Statistique Exploratoire Spatiale : Synthèse et Réflexions

Cours ENSAE ISE1 Cycle Long, 2025–2026
Par David NGUEAJIO, à l'issue d'un cours dispensé par M. Hema Aboubacar

Introduction

La statistique exploratoire spatiale (SES) représente un paradigme fondamental des analyses géographiques et statistiques contemporaine. En intégrant explicitement la dimension spatiale comme élément structurant de la variation des phénomènes, elle transcende l'approche statistique classique . Elle n'est cependant pas à confondre avec l'économétrie spatiale, proche de l'inférence traditionnelle qui suppose souvent un modèle a priori. Ici, le but de l'approche exploratoire est de faire parler les données afin qu'elles révèlent leurs structures cachées.

Ce résumé synthétise les enseignements reçus et les travaux pratiques réalisées lors du cours de SES dispensé par M. Hema Aboubacar à l'Ecole Nationale de la Statistique et de l'Analyse Economique Pierre NDIAYE de Dakar durant le semestre 1 de l'année académique 2025-2026 aux élèves d'ISE-Cycle Long. Il suit volontairement une progression thématique plutôt que chronologique, afin de mieux rendre compte de l'interdépendance du savoir diffusé dans chaque "compartiment" du dépôt github.

1 Fondements et Enjeux Méthodologiques

Au cours d'une conférence de 1969 Waldo Tobler énonce une proposition qui publiée en 1970 deviendrait la première loi de la géographie :"everything is related to everything else, but near things are more related than distant things". Ces propos suggèrent que les observations géographiquement proches tendent à être similaires. Ce serait donc se privée de précieuses informations que d'ignorer le facteur géographique dans le processus statistiques d'exploitations des données. Cet enjeu ne devint particulièrement visible vers la fin du cours. Dans l'analyse des prix immobiliers de **Baltimore (1978)**, la distribution des 211 ventes de maisons révèla un clustering spatial manifeste : les quartiers centraux et périphériques forment des clusters de prix. De même, l'étude des **quartiers de Columbus, Ohio** (49 entités administratives) indiquait que le crime et le revenu n'étaient pas distribués aléatoirement mais agrégés spatialement, suggérant des mécanismes de *contagion* ou de *concentration* géographique.

Le contexte étant posé, nous pouvons alors définir les statistiques exploratoires spatiales comme l'ensemble des méthodes et techniques permettant : - d'explorer les structures spatiales cachées dans les données géographiques

- de visualiser la distribution et les patterns spatiaux
- de quantifier l'autocorrélation spatiale (dépendance entre lieux proches)
- de générer des hypothèses sur les mécanismes géographiques sans imposer a priori un modèle.

Elles sont stratégique en ceci qu'elles peuvent permettre d'abritrer sur dès thématiques décisionnelles récurrentes, telles que l'allocation des ressources (Au cours du TP3 par exemple, l'analyse de l'accessibilité des services sociaux permettraient d'identifier les zones où implanter des infrastructures sociales aurait l'impact maximum), ou encore la prévention et les prédictions (au cours de l'examen,

il fallait à un moment analyser les corrélations entre précipitations, et rendements agricoles. L'on aurait pu aller plus loin en intégrant les indices spectraux et ainsi détecter des risques de famines, potentiellement liée à des risque de de conflits).

Enfin, il faut différencier SES d'économétrie spatiale en ceci qu'elles constituent plutôt la phase exploratoire qui permettent de générer des hypothèses et d'observer des motifs. Ce sont ses résultats qui doivent nourrir la construction des modèles et non l'inverse.

2 Méthodes computationelles : Traitement local vs cloud

3 Données Spatiales : Types, Sources et Limites

L'on distingue deux grands types de données spatiales, à savoir les données dîtes vectorielles et celles dîtes rasters.

Les données vectorielles sont des représentation d'objets géographiques par des géométries discrètes : points (comme les hopitaux dans le TP3), lignes (comme les routes et les rivières), polygones (comme les limites admnistratives tout le long du cours). Elles dominent les analyses administratives. Dans les **TPs individuels (TP1)**, la manipulation de shapefiles GADM (Angola, Côte d'Ivoire) et de données OpenStreetMap (OSM) pour le Cameroun illustre les limites pratiques : OSM offre couverture granulaire mais qualité inégale selon régions africaines ; GADM fournit structures administratives harmonisées mais potentiellement obsolètes. La transition moderne vers GeoPackage (.gpkg), résout le problème de fragmentations des shapesfiles (.shp+ .shx + .dbf + ...) en utilisant un seul fichier au lieu de 4–6) mais elle suppose une infrastructure logicielle compatible. (L'utilisation de geopackages dans GEE pourraient être un axe futur de travail dans les prochains cours).

