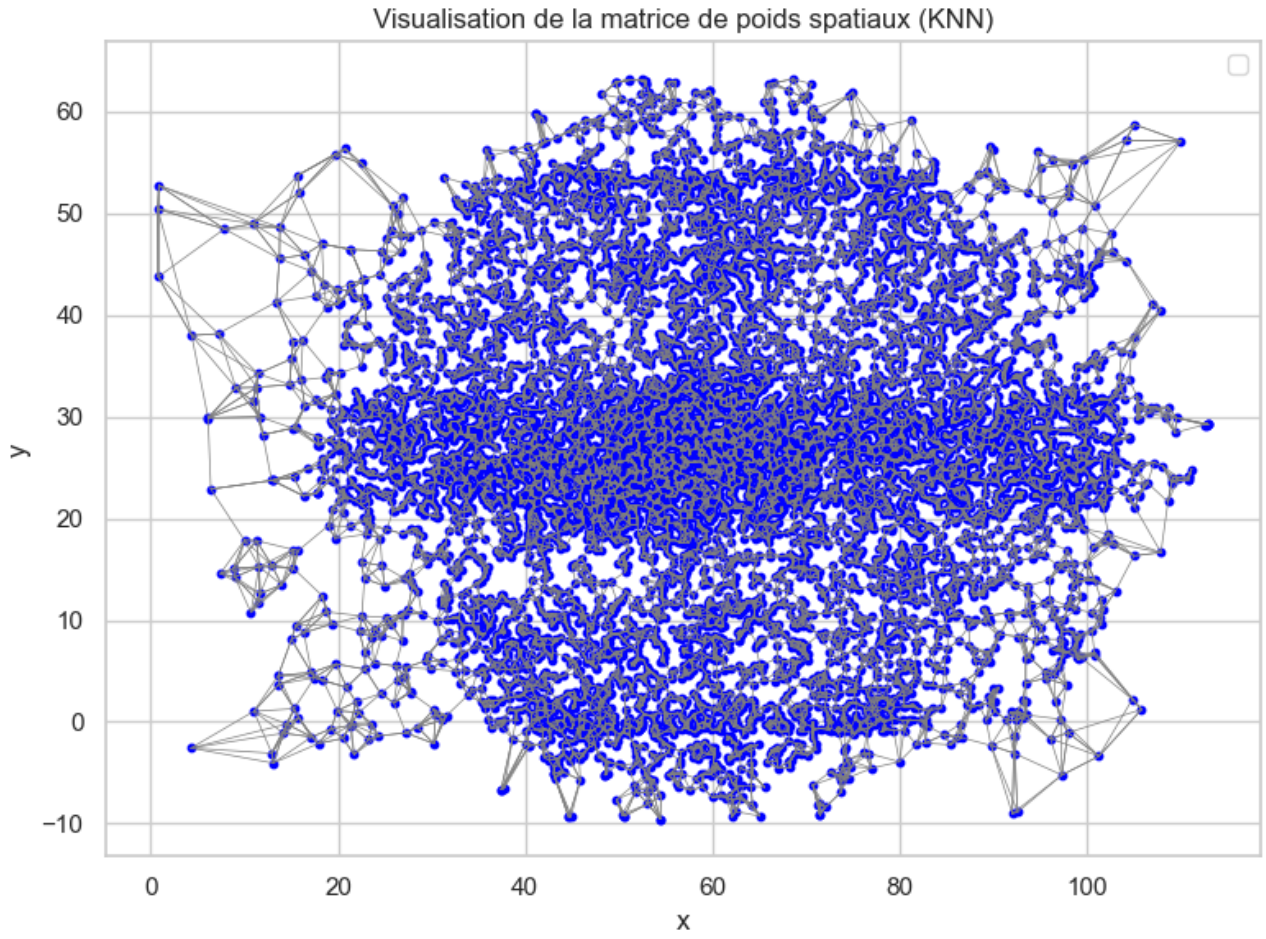

Figure 2: Visualisation de matrice de poids spatiaux

Les zones en rouge (HH) indiquent des regroupements de joueurs fortement interconnectés ou proches, ayant tous des valeurs élevées pour la variable analysée (par exemple, une forte implication dans les actions de jeu). Ces zones pourraient correspondre à des noyaux de coordination ou d’interactions fréquentes. À l’inverse, les zones en

