

DM - Econométrie Spatiale

Préambule.....	1
1 - Base de données.....	2
1.1 - Nettoyage de la base de données.....	2
1.2 - Choix des variables.....	3
2 - Statistiques descriptives.....	5
3 - Modèle d'économétrie spatiale.....	8
3.1 - Estimation de base avec un modèle OLS.....	8
3.2 - Construction de la matrice de poids.....	9
3.3 - Nature des interactions spatiales.....	10
3.4 - Comparaison des modèles spatiaux.....	11
4 - Interprétation des résultats et diagnostic.....	13
4.1 - Etude statistique des impacts des variables.....	13
4.1 - Cartographie des impacts.....	15
4.2 - Comparaison des impacts à Chartres contre Paris.....	17
4.3 - Conclusion de nos hypothèses.....	18
Exercice :.....	19
(a) Ecrire le modèle SDM (Spatial Durbin Modèle).....	19
(b) Ré-écrire ce modèle en utilisant le développement de la matrice inverse $(I-\rho W)^{-1}$...	19
(c) Donner la matrice d'impact associée à la i ème variable indépendante.....	20
(d) Préciser l'expression de $\partial y_i / \partial x_{jl}$. La comparer avec celle obtenue dans le modèle SAR, quelles différences remarquez-vous ?.....	20

Préambule

Cette étude vise à explorer les facteurs qui influencent l'inscription au sport en club, un engagement qui m'a personnellement touché dès mon enfance, et auquel mes parents m'ont encouragé. Nous nous intéressons ici aux éléments qui facilitent ou limitent cet accès, et nous interrogeons les dynamiques territoriales et sociales qui peuvent influencer la pratique sportive.

Pour comprendre les disparités de pratique sportive entre territoires, il est essentiel d'analyser si des individus géographiquement proches partagent des taux de licences similaires, ce qui pourrait indiquer un effet spatial dans les comportements sportifs. Par ailleurs, les infrastructures sportives jouent-elles un rôle déterminant dans la pratique ? Nous vérifierons si les territoires moins bien dotés en équipements sportifs affichent une pratique sportive plus faible.

Au-delà des infrastructures, nous investirons l'influence de la structure démographique. Existe-t-il une relation entre la proportion de jeunes dans une zone et le taux de licences sportives ? Nous chercherons également à savoir si la densité de population interagit avec la disponibilité des équipements sportifs et la proportion de jeunes pour expliquer les taux de licences, ce qui pourrait mettre en évidence des contextes territoriaux favorisant ou freinant l'accès au sport.

En croisant ces dimensions spatiales et sociales, notre objectif est d'identifier les principaux déterminants de la pratique sportive et de proposer une cartographie des disparités. À l'issue de ce projet, nous espérons ainsi mieux comprendre comment l'accès au sport est façonné par l'environnement local, et comment ces facteurs interagissent pour influencer le choix de s'engager dans une activité sportive licenciée.

1 - Base de données

1.1 - Nettoyage de la base de données

Nous disposons d'une base de données de 1 255 observations réparties en 6 colonnes, contenant le code et la géolocalisation de chaque observation. Ces données proviennent de l'Observatoire des Territoires, garantissant leur qualité. Cependant, nous avons dû les convertir au format numérique, car elles étaient initialement enregistrées sous forme de texte.

Notre étude étant centrée sur la France métropolitaine, nous avons exclu les observations relatives aux îles françaises et avons vérifié l'absence de doublons dans les codes.

Pour le fond de carte, nous avons utilisé une source de data.gouv.fr, qui recense les 2 190 EPCI (Établissements Publics de Coopération Intercommunale) de France métropolitaine.

Un EPCI est une entité administrative qui regroupe plusieurs communes pour gérer des services et projets communs. Ce fond de carte est l'un des plus détaillés disponibles, car certains EPCI sont divisés en plusieurs zones ayant des codes identiques mais des données géographiques différentes.

Nous avons relevé 24 valeurs manquantes, correspondant à des EPCI qui n'existent plus ou qui ont fusionné avec d'autres. Ces valeurs manquantes proviennent de l'Observatoire des Territoires, dont la base de données présente une granularité légèrement inférieure pour les grandes métropoles ou grands EPCI. Dans certains cas, les EPCI divisés en plusieurs zones dans nos données n'avaient qu'un code unique dans la base de l'Observatoire. Pour assurer une jointure correcte, nous avons renommé la variable 'Code' en 'CODE_EPCI' et avons dû rendre chaque code unique, car la granularité du fond de carte était donc plus faible que celle de nos données.

Nous avons regroupé ces zones fragmentées en utilisant la moyenne des données pour chaque territoire. Les lignes concernées ont été corrigées directement dans le dataframe pour assurer les correspondances finales.

1.2 - Choix des variables

J'ai sélectionné les 4 variables suivantes car avant de commencer le foot en club, je jouais souvent au city-stade de ma ville. C'est là que mes parents ont vu mon intérêt pour le sport et ont décidé de m'inscrire en club, ce qui m'a permis d'obtenir ma première licence. Tous les enfants du quartier jouaient également sur ce terrain, et beaucoup ont pris une licence dans les clubs aux alentours. Cela me fait penser que le nombre d'équipements sportifs dans une zone peut influencer le nombre de licenciés, car ces infrastructures encouragent la pratique régulière. J'ai choisi d'inclure toutes les infrastructures des différents sports, car je pense que mon expérience pourrait se retrouver chez d'autres personnes avec différents sports.

De plus, cet effet est renforcé par les interactions sociales. À force de jouer ensemble et de discuter, l'idée de s'inscrire au club se propage et plus il y a de monde, plus les enfants sont motivés à faire du sport ensemble. Cela contribue naturellement à augmenter le nombre de licenciés. En revanche, un manque de pratiquants peut entraîner la fermeture de certains clubs, ce qui montre bien qu'un plus grand nombre de pratiquants favorise la durabilité des clubs et donc le nombre de licences.

