Aurélien Bornes
M1 Géopolitique-Geoint
Sorbonne Université

Analyse de données - M1

Rapport d'activité

Séance 2 - Les principes généraux de la statistique

La géographie est une discipline en renouvellement perpétuel. Sa structuration et son évolution progressive, au travers de siècles de travaux illustrent ce constat. Néanmoins, elle a pendant longtemps méprisé certaines caractéristiques des statistiques, ne s'inscrivant pas traditionnellement dans son champ disciplinaire. Néanmoins, face à l'importante quantité de données générées par l'ensemble des travaux, force est de constater que seul cet outil statique peut aider à la lecture, et parfois la compréhension, de ces données. Cette complexité de relations a souvent abouti à une sous-estimation du potentiel de cet outil. À l'ère de la massification des données et du Big Data, il n'a jamais été autant important de s'en saisir, notamment sous un spectre d'automatisation. L'information géographique, ou la représentation d'un objet ou d'un phénomène localisé dans l'espace et dans le temps, se décompose traditionnellement en deux séries statistiques distinctes. Le premier se compose de tout ce qui peut caractériser l'ensemble délimité par des éléments de géographie humaine (population, caractéristiques socio-économiques, etc.) ou physique. Le second s'attache à étudier la morphologie même de ces différents ensembles. Dans un SIG, le premier correspond à la base attributaire, le second aux données géométriques. La géographie, bien qu'elle ne produit pas tout le temps ses propres données, exprime des besoins en termes d'analyse de données. Tout d'abord, une nomenclature clairement et préalablement définis, permettant le recueil de l'information. Ensuite, des métadonnées, qui offrent des informations importantes par rapport aux différentes données. Garant de la fiabilité des données, on y trouve classiquement la définition et la nomenclature utilisées, les lieux et dates de l'observation, etc.

Une distinction s'opère entre statistique descriptive et explicative. Les statistiques descriptives, au travers d'une étude des données, cherchent à identifier des propriétés remarquables par rapport à une distribution théorique connue. En résumant et en mettant de l'ordre dans les données, elles permettent d'obtenir une simplification d'un phénomène. En revanche, elles ne cherchent pas la prédiction. En parallèle, les statistiques explicatives permettent d'expliquer, de prévoir ou d'influencer la valeur d'une variable-réponse ou dépendante. En géographie, les types de visualisation des données statistiques sont multiples (histogrammes, diagrammes, boîtes à moustaches, courbes, cartes, etc.). Le choix de tel ou tel type dépend de plusieurs paramètres : la nature de la variable, le type de distribution ou encore l'objectif même du travail. D'autres méthodes d'analyse des données existent. On retrouve à la fois des méthodes :

- "descriptives" à l'image des ACP, AFC, ACM, AFDM, AFM, CAH ou encore nuées dynamiques
- "explicatives" comme des régressions simple, multiple, analyses discriminantes, ANOVA ou segmentation
- "de prévision" telles que des analyses et prévisions de séries chronologiques.

Les méthodologies et les outils statistiques utilisent des termes spécifiques, dont les définitions sont à garder à l'esprit. De manière synthétique, on pourrait définir une population statistique comme l'ensemble des objets (individus ou unités statistiques) sur lesquels se porte l'étude. Par exemple, les habitants de la commune de Nanterre. L'individu statistique est lui un élément à part entière de l'ensemble de cette population statistique. Par exemple, un étudiant vivant à Nanterre. Ces individus ont des caractéristiques particulières, appelées

caractéristiques d'un individu. L'étudiant en question a, par exemple, les yeux marron. Chacun des individus est associé à une modalité, qui s'apparente comme l'ensemble des valeurs prises par un caractère. Cette caractérisation se scinde en deux types. Elle peut être qualitative - désignant une qualité ou une éventualité non chiffrée - ou quantitative - désignant une quantité chiffrée. Elles présentent chacune deux variables de sous-types :

- qualitatives nominales, décrivant et qualifiant les données
- qualitatives ordinales, décrivant la relation entre les données
- quantitatives discrètes, décrivant et comptant une liste finie et isolée de valeurs
- quantitatives continues, décrivant et mesurant les valeurs d'un intervalle.