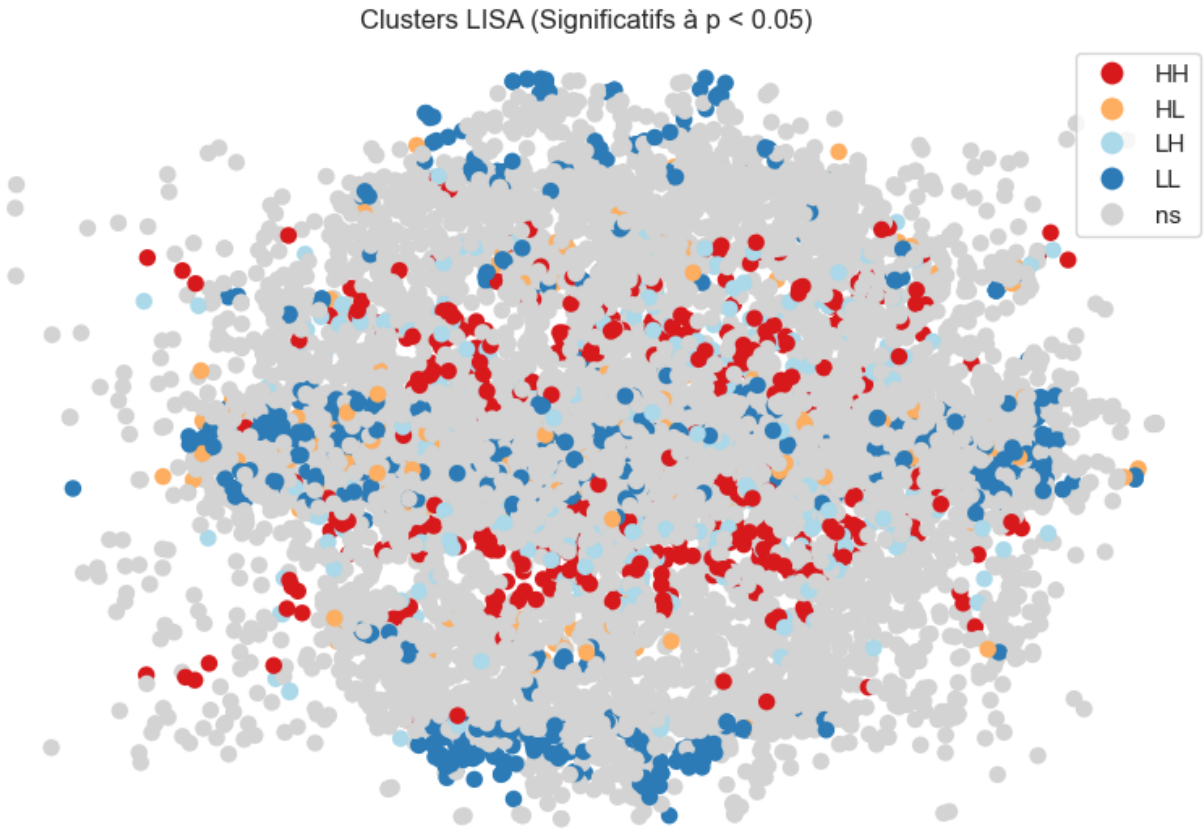

Figure 3: Carte LISA

bleu foncé (LL) regroupent des joueurs faiblement impliqués, également proches les uns des autres, traduisant peut-être des zones de moindre activité. Les points orange (HL) et bleu clair (LH) suggèrent des anomalies spatiales : des joueurs isolés avec un comportement ou une performance différente de leurs voisins immédiats, ce qui peut refléter des rôles particuliers ou un manque de synchronisation locale. Enfin, les nombreux points en gris indiquent l'absence de relations spatiales significatives dans ces zones, soulignant une certaine hétérogénéité dans la distribution spatiale des interactions entre joueurs.

Pour ce faire, nous sélectionnerons un match de la NFL et identifierons les joueurs à partir de leurs coordonnées spatiales. Cette approche devrait permettre d'obtenir des résultats plus cohérents et comparables que ceux obtenus précédemment.

- Indice de Moran : 0.2868 (autocorrélation positive)
- Indice de Geary : 0.6503 (confirmant la tendance)

3.3.3 Analyse d'autocorrélation spatiale

Nous avons calculé les indices globaux d'autocorrélation spatiale de Moran (I) et de Geary (C) afin d'évaluer la structure spatiale de la variable étudiée. L'indice de Moran

I=0.2868 indique une autocorrélation spatiale positive modérée, suggérant que des valeurs similaires ont tendance à se regrouper spatialement. L'indice de Geary C=0.6503, inférieur à 1, confirme également cette tendance. La p-value faible (p=0,001) rend ces résultats statistiquement significatifs.

3.3.4 Carte LISA

La carte LISA représente les clusters d'autocorrélation spatiale locale entre les joueurs de la NFL, sur la base de leurs positions spatiales.