Dans les zones à forte densité de population, les habitants sont souvent exposés à un plus large éventail d'activités sportives et d'événements, ce qui encourage la pratique sportive en tout genre. J'ai également sélectionné la variable des jeunes de moins de 15 ans, car comme explicité précédemment, les parents et les écoles encouragent généralement les jeunes à pratiquer des activités sportives pour des raisons de santé, de développement des compétences sociales, et de discipline par exemple. De plus, les villes proposent souvent des programmes et installations spécifiques pour ce public. Les jeunes qui commencent tôt sont plus susceptibles de continuer dans le sport en grandissant, ce qui fait de la tranche

des moins de 15 ans une population clé pour comprendre les premières adhésions aux licences sportives.

Je soupçonne donc que ces variables aient un impact sur ma variable à expliquer, qui est le nombre de licences sportives pour 100 habitants.

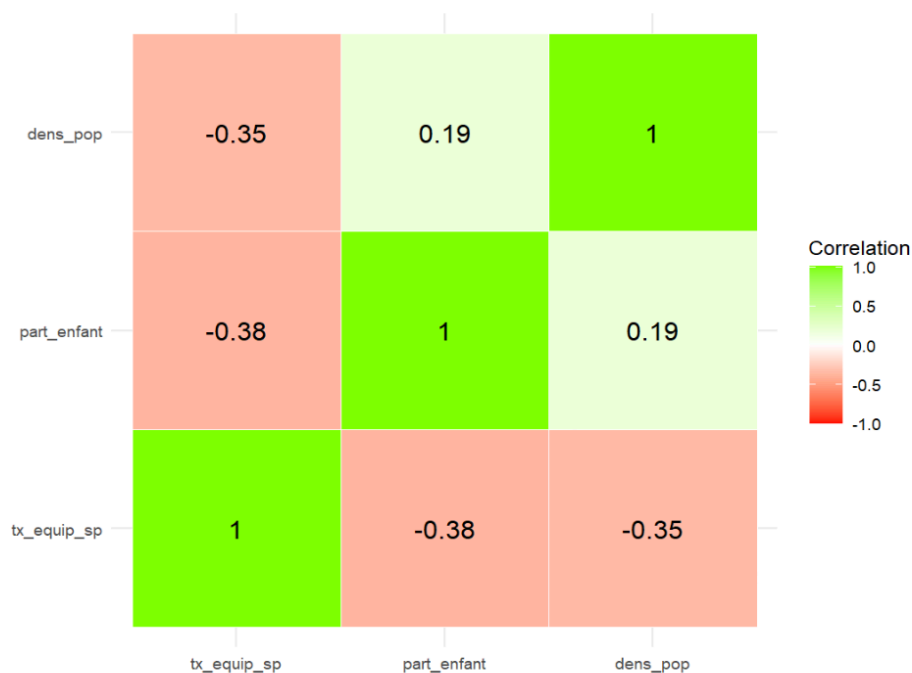
Pour résumer, J'ai choisi ces 4 variables :

Densité de population 2021	Part des moins de 15 ans 2021	Taux d'équipements sportifs pour 1 000 habitants 2023	Nombre de licenciés sportifs pour 100 habitants 2021
----------------------------	-------------------------------	---	--

Je n'ai pas inclus de variables liées au revenu, car d'après mon expérience, le revenu ne joue pas un rôle déterminant dans la décision de prendre une licence même s'il peut être pertinent pour certains sports. De plus, la granularité des données disponibles ne permettait pas d'obtenir des informations suffisamment détaillées sur cette variable.

J'ai choisi une granularité au niveau des EPCI plutôt qu'au niveau des communes ou des départements. Cela me permet d'avoir toutes les données nécessaires sans diluer l'information, comme cela aurait pu être le cas avec une granularité plus large, telle que les départements. Cette granularité est d'autant plus pertinente pour mon sujet.

Pour éviter les redondances entre les variables (celles qui seraient 100 % corrélées et apporteraient donc la même information), nous établirons une matrice de corrélations.



Nous constatons qu'il n'y a pas de corrélation parfaite ce qui est positif pour notre étude, car cela signifie que chaque variable apportera des informations uniques au modèle.

Nous avons également calculé le VIF (Variance Inflation Factor) de notre modèle, et voici les valeurs obtenues :

tx_equip_sp	part_enfant	dens_pop
1.287227	1.165813	1.141705

Ce qui signifie qu'il n'y a pas de **multicolinéarité forte** dans notre modèle, et donc que nous n'avons pas besoin de retirer ou d'ajuster nos variables explicatives en raison de ce problème.

2 - Statistiques descriptives

Pour mieux comprendre notre jeu de données, nous allons réaliser quelques statistiques univariées.

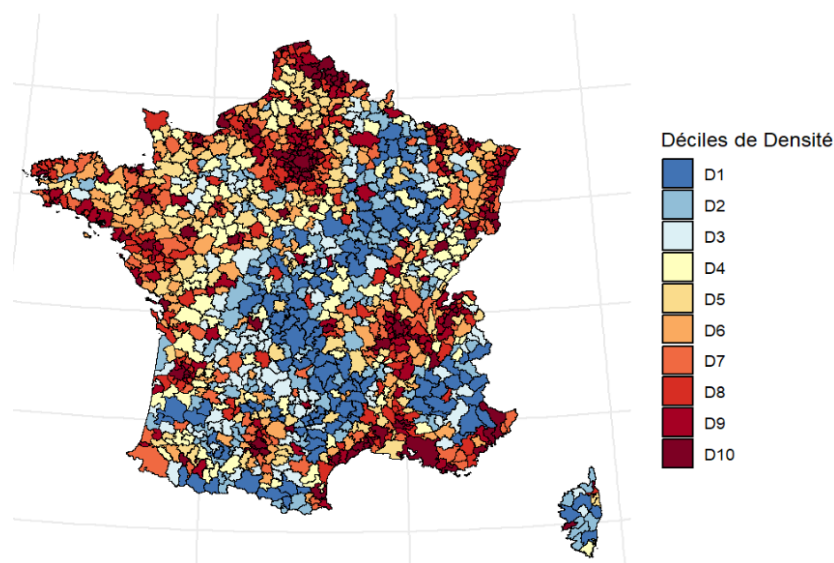
Commençons par la variable de densité de population :

Nous avons 29 EPCI qui comptent plus de 1 000 habitants par km², avec la Métropole du Grand Paris en tête affichant plus de 8 700 habitants par km². La moitié des EPCI ont une densité supérieure à 110 habitants par km², et l'autre moitié a une densité inférieure.