Il peut exister une hiérarchie entre elles, fondée notamment sur la richesse d'analyse possible. En effet, les variables quantitatives peuvent parfois s'apparenter comme supérieures aux variables qualitatives ordinales et, in fine, nominales.

L'amplitude et la densité d'une série statistique est mesurable. La première, peut se calculer par la différence entre la valeur maximale et minimale, autrement dit les longueurs *b* et *a*. La densité (*d*) est, elle, le rapport entre l'effectif (*ni*) et l'amplitude de la classe décrivant une modalité, où :

$$d = \frac{n_i}{b - a}$$

Les valeurs des classes peuvent être estimées approximativement lors d'une discrétisation, en suivant les formules de Sturges (a) ou Yule (b) :

a. $k \approx 1 + 3,2222 \times \log_{10} n$ b. $k \approx 2,5 \sqrt[4]{n}$

De même manière, il est possible de définir un effectif, qui est le nombre d'occurrences d'une variable dans la population statistique, ou encore de calculer une fréquence. Cette dernière est le rapport entre l'effectif *ni* et l'effectif total *n* (somme de *ni*). Notée i, elle vaut :

$$f_i = \frac{n_i}{n}$$

La fréquence cumulée est la somme des effectifs associés aux valeurs du caractère qui sont inférieures ou égales à *k* :

$$f_i = \sum_{i=1}^{k} n_i \leq k$$

La distribution statistique est quant à elle la répartition observée des fréquences d'un caractère. Elle permet notamment d'identifier la loi de probabilité associée.

Séance 3 - Les paramètres statistiques élémentaires

La variabilité des valeurs d'une série statistique s'illustre au travers des différents paramètres de position, de concentration, de dispersion et de forme. L'existence de plusieurs types de moyenne s'explique par la variation du type de données et d'objectifs d'analyse. Chacune met en avant un aspect différent. Par exemple, la moyenne arithmétique met en avant une valeur globale, mais reste sensible aux extrêmes, la moyenne pondérée tient en compte l'importance différente des valeurs et fournit une mesure plus précise lorsque les données ont des poids différents, la moyenne harmonique prend en considération les vitesses, etc. La médiane, aussi appelée "moyenne du milieu", partage une série de données en deux parties comprenant exactement le même nombre de données. Son calcul permet d'avoir une valeur centrale représentative et robuste. Elle est notamment utile dans les séries statistiques où les valeurs sont déséquilibrées (ex : le salaire des ménages). En revanche, elle ne peut pas être utilisée dans tous les cas. Le mode, simplifié comme une moyenne de fréquence, peut être calculé dès qu'une valeur apparaît plus souvent que les autres dans une série. Il permet en partie d'identifier la valeur la plus fréquente ou celle qui a la plus forte densité de probabilité. Dans le cas où plusieurs modes sont identifiés, on peut parler de série bi-modale ou pluri-modale. La médiale est une valeur centrale partageant la masse d'une variable en deux parties de même poids. C'est une médiane calculée relativement aux valeurs globales. Elle montre la répartition de l'ensemble de cette masse dans la série statistique. La comparaison entre la médiane et la médiale permet d'identifier les effets de concentration d'une population statistique, décrite par la courbe de Gini. Son utilité trouve un écho certain dans la mesure des inégalités au sein d'une distribution, en indiquant notamment les points où les valeurs se concentrent entre quelques individus.

L'unique calcul des écarts à la moyenne ne permet pas de mesurer la dispersion, puisque le somme de ce calcul équivaut toujours 0, notamment puisque les écarts positifs compensent les écarts négatifs. La variance, ou la moyenne des carrés des écarts à la moyenne, permet de mesurer la dispersion globale. Présentée comme la meilleure caractéristique de dispersion, elle évite cette compensation en transformant tous les écarts en valeurs positives. D'un autre côté, on peut aussi remplacer la variance par l'écart-type. S'exprimant en unité au carré, l'utilisation de la variance peut parfois rendre l'interprétation de certaines variables peu intuitive. Par exemple, la variance de la taille (en cm) peut donner des valeurs en cm2. La racine carrée de cette variance, aussi appelée l'écart-type, conserve la même unité que la variable d'origine. En indiquant l'écart "typique" à la moyenne, il est ainsi plus facile d'interpréter certaines variables. Il permet aussi d'identifier le pourcentage de la population appartenant à un intervalle centrée sur l'espérance mathématique. L'étendue est également une caractéristique significative du mesure de la dispersion. Facile à calculer, s'apparentant comme la différence entre la plus grande et la plus petite des valeurs, il évalue l'amplitude de variation d'une série, et ainsi ouvre le champ à la comparaison de variabilité entre plusieurs séries. Autres éléments importants : les quantiles. Ils permettent de savoir où se situent les individus d'une population. Ils offrent aussi une comparaison entre différents "groupes" préalablement définis, nécessitant la construction de seuils.