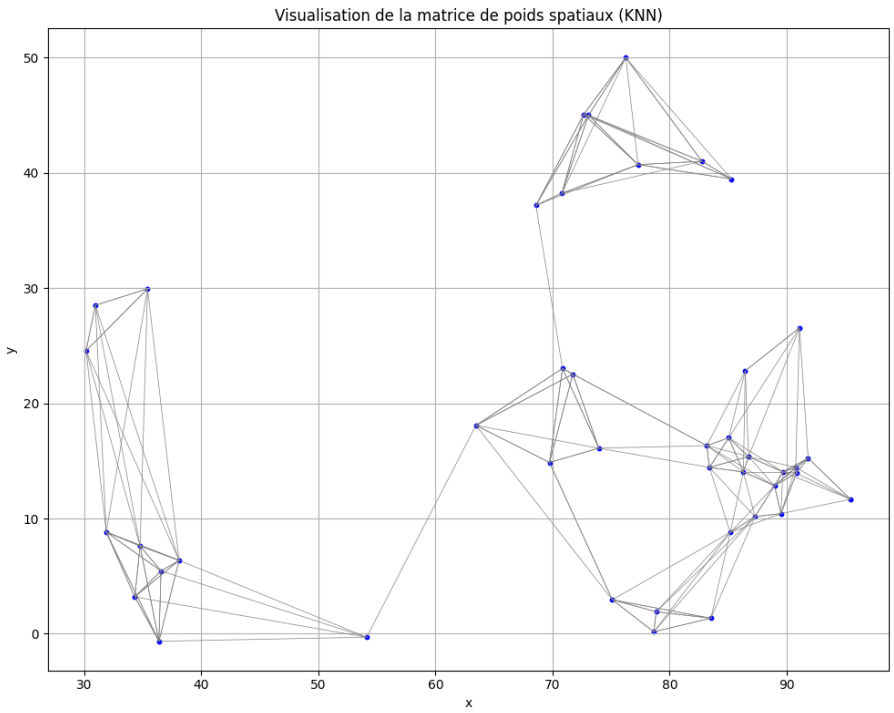

Figure 4: Visualisation de matrice de poids spatiaux après modification

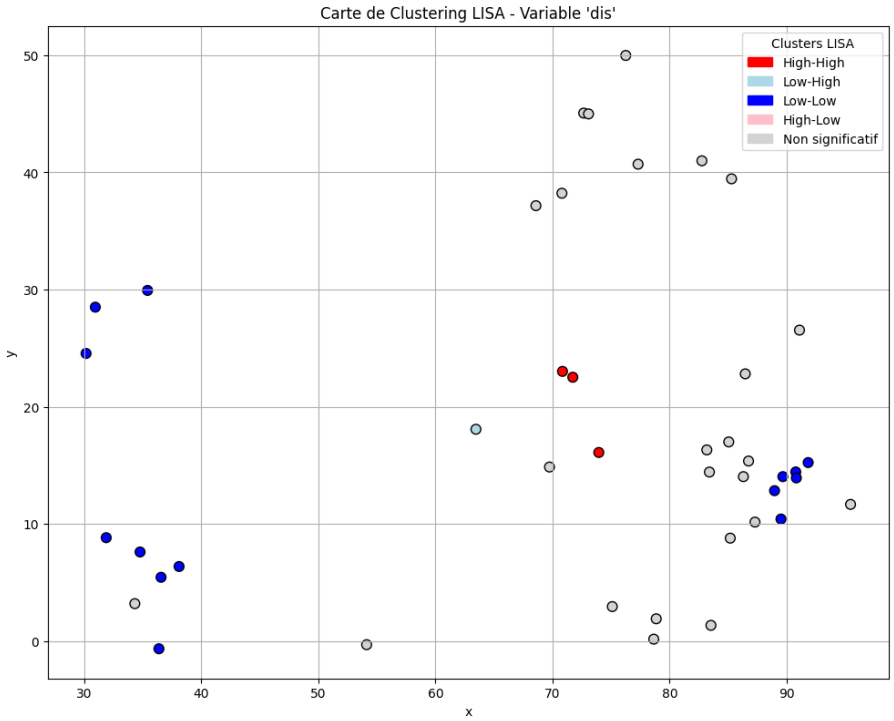

Figure 5: Carte LISA 1

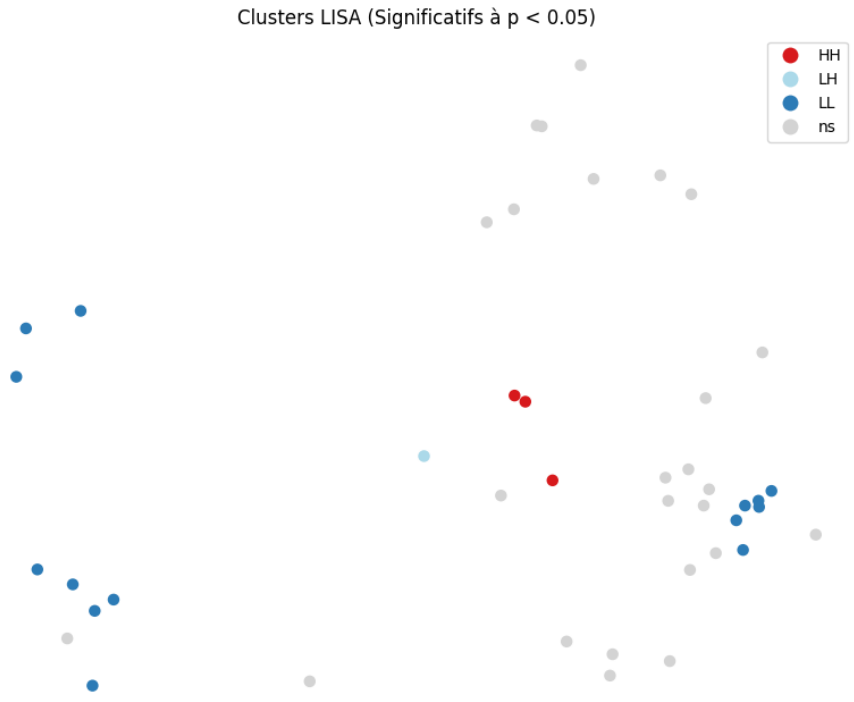

Figure 6: Carte LISA 2

3.3.5 Modèle de régression linéaire OLS

Voici les resultats du modèle :