La moyenne, bien qu'elle puisse ne pas être très révélatrice en raison de la présence de valeurs extrêmes, est de 154 habitants par km².

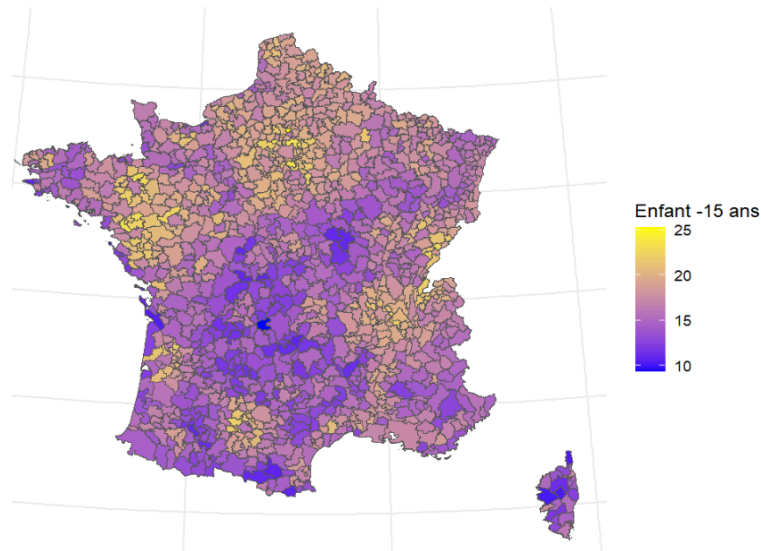
Les déciles de notre variable :

10%	20%	30%	40%	50%	60%	70%	80%	90%
19.392	26.900	36.192	48.108	63.550	86.068	116.792	162.420	305.434



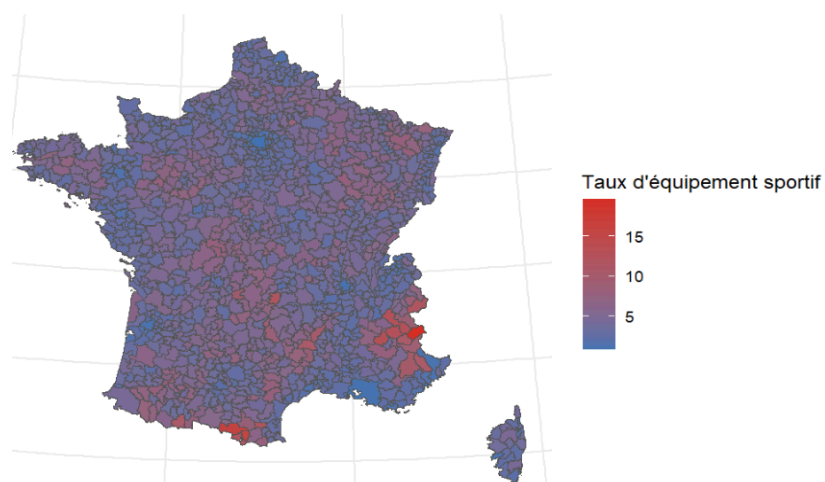
Nous parvenons à mettre en évidence la dynamique de la densité de population en France. Il apparaît clairement que Paris et la région parisienne, ainsi que les côtes méditerranéennes sont fortement peuplées. De même, les grandes métropoles et les zones frontalières avec la Suisse et la Belgique présentent également des densités élevées.

Répartition des enfants de -15 ans :



Les EPCI les plus densément peuplés d'enfants de moins de 15 ans se trouvent principalement au centre de la France, à proximité de Paris, ainsi qu'à l'ouest, avant la Bretagne. On observe également une densité importante dans les régions autour de Lyon et Toulouse. Les zones de très faible densité coïncident en grande partie avec celles où la densité de population globale est également la plus faible notamment le long de la "diagonale du vide", une région où la population semble être plus âgée. L'EPCI la plus peuplée d'enfants de moins de 15 ans avec une proportion de 25 %, est la Communauté de communes de la Vallée de Chamonix-Mont-Blanc.

Enfin, pour notre variable du taux d'équipement par tranche de 1 000 personnes, voici la répartition observée en France :



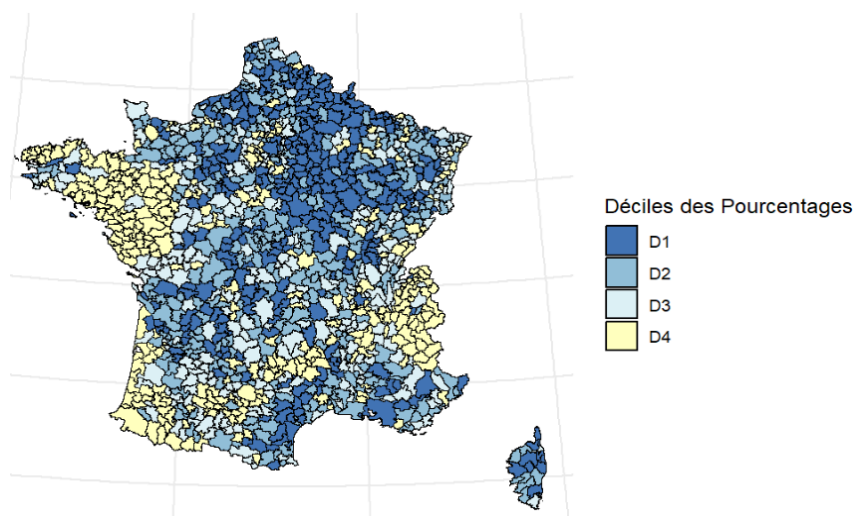
Le Sud-Est de la France semble mieux équipé pour 1 000 habitants que les autres régions. Étonnamment, Paris est moins bien représentée, probablement en raison de sa forte densité de population ce qui fait que les équipements sportifs ne peuvent pas couvrir un nombre d'habitants suffisamment élevé pour obtenir un ratio équilibré. La zone frontalière avec l'Espagne est également bien équipée. Pour le reste du territoire la répartition des équipements est assez homogène. Nous avons une moyenne de 4 équipements par ECPI.