Parmi les quantiles les plus utilisés, on retrouve :

- la médiane, scindant la population en deux parts égales
- les quartiles, coupant la population en quatre parts égales (avec un Q1 égal 25 %, un Q2 égal à la médiane et un Q3 égal à 75 %)
- les déciles, découpant la population en dix parts égales (où chaque décile, de D1 à D9, équivaut à 10 %)

Les étendues interquartiles et interdéciles représentent la différence entre deux quartiles ou déciles. Tous ces éléments peuvent être agrégés dans une boîte de dispersion, aussi appelée boîte à moustaches. C'est la seule figure où l'on peut voir la valeur maximale, minimale, la médiane, le 1er et le 3e quartile, ainsi que le 1er et le 9e décile. Elle s'apparente alors comme une sorte de synthèse de la distribution et met en avant sa dispersion. Sa lecture est plutôt simple, la taille de la boîte représente l'écart interquartile (soit Q3-Q1). Plus elle est longue, plus la dispersion est forte. On y retrouve la médiane, où son décentrement témoigne d'une répartition asymétrique. Les moustaches représentent les valeurs minimales et maximales. La plus ou moins grande longueur de ces dernières sous-entend la concentration des valeurs d'un côté ou de l'autre et une dispersion hors des valeurs de la zone centrale (et donc de la médiane).

Les paramètres de formes caractérisent la symétrie, l'aplatissement, etc. de la loi de distribution statistique de la variable étudiée. Les moments centrés, calculés à partir de la moyenne et de ses écarts servent à analyser la forme de la distribution (dispersion, aplatissement, asymétrie, etc.). Les moments absolus, fondés sur la valeur absolue des écarts (variance), mesurent la dispersion réelle et sans compensation entre écarts positifs et négatifs. Complémentaires, les moments centrés répondent davantage à une analyse assez théorique de la distribution, les moments absolus à une description plus robuste.

Il est important de vérifier la symétrie d'une distribution, puisqu'elle influence le choix et la pertinence des indicateurs et l'interprétation correcte de la dispersion. On peut la vérifier à l'aide des coefficients β1 et β2 de Pearson et de Fisher. La mesure de la dissymétrie se fait avec cette formule :

$$\beta_1 = \frac{\mu_3}{{\mu_2}^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3}$$

Où une valeur de β1 > 0 montre l'étalement sur la droite de la distribution est étalée sur la droite, la dissymétrie est dite positive. Si β1 < 0 alors la distribution est étalée sur la gauche. La dissymétrie est dite négative. Enfin, si β1 = 0 alors la distribution est symétrique. La mesure d'aplatissement β2 se réalise avec cette formule :

$$\beta_2 = \frac{\mu_4}{{\mu_2}^4} - 3$$

Où une valeur de β2 > 0 illustre le caractère platicurtique de la distribution. Si β2 < 0 alors la distribution est dite leptocurtique. Enfin, si β2 = 0 alors la distribution est dite mésocurtique.