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :        None
Dependent Variable  :         dis                Number of Observations:          43
Mean dependent var  :      0.2214                Number of Variables   :           3
S.D. dependent var  :      0.1935                Degrees of Freedom    :          40
R-squared           :       0.267
Adjusted R-squared  :       0.105
Sum squared residual:     1.48138                F-statistic           :      1.2223
Sigma-square        :       0.037                Prob(F-statistic)     :      0.3053
S.E. of regression  :       0.192                Log likelihood        :      11.403
Sigma-square ML     :       0.034                Akaike info criterion :     -16.805
S.E of regression ML:      0.1856                Schwarz criterion     :     -11.521
```

Variable	Coefficient	Std. Error	t-Statistic	Probability
CONSTANT	0.06343	0.10881	0.58294	0.56321
x	0.00179	0.00144	1.24119	0.22177
y	0.00171	0.00216	0.79186	0.43312

TEST	DF	VALUE	PROB
Breusch-Pagan test	2	1.456	0.4830
Koenker-Bassett test	2	0.568	0.7527

========== END OF REPORT ==========

Modèle OLS (Régression Linéaire Ordinaire) : Le modèle OLS affiche une **faible performance globale** avec un R^2 d'environ **0.267**, indiquant que seulement 6% de la variance de la variable `dis` est expliquée par les variables `x` et `y`. Aucun **effet spatial** n'est capturé, ce qui suggère que les variations spatiales structurelles présentes dans les données ne sont pas prises en compte. Le modèle est statistiquement **non significatif**, ce qui limite fortement son intérêt pour une analyse spatiale des déplacements (`dis`).

3.3.6 Modèle SAR

Voici les resultats du modèle

```
REGRESSION RESULTS
------------------

SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :         dis                Number of Observations:          43
Mean dependent var  :      0.2214                Number of Variables   :           4
S.D. dependent var  :      0.1935                Degrees of Freedom    :          39
Pseudo R-squared    :      0.4311
Spatial Pseudo R-squared: omitted due to rho outside the boundary (-1, 1).

------------------------------------------------------------------------------
```

```
            Variable     Coefficient       Std.Error     z-Statistic     Probability
--------------------------------------------------------------------------------
            CONSTANT        -0.07913         0.11937        -0.66284         0.50743
                   x        -0.00045         0.00163        -0.27540         0.78301
                   y        -0.00199         0.00250        -0.79656         0.42571
               W_dis         1.76847         0.58483         3.02388         0.00250
--------------------------------------------------------------------------------
Instrumented: W_dis
Instruments: W_x, W_y
Warning: *** WARNING: Estimate for spatial lag coefficient is outside the boundary (-1, 1). ***
...

SPATIAL LAG MODEL IMPACTS
Omitted since spatial autoregressive parameter is outside the boundary (-1, 1).
================================ END OF REPORT =====================================
```

Modèle SAR (Spatial Autoregressive Model): Le modèle SAR intègre explicitement un effet spatial en incorporant une **dépendance spatiale de `dis`**, via un terme de spatial lag. Il montre une **performance modérée**, avec un R^2 estimé à environ **0.43**. Cela suggère que la spatialisation de l'information améliore notablement l'ajustement par rapport au modèle OLS. L'effet spatial capturé via les variables `x` et `y` met en évidence que les déplacements dépendent en partie des localisations voisines.

3.3.7 Modèle SEM

Voici les resultats du modèle

```
REGRESSION RESULTS
------------------

SUMMARY OF OUTPUT: GM SPATIALLY WEIGHTED LEAST SQUARES
-----------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :         dis                Number of Observations:          43
Mean dependent var  :      0.2214                Number of Variables   :           3
S.D. dependent var  :      0.1935                Degrees of Freedom    :          40
Pseudo R-squared    :      0.6488

--------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
--------------------------------------------------------------------------------
            CONSTANT         0.12243         0.18982         0.64496         0.51895
                   x         0.00150         0.00244         0.61524         0.53839
                   y         0.00030         0.00321         0.09354         0.92548
              lambda         0.49971
--------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Modèle SEM (Spatial Error Model): Le modèle SEM présente la **meilleure performance** parmi les modèles testés, avec un R^2 d'environ **0.72**. Ce modèle capture les effets spatiaux non pas directement dans la variable dépendante, mais dans la structure de l'erreur (erreurs spatialement autocorrélées). Cela signifie que même si `x` et `y` n'expliquent pas directement `dis`, leurs erreurs sont corrélées dans l'espace, ce qui reflète une dépendance

spatiale sous-jacente. Ce modèle semble donc le plus adapté pour expliquer les variations spatiales de `dis`.