Cette variable inclut les infrastructures permettant la pratique d'activités sportives, qu'elles soient spécialisées ou non, et qu'elles soient intérieures ou extérieures. Seules les installations accessibles au public, que ce soit individuel ou via une structure publique ou privée (associative ou commerciale), pour la pratique d'activités physiques et/ou sportives, sont prises en compte.

Passons maintenant à la variable expliquée, à savoir le pourcentage de licenciés sportifs :

Les quartiles de notre variable :

25%	50%	75%	100%
15.90	18.80	22.15	57.70



Cette carte révèle des tendances claires concernant le pourcentage de licenciés sportifs par EPCI. On remarque que la Bretagne se distingue par un taux de licenciés élevé. Le sud de la France, en particulier vers Hendaye, présente également un très bon pourcentage de sportifs suivi par le Sud-Est. En revanche, le Nord semble moins bien loti.

Le pourcentage le plus élevé, bien au-dessus de la moyenne de 18,80%, se trouve à nouveau dans l'EPCI de la Communauté de communes de la Vallée de Chamonix-Mont-Blanc, ce qui peut probablement s'expliquer par la pratique courante du ski dans cette région.

3 - Modèle d'économétrie spatiale

Dans cette troisième partie, nous adopterons une approche structurée de plusieurs étapes pour identifier le modèle approprié permettant de modéliser les relations spatiales entre les variables sélectionnées.

3.1 - Estimation de base avec un modèle OLS

Nous avons réalisé une régression linéaire multiple pour déterminer si nos données nécessitent un modèle spatial. Cela nous permettra d'obtenir une référence de base et de vérifier si la variable dépendante est correctement expliquée par les variables explicatives choisies sans prendre en compte l'effet spatial. Si le test de Moran révèle une autocorrélation spatiale significative, cela indiquerait que les erreurs résiduelles du modèle OLS sont corrélées spatialement, ce qui pourrait biaiser et rendre inefficaces les résultats du modèle OLS. Dans ce cas, cette régression classique risquerait de sous-estimer ou surestimer les coefficients, en ignorant des effets importants liés à la structure spatiale des données.

Dans notre modèle, la part des enfants est la seule variable significative avec un effet positif sur le nombre de licences sportives. Cependant, la faible qualité du modèle (R^2 de 5 %) pourrait suggérer la présence d'une dépendance spatiale non prise en compte. Les résidus du modèle présentent une dispersion importante allant de $-10,52$ à $38,38$, ce qui indique une variabilité considérable non expliquée par les variables incluses. De plus, l'écart moyen entre les prédictions et la réalité est d'environ $\pm 4,65$ ce qui reflète une précision limitée.

Pour explorer cette hypothèse, nous avons effectué un test de Moran sur les résidus. Avec une statistique de Moran I de $0,5260$ et une p -valeur extrêmement faible, nous rejetons l'hypothèse nulle d'absence d'autocorrélation spatiale. Cela signifie que les résidus ne sont pas distribués de manière aléatoire dans l'espace, alors le modèle OLS n'a pas capturé une structure spatiale sous-jacente dans les données.

Ces résultats soulignent la nécessité de recourir à un modèle de régression spatiale pour intégrer les effets spatiaux et améliorer la qualité de l'analyse.

3.2 - Construction de la matrice de poids

L'emploi d'une matrice de poids optimale, notée W , est essentiel pour assurer la pertinence des résultats dans notre futur modèle. Une matrice adéquate permet de saisir avec précision les dépendances spatiales entre les unités, en prenant en compte les mécanismes spécifiques de propagation ou d'influence spatiale, ce qui renforce la robustesse et l'interprétation des estimations. Pour déterminer la matrice W la plus appropriée, il est pertinent d'explorer plusieurs méthodes de construction :

- Approche basée sur les distances : Cette méthode repose sur un seuil r de distance maximale, permettant de définir les relations de voisinage uniquement entre les entités géographiquement proches.

- Approche fondée sur les frontières : En tenant compte des EPCI partageant une frontière commune, cette méthode saisit les interdépendances administratives ou géographiques, souvent significatives dans les contextes où les ressources ou les infrastructures sont partagées entre des territoires adjacents.
- Approche des k plus proches voisins : Cette méthode est appropriée pour modéliser les influences des unités les plus proches, sans se baser sur un seuil strict de distance ou de frontières. Elle est particulièrement utile dans des contextes où l'intensité des relations spatiales varie selon les caractéristiques spécifiques des entités.

Comparer ces trois approches permet de choisir la matrice W la plus adaptée au phénomène étudié. Cela garantit que la structure spatiale retenue reflète au mieux les mécanismes d'interdépendance en jeu, assurant ainsi une modélisation plus précise des externalités spatiales liées à nos variables.

Pour la régression avec les frontières communes, l'indice de Moran (significatif, avec une p -value très faible) est de 0,5249, ce qui suggère une autocorrélation spatiale positive modérée. Cela signifie que les zones géographiquement proches tendent à avoir des valeurs similaires pour le pourcentage de licenciés sportifs. Une conclusion similaire est obtenue avec le test de Monte-Carlo. En testant également avec les méthodes basées sur la distance et les k plus proches voisins (KNN), les trois approches révèlent des autocorrélations spatiales positives avec des résultats quasiment identiques. Autrement dit, les zones partageant des voisins (qu'il s'agisse de frontières communes, de voisinages définis par la distance ou de KNN) ont tendance à présenter des valeurs similaires pour la variable "pourcentage de licenciés sportifs".