Les données raster s'avèrent indispensables dès qu'on s'intéresse à la densité ou aux variables continues. La **densité de population WorldPop**, utilisée dans tous les TPs, offre granularité inégalée pour l'analyse d'accessibilité. Dans le **TP3)**, l'intersection raster population avec polygones hôpitaux (via extraction zônale) produit estimations de population desservie—impossible par données administratives seules. Les données satellitaires (Landsat, Sentinel-2, MODIS) révèlent une mutation technologique majeure. Le **TP4 (Ethiopie - Identification Terres Arables)** fusionne les indices de végétations, les MNT (pentes), la couverture forestière Hansen, et les zones protégées vectorielles pour délimiter terres réellement cultivables. Cette approche multi-source—impossible localement il y a 10 ans—est rendue viable par Google Earth Engine (GEE). Le cloud computing distribue calculs massifs sans upload/téléchargement.

Fusion Vecteur + Raster : le cas des données hybrides

Problème : Comment combiner données vectorielles discrètes (limites administratives, zones protégées) avec données raster continues (densité de population, NDVI satellitaire) ? Nous avons rencontré cette question dans trois contextes :
(i) limites administratives (vecteur/polygones) fusionnées avec densité population (raster/grille) ;
(ii) réseau routier (vecteur/lignes) superposé NDVI satellite (raster) ;
(iii) zones protégées (vecteur) croisées classification occupation du sol (raster).
Trois solutions opérationnelles :
L' *Extraction zônale* : qui résume les statistiques rasters dans les polygones. Par exemple prendre la population totale par commune agrégée depuis une grille raster de WorldPop.
La *Rasterisation* : convertit vecteur en raster pour analyses uniformes—utile quand logique raster domine.
La *Vectorisation* : qui transforme raster en polygones pour exploiter topologie vectorielle.

4 Autocorrélation Spatiale et Quantification

Mesurer explicitement la dépendance spatiale constitue le cœur de la statistique exploratoire. L'**indice de Moran** formalise ce concept : valeur unique résumant si valeurs proches sont similaires (positive) ou opposées (négative).

Le cours démontre l'importance de définir le voisinage. Trois approches structurent les TPs : la contiguïté topologique (queen : frontière partagée, y compris diagonales), k-plus proches voisins (kNN : régularité garantie), seuil de distance (flexibilité). Dans le cas de l'étude de la Caroline du Nord (une centaine de comtés), la contiguïté Queen a révèlé une moyenne 5 voisins par comté, structure variable (coins = moins de voisins). Cette variation affecte directement résultats spatiaux : un indice de Moran élevé peut refléter matrice de voisinage choisie autant que structure réelle.

L'indice de Moran lie directement à régression spatiale. Sur **données Columbus (crime, revenu)**, démonstration analytique puis numérique établit : si matrice poids W normalisée par ligne et variable X centrée, alors $I = \beta$ où β est pente du diagramme de Moran (WX vs X). Cette propriété révèle profonde unité entre indice global et régression : visualiser la pente du diagramme de Moran c'est interpréter l'autocorrélation immédiatement.

Cependant, indice global masque hétérogénéité spatiale. Au cours du devoir des questions ont été posé sur l'approche LISA qui justement palie à cette limite.

L'approche **LISA (Local Moran)** désagrège l'indice par région, classant en quatre quadrants : HH (haute valeur + voisins hauts = cluster positif), LL (basse + basse = cluster négatif), HL/LH (outliers spatiaux). Sur Columbus crime, cette distinction sépare genuine hotspots (HH : quartiers pauvres/criminels entourés similaires) d'anomalies (HL : quartier riche îlot dans pauvreté). Pour policymakers, cette distinction revêt importance opérationnelle : intervention sur hotspot vs outlier demandent stratégies différentes.

5 Réflexions Critiques et Limitations

Le cours révèle limitations fondamentales méritant discussion. D'abord, le conflit **indépendance vs structure spatiale** demeure mal compris. Un indice Moran élevé peut refléter processus spatial réel (contagion crime, diffusion maladie) ou simplement partage covariable cachée (l'équivalent de

l'endogénéité en regréssion linéaire). Ainsi, deux quartiers voisins pourraient être riches non parce qu'ils sont voisins mais simplement parce qu'ils partagent la même histoire urbaine. Identifier une cause exige toujours le développement d'une théorie extérieure : la simple correlation échouera à coup sûr. Sur Baltimore par exemple les prix immobiliers, peuvent refléter les préférences spatiales réelles ou simplement la proximité aux services (écoles, transports). Les données seules sont insuffisantes, avant toute interprétation, le contexte théorique et historique est indispensable.