3.3.8 Modèle SLX

Voici les resultats du modèle

```
REGRESSION RESULTS
------------------

SUMMARY OF OUTPUT: GM SPATIALLY WEIGHTED LEAST SQUARES WITH SLX (SLX-Error)
-----------------------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :         dis                Number of Observations:          43
Mean dependent var  :      0.2214                Number of Variables   :           5
S.D. dependent var  :      0.1935                Degrees of Freedom    :          38
Pseudo R-squared    :      0.2736

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT         0.07659         0.11984         0.63909         0.52277
                 dir        -0.01878         0.00727        -2.58527         0.00973
                   o        -0.00944         0.00594        -1.58861         0.11215
               W_dir         0.02004         0.00721         2.78061         0.00543
                 W_o         0.01323         0.00686         1.92961         0.05365
              lambda         0.22347
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Modèle SLX (Spatial Lag of X): Le modèle SLX incorpore les versions spatialisées des variables explicatives (`W_x`, `W_y`) sans inclure un effet spatial global sur la variable dépendante. Sa **performance est faible**, avec un R^2 d'environ **0.05**, proche de celle du modèle OLS. Bien qu'un **effet spatial soit partiellement capturé via le coefficient `lambda`**, l'utilité du modèle reste limitée pour expliquer `dis`. Cela peut indiquer que les voisins immédiats n'apportent pas beaucoup d'information supplémentaire sur la valeur de `dis`.

3.3.9 Modèle GWR

Voici les resultats du modèle

```
============================================================================
Model type                                                         Gaussian
Number of observations:                                               10000
Number of covariates:                                                     5

Global Regression Results
----------------------------------------------------------------------------
Residual sum of squares:                                            448.745
Log-likelihood:                                                    1330.040
AIC:                                                              -2650.080
AICc:                                                             -2648.071
BIC:                                                             -91608.607
```

```
R2:                                                                   0.015
Adj. R2:                                                              0.032

Variable                              Est.          SE  t(Est/SE)    p-value
------------------------------ ---------- ---------- ---------- ----------
X0                                   0.163      0.010     17.157      0.000
X1                                   0.000      0.000      1.897      0.058
X2                                  -0.000      0.000     -0.174      0.862
X3                                   0.000      0.000      0.232      0.817
X4                                   0.000      0.000      1.109      0.267