Pour compléter mon analyse, j'ai effectué les tests LM Error et LM Lag afin de confirmer l'existence d'autocorrélation spatiale. Les résultats de ces tests révèlent des p -values extrêmement faibles donc il existe une dépendance spatiale marquée, tant au niveau des erreurs (LM Error) que dans la variable dépendante elle-même (LM Lag).

3.3 - Nature des interactions spatiales

Selon notre premier test de Lagrange, on aurait choisi de tester ces modèles car je suspecte la présence de diverses relations spatiales, chaque modèle permettant de capturer un aspect spécifique de l'autocorrélation spatiale. En explorant ces différentes options (SAR, SLX, SEM, SDM, SDEM), nous nous assurons de couvrir l'ensemble des dynamiques spatiales possibles.

Il serait pertinent de creuser davantage l'analyse de ces cinq modèles :

- Le **SAR** est un modèle approprié si on souhaite capturer les effets de contagion spatiale, où les caractéristiques d'une unité influencent directement celles de ses

voisines. Par exemple, si le taux de licenciés sportifs dans une zone géographique est influencé par celui des zones voisines, un modèle SAR peut aider à ajuster ces effets.

- Le modèle **SLX** est similaire au SAR, mais il capture la dépendance spatiale dans les variables explicatives plutôt que dans la variable dépendante. Ce modèle ajoute les valeurs des variables explicatives des zones voisines comme prédicteurs dans le modèle. Il est particulièrement utile lorsqu'il est supposé que les caractéristiques des zones voisines (et non seulement les caractéristiques de la zone elle-même) influencent le résultat de la zone centrale.
- Le **SEM** est une bonne option lorsque nous avons identifié des résidus spatialement autocorrélés, comme c'est le cas avec les tests de Moran sur les résidus dans notre modèle OLS. Il permet de modéliser l'autocorrélation spatiale dans les erreurs, ce qui permet d'améliorer la précision des estimations et de traiter les biais causés par les relations spatiales non observées.
- Le **SDM** combine les caractéristiques du modèle SAR et du SLX. Ce modèle permet de capturer à la fois l'autocorrélation spatiale dans la variable dépendante (comme le SAR) et l'autocorrélation spatiale dans les variables explicatives (comme le SLX). Il est un modèle plus flexible que le SAR et le SLX en ce sens qu'il capture à la fois les effets directs et les effets indirects de la spatialité.
- Le **SDEM** est une combinaison du SEM et du SDM. Il est utilisé lorsque l'on suspecte que les erreurs de notre modèle sont autocorrélées, mais que des influences spatiales existent également entre nos variables explicatives et notre variable dépendante.

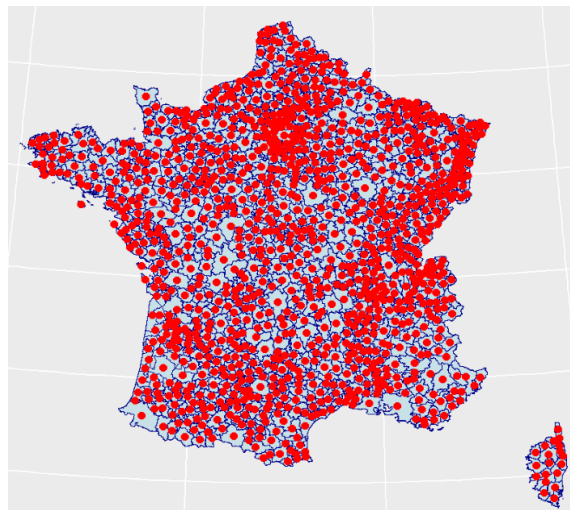
Cependant, selon la source **Anselin en 1988 dans l'ouvrage 'Spatial Regression'**, LM error et LM lag étant significatif via le test du multiplicateur de Lagrange, nous devons effectuer un test robuste de ces mêmes statistiques. Les résultats de ce test indiquent cette fois-ci que LM err robuste est significatif avec une pvalue proche de 0 donc on rejette l'hypothèse nulle du test qui indique qu'il n'y a pas de dépendance spatiale dans les erreurs après avoir pris en compte une potentielle dépendance lag. Cependant LM lag robuste n'est pas significatif et présente une pvalue de 0.1365, on ne peut donc pas rejeter l'hypothèse nulle indiquant qu'il n'y a pas de dépendance lag après avoir pris en compte la dépendance dans les erreurs. Cela indique que la dépendance spatiale est principalement dans les erreurs résiduelles et non dans la variable dépendante.

3.4 - Comparaison des modèles spatiaux

Compte tenu des relations spatiales de nos données, nous avons opté pour l'évaluation de 2 modèles spatiaux qui sont le modèle SEM et le modèle SDEM car respectivement, le SEM est un modèle qui capture la dépendance spatiale dans les erreurs résiduelles du modèle de régression, et, le SDEM capture à la fois la dépendance spatiale dans les erreurs (comme le SEM) ainsi que les effets potentiels des variables explicatives des régions voisines sur la variable dépendante. En les associant à trois types de matrices de poids définis précédemment. Les modèles seront testés avec chacune des matrices qui seront optimisées

pour chacun des modèles en testant des plages de paramètres étendues. Nous comparerons chacun des modèles spatiaux avec les différentes matrices de poids optimisées. Donc, pour sélectionner le bon modèle avec la meilleure matrice, nous procéderons à une comparaison des performances des modèles à travers la minimisation du critère d'information d'Akaike (AIC) et de la log vraisemblance. Nous aurons donc 2 modèles à comparer et 3 paramétrages différents pour chacun des modèles afin de sélectionner le modèle final. Notons que comparer les maximums de vraisemblance revient à utiliser les critères classiques de comparaison de modèles tels que AIC ou BIC. En effet, en général, la complexité du modèle reste identique.