Deuxièmement, **les données satellitaires ne créent pas ce qui n'existe** : NDVI indique végétation et non le type culture. Durant le TP4 nous identifions des « terres arables » mais sans validation terrain, qu'en sait on réellement ? La fusion de multiples sources raster augmente confidence mais n'élimine pas erreur systématique. De plus, NDVI 2020 reflète usage actuel non potentiel : une terre déboisée récemment compte comme « arable » si la pente et le climat sont favorables, mais pourrait tout aussi bien nécessiter un défrichage ou un térrassement et donc des investissements massif.

Troisièmement, **le cloud computing crée dépendances** : Les algorithmes de GEE algorithmes sont optimisés Google, et ne sont pas totalement transparents (les API servent à faire des requêtes, mais quid de la manipulation mémoire, des formules réelles de calcul, des arbitrages vitesse/Accésibilité ?). Un travail éffectué localement (R/Python) est reproduisibles par quiconque ; GEE dépend Google infrastructure. Pour des chercheurs travaillant pour des gouvernements en Afrique, cette dépendance pose des risques stratégiques non négligeables à moyen terme.

Enfin, **l'intégration raster-vecteur** reste problématique. Qu'on pense par exemple aux débordements fréquents (raster excèdant limites administratives), extractions zônales supposant des polygones valides (en excluant les overlaps). Ou encore l'arbitrage en cas de conflits de données (OSM vs cadastre officiel, WorldPop vs census local qui dit la vérité) ces élements méthodologiques non totalement géré.

Conclusion

Les méthodes de statistique exploratoire spatiale demandent une certaine rigueur. Choisir matrice voisinage, définir une classe raster, valider sur terrain : ce sont des décisions non techniques qui orientent toute l'analyse substrat analytique.

Elle transcende la technique en redonnant l'agentivité aux données. Plutôt que imposer modèle, l'analyse exploratoire écoute données, questionne hypothèses, découvre mondes. En Afrique où institutions faibles, données rares, cette approche humble et itérative s'avère plus productive que modèles sophistiqués reposant hypothèses fragiles.

Synthèse Finale : Décisions Pratiques et Bonnes Pratiques

Quand utiliser quelle technologie ? Le choix d'outils dépend volume données et objectif : analyses < 5 GB sur données locales privilégient R/Python (flexibilité, reproducibilité, documentation via R Markdown) ; données > 50 GB de rasters satellitaires exigent Google Earth Engine (scalabilité cloud, traitement distribué sans infrastructure locale) ; modélisation statistique avancée (autocorrélation spatiale, inférence) mobilise packages R spatialisés (sf, terra, spdep) ; communication policymakers requiert QGIS (cartes statiques professionnelles) et HTML interactif (Folium, Leaflet) ; production opérationnelle 24/7 (monitoring temps réel, alertes) nécessite serveurs cloud et dashboards automatisés.

Checklist avant toute analyse : Vérifier (1) *harmonisation CRS* : tous vecteur/raster identical projection (st_crs() R, crs() Python) ; (2) *validité géométries* : st_is_valid()=TRUE, exclure polygones autointersectants ; (3) *valeurs manquantes documentées* : comprendre source NaN (collecte incomplète vs absence réelle) ; (4) *outliers investigués* : données réelles ou erreurs saisie ? ; (5) *voisinage justifié* : pourquoi matrice Queen vs kNN vs distance ? Tester robustesse matrices alternatives ; (6) *significativité testée* : indice Moran accompagné p-value, rejetter H0 si $p < 0.05$; (7) *résultats validés* : comparaison sources alternatives (OSM vs census, WorldPop vs terrain checks), impossibilité terrain valider tous pixels justifie prudence résultats.

Ressources écosystème : Sources données gratuites : GADM (limites administratives mondiales), OpenStreetMap (infrastructure OSM), WorldPop (densité population 100m), MODIS/Sentinel (satellites couverts GEE), ERA5/CHIRPS (données climatiques). Écosystème R spatial : packages sf (vecteur), terra (raster), spdep (autocorrélation spatiale), tmap (cartographie). Écosystème Python : GeoPandas (vecteur), Rasterio (raster), PyQGIS (scripting SIG).