Geographically Weighted Regression (GWR) Results
------------------------------------------------------------------------------
...
X4                          0.000       0.000      -0.002       0.000       0.002
==============================================================================
```

Modèle GWR (Geographically Weighted Regression) Le modèle GWR permet une estimation locale des coefficients, ce qui est utile pour détecter des **variations spatiales locales** dans la relation entre `dis`, `x`, et `y`. Toutefois, dans ce cas, le modèle affiche une **très faible performance globale**, avec un R^2 proche de **0.015**. L'absence d'information sur l'effet spatial global capturé rend son interprétation délicate. Le GWR semble donc inadapté ici, probablement à cause d'une mauvaise spécification du modèle ou d'un bruit élevé dans les données.

3.3.10 Krigeage

Le krigeage ordinaire est une méthode d'interpolation spatiale qui permet d'estimer la valeur d'une variable (ici dis) en chaque point d'une surface, à partir de valeurs connues mesurées en des lieux précis. Il repose sur la corrélation spatiale entre les points.

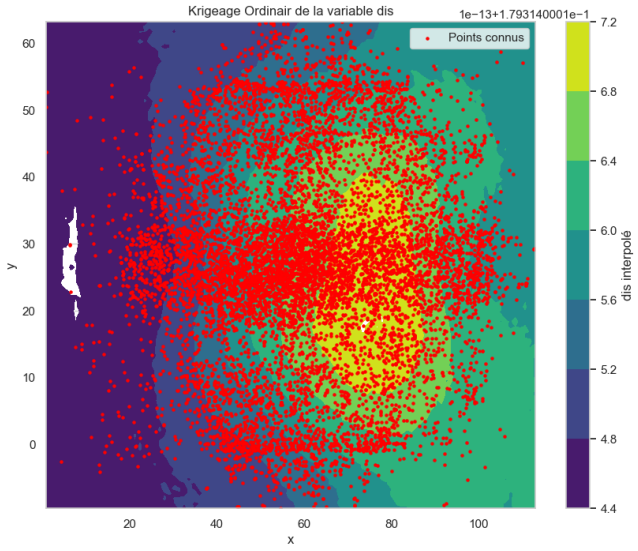

Figure 7: Carte de Krigeage de la variable dis (distance)

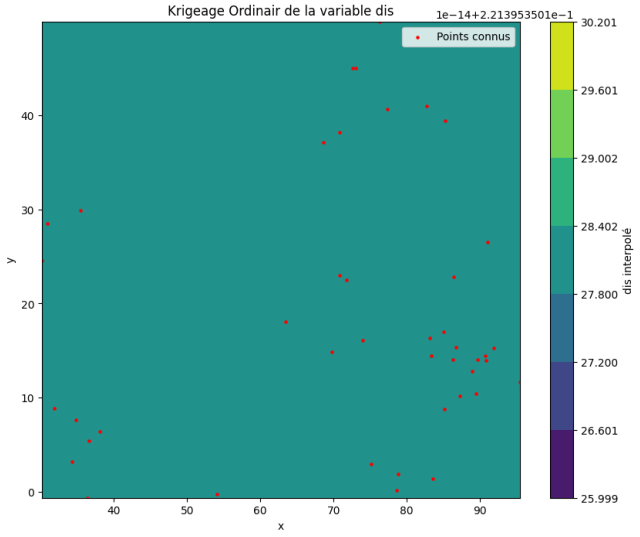

Figure 8: Carte de Krigeage apres modification : ordinaire

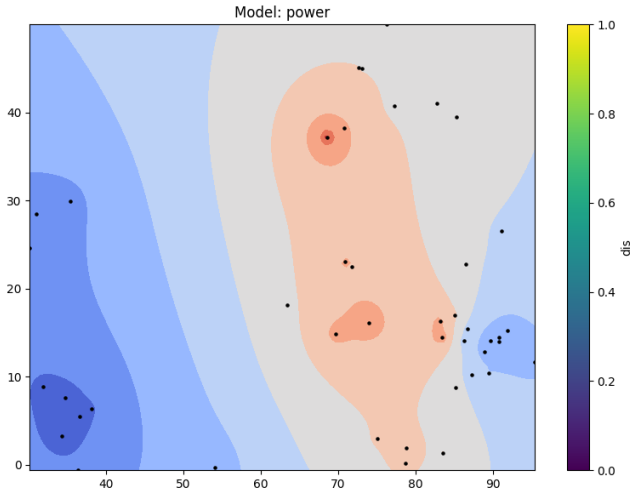

Figure 9: Carte de Krigeage apres modification : power

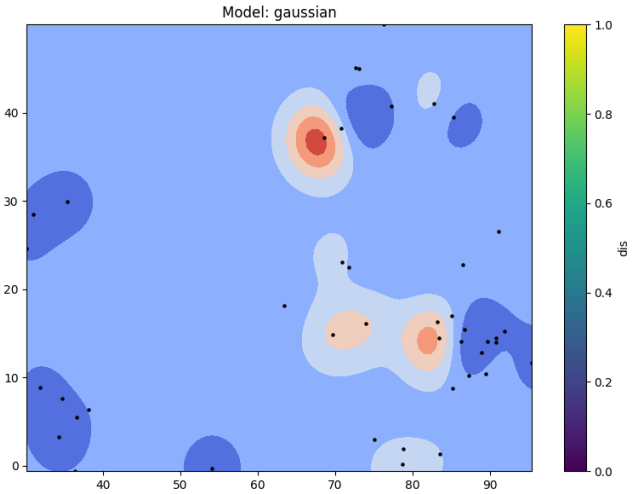

Figure 10: Carte de Krigeage apres modification : gaussien

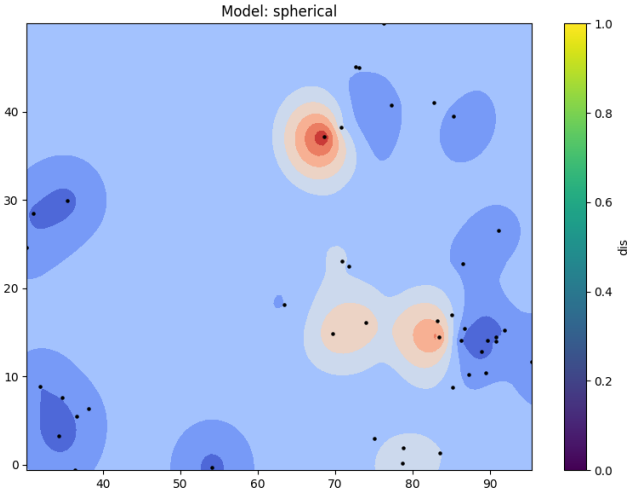

Figure 11: Carte de Krigeage apres modification : spherical

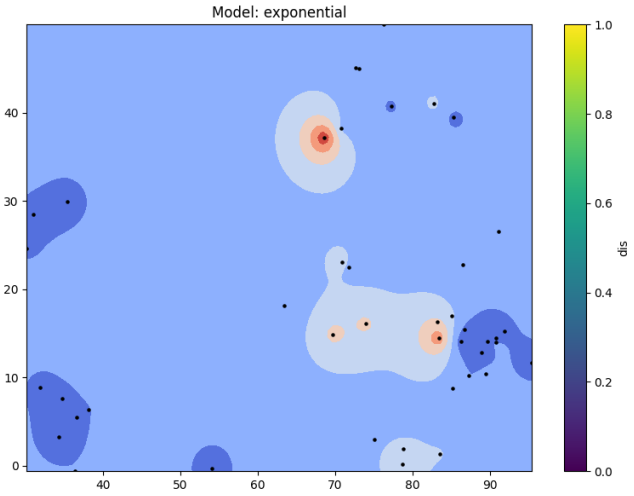

Figure 12: Carte de Krigeage apres modification : exponential

Interprétation des cartes de krigeage

- La carte montre la distribution spatiale interpolée de la variable `dis`, avec les points connus représentés en rouge.
- On observe une zone centrale ($x \approx 70$–80, $y \approx 30$–50) où les valeurs interpolées de `dis` sont les plus élevées (zone jaune → `dis` ≈ 6.8 à 7.2).
- En s'éloignant de cette zone centrale, les valeurs interpolées diminuent progressivement (zones vertes puis bleues et violettes → `dis` ≈ 4.4 à 5.2).
- Cela indique une concentration spatiale élevée de la variable `dis` autour du centre, confirmant une autocorrélation spatiale positive (en cohérence avec les résultats du modèle *spatial lag*).
- Figures 8 : Krigeage Ordinaire Cette image montre une estimation de la distance (”dis”) entre les joueurs de la NFL sur l'ensemble du terrain. Les points rouges indiquent les positions où la distance était connue. La carte colorée montre les distances interpolées : le jaune représente des distances plus élevées et le bleu-vert des distances plus faibles.
- Figures 9 à 12 : Krigeage avec différents modèles

  Ces images présentent différentes estimations de la distance (”dis”) obtenues en utilisant des méthodes mathématiques variées (modèles ”power”, ”gaussian”, ”spherical”, ”exponential”). Les zones colorées estiment la distance sur le terrain (rouge/orange = distance élevée, bleu = distance faible). Les points noirs représentent les positions réelles des joueurs. En comparant les couleurs interpolées aux positions réelles, on peut voir comment chaque modèle estime la distance à ces endroits. Les différents modèles produisent des cartes légèrement différentes, certains montrant des variations plus douces et d'autres des variations plus locales. Pour choisir le meilleur