Concernant les matrices W utilisées, nous avons testé pour chaque modèle les KNN de 1 à 30 voisins (pas plus, car la puissance de calcul était trop importante). Pour les frontières communes, nous avons utilisé la technique classique des frontières partagées par EPCI. Et pour la distance, nous avons cherché les distances optimales afin de trouver les meilleures plages à estimer pour obtenir les meilleurs résultats (et que l'ordinateur utilisé puisse suivre). Pour cela, il a fallu calculer les distances moyennes des EPCI entre elles en fonction de leur centre, que l'on peut afficher sous la représentation suivante :



Nous avons une moyenne des distances entre EPCI de 408 km et une médiane de 398 km. Les paires d'EPCI, entre les variables influençant les variables à la hausse, restent cohérentes pour tout le territoire français. Nous supposons que les relations pertinentes se situent entre voisins proches (également par l'étude des frontières communes). Nous allons donc explorer des distances plus petites et tester progressivement des seuils de 10 à 50 km (on pourrait tester un éventail plus large, mais ma machine ne le permet pas).

Après avoir testé tous nos modèles, voici ce qu'il en ressort :

Modèle	Configuration	Log-vraisemblance
SEM	Distance 40,000	-3317.972
	KNN = 7	-3305.287
	Frontière commune	-3308.099
SDEM	Distance 40,000	-3308.42
	KNN = 7	-3288.884
	Frontière commune	-3299.868

La problématique étant maintenant de départager nos modèles basés sur la matrice W construite avec les KNN et optimale avec 7 voisins. Nous allons effectuer le test du rapport de vraisemblance, qui permet de déterminer si l'ajout de paramètres ou de complexité dans un modèle améliore significativement son ajustement aux données.

Ainsi nous trouvons que la statistique de test étant égale à $LR = -2(-3305.287 - (-3288.884)) = 32.806$, et que la stat du khi deux à 5% pour 1 ddl est 3.84, alors LR est supérieur à la valeur critique du khi-deux, donc on rejette l'hypothèse nulle et on conclut que le modèle SDEM améliore significativement la vraisemblance par rapport au modèle SEM ce qui justifie l'ajout des termes supplémentaires (influence spatiale des variables indépendantes via WX). On dit qu'il considère à la fois l'influence spatiale des variables indépendantes et des résidus.

Nous aurions le modèle sous la forme suivante :

$$Y = \rho W_1 Y + X\beta + W_2 X\delta + v, \text{ avec } v = \lambda W v + \varepsilon \text{ et } \varepsilon \sim N(0, \sigma^2 I_n)$$

Avec pour chaque terme :

- **Autocorrélation spatiale de y** : $\rho W_1 Y$
- **Effets spatiaux des variables explicatives** : $W_2 X\delta$
- **Effets directs des variables locales** : $X\beta$

4 - Interprétation des résultats et diagnostic

4.1 - Etude statistique des impacts des variables

Afin d'analyser notre modèle SDEM, il est crucial d'examiner les impacts directs, indirects et totaux de nos variables explicatives. Ici, on a calculé l'impact moyen à l'aide de la fonction impacts(), qui donne un aperçu général de l'effet de chaque variable explicative. Ces impacts permettent de comprendre non seulement l'effet immédiat (direct) des variables sur la variable dépendante, mais aussi les effets diffusés à travers les entités spatiales voisines (indirects) et l'effet global combiné (total).

Les statistiques du modèle et de ces coefficients révèlent donc différentes informations. Dans un premier temps, ce qui est rassurant, c'est que le modèle SDEM montre une

amélioration significative par rapport au modèle OLS en terme de critères d'information AIC avec 6595.8 comparé à AIC de 7295.4 pour l'OLS. Par le fait, le paramètre Lambda (0.79) qui est le paramètre spatial d'erreur indiquant la présence d'autocorrélation spatiale positive dans les résidus du modèle est significatif.

Table 1: Mesure des Impacts (SDEM)

Variable	Direct		Indirect		Total	
Estimates						
dens_pop	−0.002 014	***	0.004 989	***	0.002 975	**
part_enfant	0.344 216	***	0.418 365	**	0.762 581	***
tx_equip_sp	0.023 791		0.528 415	*	0.552 206	*
Standard Errors						
dens_pop	0.000 328		0.001 001		0.001 038	
part_enfant	0.056 178		0.149 528		0.163 523	
tx_equip_sp	0.075 919		0.213 584		0.235 719	

Note :

*** significatif à 1%, ** significatif à 5%, * significatif à 10%.

Il convient de noter que l'intercept n'est pas significatif, mais cela ne constitue pas un problème car notre modèle est bien ajusté. Cette absence de significativité pourrait être due à la présence de variables omises.

En ce qui concerne **dens_pop**, l'impact direct est négatif et significatif indiquant qu'une densité de population plus élevée dans une zone entraîne une diminution du taux de licenciés sportifs. Cependant, l'impact indirect via les zones voisines est positif et significatif, indiquant que la densité de population des régions avoisinantes a un effet favorable sur la pratique sportive. L'impact total, combinant les effets directs et indirects reste positif et significatif soulignant l'importance des influences locales et de voisinage.

Pour **part_enfant**, l'impact direct est fortement positif et significatif, ce qui signifie qu'une augmentation de la proportion de jeunes dans une zone est associée à une forte augmentation du taux de licenciés sportifs. De plus, l'impact indirect des jeunes dans les zones voisines est également positif et significatif renforçant l'idée que la présence de jeunes dans une région influence positivement la pratique sportive dans les zones voisines. L'impact total est d'autant plus élevé, mettant en lumière l'influence importante de la proportion de jeunes, tant au niveau local qu'à travers les territoires voisins.

Concernant **tx_equip_sp**, bien que l'impact direct des équipements sportifs sur le taux de licenciés sportifs ne soit **pas significatif**, l'impact indirect à travers les zones voisines l'est, indiquant que l'amélioration des équipements sportifs dans les régions avoisinantes joue un rôle crucial dans l'augmentation de la pratique sportive. L'impact total est également significatif soulignant l'importance des infrastructures sportives dans les zones voisines pour favoriser la pratique sportive ce qui paraît logique si nous devons pratiquer un sport, il est probable que les infrastructures soit dans des territoires voisins.