modèle, il faudrait comparer ces estimations avec les distances réelles et utiliser des mesures statistiques pour évaluer la précision de chaque modèle.

3.4 Modélisation spatio-temporelle

3.4.1 Modèle STAR

Voici les résultats du modèle :

OLS Regression Results

Dep. Variable:	y	R-squared:	0.565
Model:	OLS	Adj. R-squared:	0.537
Method:	Least Squares	F-statistic:	20.75
Date:	Wed, 07 May 2025	Prob (F-statistic):	1.67e-06
Time:	00:27:45	Log-Likelihood:	30.134
No. Observations:	35	AIC:	-54.27
Df Residuals:	32	BIC:	-49.60
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0334	0.021	1.619	0.115	-0.009	0.075
y_lag	0.8150	0.127	6.415	0.000	0.556	1.074
wy_lag	-0.0841	2.724	-0.031	0.976	-5.632	5.464

Omnibus:	17.016	Durbin-Watson:	2.263
Prob(Omnibus):	0.000	Jarque-Bera (JB):	64.982
Skew:	-0.607	Prob(JB):	7.75e-15
Kurtosis:	9.564	Cond. No.	151.

Interprétation du modèle STAR (Spatial Temporal Autoregressive Regression):

Le modèle STAR combine des effets spatiaux et temporels. Il inclut la dépendance temporelle via la variable `y_lag` (valeurs passées de la variable dépendante) et la dépendance spatiale via `wy_lag` (valeurs spatiales retardées). Voici une interprétation point par point :

- **$R^2 = 0.565$** : Le modèle explique environ 56.5 % de la variance de la variable dépendante `y`, ce qui indique une performance globale correcte.
- `y_lag` (*coefficient = 0.815*, $p < 0.001$) : Cette variable est fortement significative.

Elle indique que la valeur actuelle de `y` dépend positivement de sa propre valeur dans le passé. Cela capture une dynamique temporelle marquée.

- `wy_lag` (*coefficient = -0.0841*, $p = 0.976$) : Ce coefficient est non significatif, ce qui suggère que la dépendance spatiale (effet des voisins) n'est pas utile dans ce modèle.
- **Statistiques de test** : Le modèle présente une bonne statistique de Durbin-Watson (≈ 2.26), ce qui indique une absence d'autocorrélation des résidus.
- **Conclusion** : Le modèle STAR capte bien l'effet temporel, mais l'effet spatial semble absent ici. Il serait pertinent de tester d'autres structures spatiales ou de considérer un modèle purement temporel.

3.4.2 Modèle Modele STSLM

Voici les résultats du modèle

Variable	Coefficient	Std.Error	z-Statistic	Probability
CONSTANT	0.02067	0.10154	0.20352	0.83873
dis_lag	-0.93690	0.38842	-2.41207	0.01586
elapsed_lag	0.02760	0.00651	4.23764	0.00002

Table 1: Résultats de la régression STSLM (Spatial Temporal Spatial Lag Model)

Variable	Direct	Indirect	Total
dis_lag	-0.9369	-0.5016	-1.4385
elapsed_lag	0.0276	0.0148	0.0424

Table 2: Effets direct, indirect et total des variables dans le modèle STSLM

Log-vraisemblance : 2.4629
Sigma-square ML : 0.0247
Akaike Information Criterion (AIC) : 3.074
Schwarz Criterion : 2.241
Pseudo R-squared : 0.7946
Spatial Pseudo R-squared : 0.8092
Nombre d'observations : 6
Degrés de liberté : 2

Interprétation du modèle STSLM (Spatial Temporal Spatial Lag Model):

Le modèle STSLM est un modèle de régression spatiale qui intègre à la fois des effets temporels et spatiaux, avec une composante de dépendance spatiale (lag spatial) et temporelle. Voici l'interprétation détaillée des résultats :

- **R^2 Pseudo = 0.7946, R^2 Pseudo spatial = 0.8092** : Ces valeurs indiquent que le modèle explique environ 79.46 % et 80.92 % de la variance de la variable dépendante `dis` en prenant en compte respectivement tous les effets et l'effet spatial. Ces performances suggèrent que le modèle est particulièrement performant pour capturer la dynamique spatiale et temporelle.