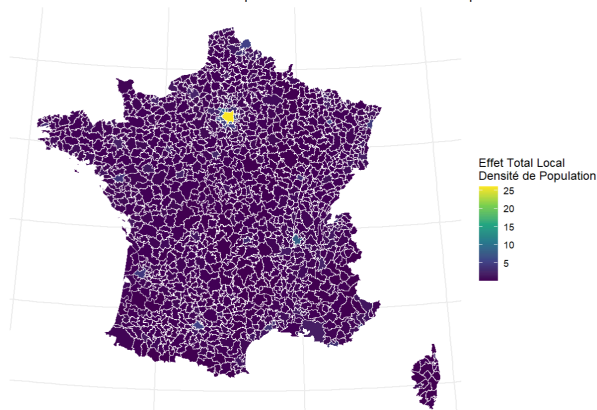
Alors, bien que **tx_equip_sp** n'ait pas d'impact direct significatif, son effet indirect et total est notable et montre l'importance des dynamiques spatiales. Les résultats suggèrent que la densité de population, la proportion de jeunes et la disponibilité des équipements sportifs dans les zones voisines influencent de manière significative l'inscription au sport en club. Ces conclusions renforcent l'idée que les politiques sportives et les décisions en matière d'infrastructure doivent tenir compte de ces interactions spatiales et démographiques pour être plus efficaces et favoriser une plus grande participation à la pratique sportive.

4.1 - Cartographie des impacts

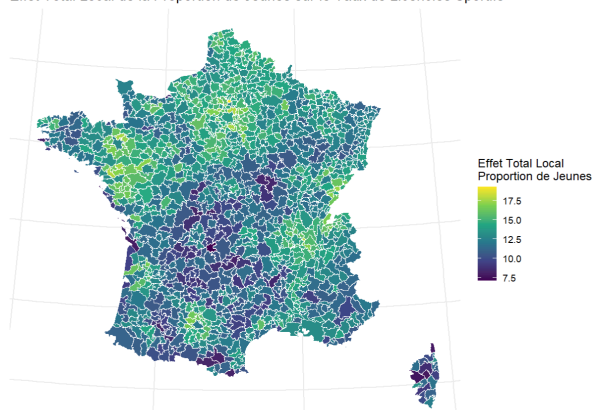
Nous avons maintenant l'idée que nos variables seules et combinées impactent en moyenne positivement le taux de licenciés sportifs. Pour mieux se rendre compte de l'impact que chaque variable a sur une EPCI, nous pourrions les cartographier afin de visualiser en détail l'impact local associé.

Les impacts totaux de chaque variable explicative peuvent être visualisés ainsi :

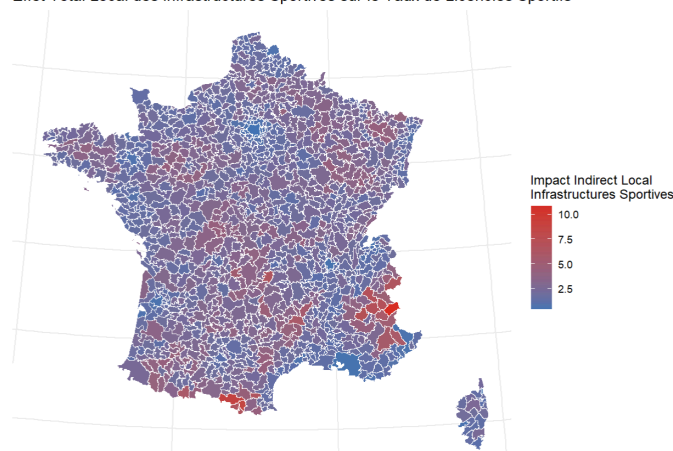
Effet Total Local de la Densité de Population sur le Taux de Licenciés Sportifs



Effet Total Local de la Proportion de Jeunes sur le Taux de Licenciés Sportifs



Effet Total Local des Infrastructures Sportives sur le Taux de Licenciés Sportifs



L'effet total montre l'influence globale de la variable explicative, intégrant comment une augmentation dans une zone et ses voisines affecte le taux de licenciés sportifs. Un effet

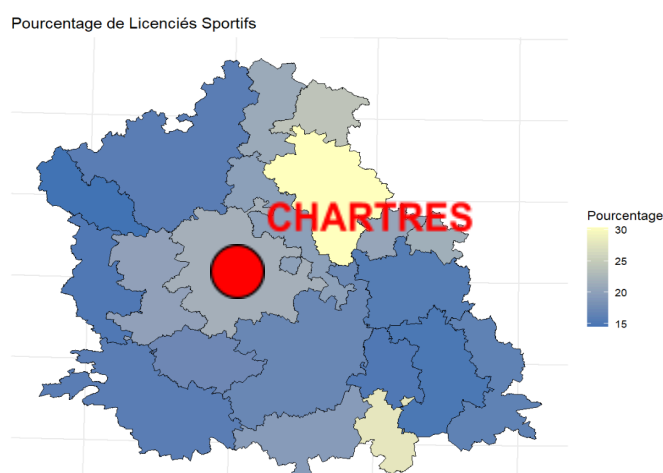
total positif signifie qu'une densité plus élevée, localement et dans les zones voisines, favorise l'inscription sportive (dans le cas où les effets sont significatifs).

Par exemple, pour la densité de population, on voit que Paris et ses voisins se démarquent largement avec un effet de densité de population qui augmente considérablement le taux de licenciés sportifs. Elle a un impact modéré pour le reste de la France hormis les très grandes villes, ce qui confirme nos hypothèses.

Concernant l'impact de la part d'individus de moins de 15 ans, cela se trouve avoir le plus d'impact plutôt dans les grandes villes également. Notons que sur la diagonale du vide, la part des enfants de -15 ans est réduite, c'est aussi ici qu'on en trouvait le moins ce qui suggère une population plus âgée et donc plus active entre elles dans ces zones. Pour le taux d'équipement sportif, nous avons le même constat qu'au endroits où le taux d'équipements le plus important se trouve être également les endroits avec les taux de licenciés sportif les plus importants.