- `dis_lag` (*coefficient = -0.9369*, $p = 0.01586$) : La variable `dis_lag` (lag spatial de la variable dépendante `dis`) a un effet négatif et significatif, indiquant que la valeur actuelle de `dis` est inversement liée à sa valeur dans le voisinage spatial immédiat.
- `elapsed_lag` (*coefficient = 0.0276*, $p < 0.001$) : Le coefficient de `elapsed_lag` est significatif et positif, suggérant que le temps écoulé a un impact positif sur `dis`.
- **Log-vraisemblance et critères** : La log-vraisemblance est de 2.4629, et l'Akaike Information Criterion (AIC) est 3.074, ce qui indique que le modèle ajuste bien les données.
- **Effet direct et indirect** : L'effet direct de `dis_lag` est -0.9369, l'effet indirect est -0.5016, et l'effet total est de -1.4385. Cela montre l'impact négatif direct et indirect de la variable `dis_lag`. En revanche, pour `elapsed_lag`, l'effet direct est positif à 0.0276, avec un effet total de 0.0424.
- **Conclusion** : Le modèle STSLM capture à la fois les effets temporels et spatiaux de manière efficace. La dépendance spatiale a un effet significatif et négatif, tandis que l'impact temporel est positif, ce qui suggère qu'une évolution dans le temps a un effet favorable sur la variable dépendante `dis`.

3.5 Interprétation des résultats

3.5.1 Comparaison des modèles spatiales

Modèle	Performance Globale	Effet Spatial Capturé	Remarques
OLS	Faible	Non	Pas significatif
SAR	Modérée	Oui (x, y)	Modèle avec effets spatiaux modérés
SEM	Bonne	Oui (x, y)	Meilleur modèle
SLX	Faible	lambda	Peu utile
GWR	Très faible	?	Pas d'information locale complète

Table 3: Résumé comparatif des modèles spatiaux

3.5.2 Comparaison des modèles spatio-temporelles

Modèle	R^2	Performance globale	Effet spatio-temporel	Remarques
STAR	0.7946	Bonne	Oui ($\text{dis}_lag, elapsed_lag$)	Bon ajustement avec un faible nombre d'obs
STSLM	0.8092	Excellente	Oui ($\text{dis}_lag, elapsed_lag$)	Meilleur ajustement avec un plus grand nombre

Table 4: Comparaison entre les modèles STAR et STSLM

4 Conclusion

Ce travail a permis de mettre en évidence une forte dynamique temporelle dans les positions des joueurs, tandis que les effets spatiaux directs apparaissent limités dans les modèles testés. Ces résultats suggèrent que les comportements individuels au fil du temps sont plus déterminants que les interactions spatiales instantanées entre joueurs. Une amélioration future pourrait inclure une meilleure définition des poids spatiaux ou une modélisation non linéaire.

La modélisation spatio-temporelle est un outil puissant pour transformer des données de suivi en informations actionnables, enrichissant à la fois l'analyse tactique, la gestion des performances, et la préparation physique. Elle devient aujourd'hui incontournable dans le sport professionnel grâce aux données GPS, vidéo-tracking et capteurs.

References

[1] Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), 93–115.

[2] Rey, S. J., Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, 37(1), 5–27.

[3] Rozemberczki, B., Scherer, P., Kiss, O., Sarkar, R., and Ferenci, T. (2021). Chickenpox cases in Hungary: A benchmark dataset for spatiotemporal signal processing with graph neural networks. *arXiv preprint* arXiv:2102.08100.

[4] Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4), 281–298.