Dans le modèle SDEM, l'impact d'une augmentation d'une unité de la variable explicative **X** sur la variable dépendante **Y** pour une observation donnée est comparable à celui d'un modèle de régression classique, mais il peut être ajusté en tenant compte de la dépendance spatiale présente dans les erreurs. Cependant, une augmentation de la variable explicative pour une observation donnée n'entraîne pas directement une modification de la variable dépendante des voisins.

Petit, étant convaincu que les joueurs et le city-stade à côté de chez ont motivé mes parents et moi à m'inscrire au football, nous allons étudier les impacts locaux des différentes variables dans la ville où j'ai grandi, Chartres.



L'impact total de la densité de population est **0.474**. Cela signifie qu'en combinant les effets directs et indirects, une augmentation d'une unité de densité de population entraîne une augmentation nette de **0.47 licenciés pour 100 habitants** dans Chartres. Bien que l'effet direct de la densité soit négatif, l'effet indirect (en raison des voisins) compense en partie cet effet, entraînant un impact global positif, ce qui pourrait suggérer que les effets de voisinage contrebalancent les effets négatifs locaux.

L'impact total de la part des moins de 15 ans est **14.11**. Une augmentation d'une unité dans la proportion des jeunes dans Chartres et ses voisins entraînera une augmentation nette de **14.11 licenciés pour 100 habitants**. Ce résultat met en évidence l'importance des jeunes dans la dynamique sportive locale et régionale, suggérant que des politiques en faveur des jeunes peuvent avoir un impact significatif sur la pratique sportive à la fois localement et dans les zones voisines.

Enfin, l'impact total des équipements sportifs est **1.71**. Cela signifie qu'une augmentation du taux d'équipements sportifs dans Chartres et ses voisins entraîne une augmentation nette de **1.71 licenciés pour 100 habitants**. Bien que l'impact direct de cette variable ne soit pas différent de 0, les effets indirects des équipements voisins ont un effet positif sur la pratique sportive à Chartres, ce qui montre que les équipements sportifs de proximité jouent un rôle important dans l'augmentation des licenciés sportifs, même si cet effet se fait via les voisins.

4.2 - Comparaison des impacts à Chartres contre Paris

En moyenne, toutes nos variables tirent le nombre de licenciés sportif vers le haut, mais ça n'est pas le cas partout. Par exemple, si nous devons comparer nos résultats à Paris, aurions-nous le même constat ?

À Chartres, la densité de population a un impact total légèrement positif (0.474), ce qui est aussi le cas à Paris. Cependant, l'impact direct est très négatif (-17.564) en raison de la saturation liée à l'hyper-densité, mais l'impact indirect est très élevé (43.514), suggérant que malgré les défis d'accès, la densité de population favorise les interactions sociales et l'engagement sportif dans différentes zones voisines potentiellement plus accessibles, résultant en un effet total positif (25.950).

La proportion d'enfants à Chartres (18,5 %) contribue à un impact total légèrement supérieur (14.108 contre 13.726).

En ce qui concerne les équipements sportifs, Chartres, avec un taux de 3.1 équipements pour 1 000 habitants, génère un impact indirect et total plus élevé (1.638 et 1.712) que Paris, où un taux plus faible (0.8 pour 1 000 habitants) limite ces effets. Ces résultats montrent que la densité d'équipements est un facteur clé pour les zones moins peuplées comme Chartres, tandis qu'à Paris, la saturation atténue l'effet des équipements. Ainsi, les équipements jouent un rôle significatif à Chartres.

Finalement, **Chartres** tire mieux parti de ses conditions locales (faible densité, part importante d'enfants, meilleure densité d'équipements) pour maximiser la pratique sportive. Alors que **Paris**, bien que plus dense et avec une forte proportion d'enfants, souffre d'un manque d'équipements adaptés à sa population ce qui réduit son potentiel sportif malgré des impacts indirects élevés liés à la densité sociale.

4.3 - Conclusion de nos hypothèses

Cette étude avait pour ambition d'explorer les déterminants de la pratique sportive en club à travers une analyse territoriale et sociale confrontant nos hypothèses initiales aux résultats obtenus. Nous avons émis les hypothèses suivantes :

L'effet de la densité de population semble trop faible pour influencer réellement l'accès au football en club, et peut-être même n'affecte-t-il pas de manière significative le nombre de licenciés sportifs. Nous pensions que la densité de population dans une zone serait un facteur déterminant dans le taux de licenciés sportifs. Cependant, les résultats suggèrent que ce n'est pas le cas. Nous avons supposé que l'hyperdensité favorise la pratique sportive en créant une plus grande concentration d'athlètes potentiels. De plus, son impact direct est négatif, car la ville est relativement petite, et une densité trop élevée entraînerait probablement un effet similaire à celui observé à Paris, avec une saturation des infrastructures.

Cependant, plusieurs facteurs viennent limiter cette influence, tels que la saturation des infrastructures, les contraintes liées à l'urbanisation, et la répartition inégale des équipements, rendant notre hypothèse difficile à confirmer.

En revanche, à Chartres, la proportion d'enfants semble avoir un impact plus marqué, probablement en raison d'une meilleure accessibilité aux équipements et aux clubs, ce qui facilite l'engagement sportif. Nous pensions que la présence d'un grand nombre d'enfants dans une zone géographique favorise la prise de licence sportive. Cette hypothèse se confirme dans le cas de Chartres (notre cas), où une forte proportion d'enfants facilite la pratique. **Mon expérience personnelle** soutient cette idée : c'est véritablement grâce à mon entourage et aux autres enfants du quartier que je me suis inscrit au foot, ce qui confirme l'importance de l'environnement social dans l'engagement sportif.

Enfin, à Chartres, la répartition plus équitable des équipements a un effet significatif, car les infrastructures sont moins saturées, permettant ainsi une utilisation optimale et libre. Nous pensions que la densité d'équipements jouerait un rôle majeur dans la pratique sportive. Cela se révèle être le cas dans les zones moins densément peuplées comme chez moi, où la meilleure disponibilité des équipements permet une meilleure pratique. Cela démontre que la disponibilité et la qualité des infrastructures sont des facteurs cruciaux pour encourager la pratique sportive, particulièrement dans les zones moins urbanisées.