Séance 4 - Les distributions statistiques

Le choix entre une distribution statistique avec des variables discrètes et une distribution avec des variables continues repose sur plusieurs critères. Premièrement, la nature de la variable étudiée. C'est l'un des critères les plus déterminants, si ce n'est le plus. Une variable est dite discrète lorsqu'elle prend un nombre fini ou dénombrable de valeurs, le plus souvent entières et isolées. Par exemple, le nombre d'épisodes d'inondations le long d'un cours d'eau. À l'inverse, une variable est dite continue lorsqu'elle prend un nombre trop important ou infini de valeurs différentes sur un intervalle réel. Par exemple, le débit des stations hydrologiques de ce même cours d'eau. Deuxièmement, les mécanismes et processus à l'origine du phénomène. En effet, ce choix dépendant aussi du processus réel et du modèle probabiliste dans lequel s'inscrit cette variable. Si le phénomène est généré par des événements plutôt rares ou faciles à compter, on optera pour une loi discrète (type binomiale, géométrique, Poisson, etc.). Ils correspondent à des événements aléatoires ou à des processus discontinus, par exemple, des glissements de terrain sur un versant. Au contraire, si le phénomène à l'origine est généré par un processus cumulatif, multi-causal ou bruité, on choisira une loi continue (type normale, log-normale, etc.). Ces phénomènes sont caractérisés par une variation continue dans l'espace ou dans le temps, par exemple, la répartition des hauteurs de précipitations. Troisièmement, la forme empirique de la distribution des données. Même si la nature de la variable est primordiale, il faut rappeler qu'il est important d'analyser l'histogramme, la symétrie et la concentration des valeurs de la variable. La forme du nuage de données influence également ce choix, où des valeurs concentrées sur quelques entiers devraient s'accompagner d'un caractère discret, et des valeurs réparties continûment, d'un caractère continu. Enfin, le choix entre variables discrètes ou continues peut aussi se faire en fonction du nombre de paramètres (certaines lois ont peu de paramètres), de la taille de l'échantillon, et de l'objectif final.

En géographie, certaines lois statistiques sont davantage utilisées. C'est le cas de l'incontournable loi normale (ou de Gauss). Référence pour les variables continues issues de facteurs assez indépendants, elle est massivement utilisée pour les variables climatiques moyennes, les variations d'altitudes locales, les erreurs de mesures, etc. La loi log-normale est fréquemment employée dans le cas de phénomènes résultant de processus multifactoriels, et notamment en géographie physique pour la surface des bassins versants ou des parcelles agricoles, la granulométrie ou encore la sédimentologie. Toujours en géographie physique, on observe le recours à la loi exponentielle dans les cas de travaux liés aux durées et notamment aux temporalités entre deux événements. Par exemple, le temps entre deux crues ou deux avalanches. En géographie urbaine, la loi de Zipf / Zipf-Mandelbrot est très utilisée puisqu'elle permet notamment la description de la relation entre la taille et le rang d'une ville. Son application dans l'analyse des systèmes urbains peut d'ailleurs être adaptée à l'ensemble des autres types de territoires. D'autres lois sont également utilisées, comme celle de Poisson. En revanche, ces lois sont souvent peu expliquées dans la littérature scientifique. Bien souvent, les géographes ont recours à des hypothèses statistiques qui correspondent à ces lois, sans forcément les citer explicitement. En rappelant que, malgré l'incontournable loi normale, certaines distributions ne le suivent pas.

Résultats :

Les différents graphiques générés par le script Python montrent que la distribution uniforme est caractérisée par une fréquence relativement constante pour l'ensemble des classes. Cette homogénéité confirme que chaque valeur de l'intervalle a la même probabilité d'être tirée. La moyenne calculée par le programme est proche du centre de l'intervalle, ce qui est cohérent avec la symétrie de la loi uniforme. L'écart-type traduit une dispersion modérée et constante, sans concentration particulière autour d'une valeur spécifique.

Par exemple, pour la loi normale, le graphique produit par le script met en évidence une distribution symétrique en forme de cloche, centrée autour de la moyenne. Les valeurs sont majoritairement regroupées autour d'elle, tandis que les fréquences diminuent progressivement lorsque l'on s'en éloigne. L'écart-type calculé montre l'étalement de la distribution : plus il est élevé, plus la courbe est large. Les résultats affichés par le script sont cohérents avec la forme du graphique, ce qui valide à la fois le calcul de la moyenne et de l'écart-type et la pertinence de la simulation.

Séance 5 - Les statistiques inférentielles

L'échantillonnage consiste à étudier une partie de la population, appelée échantillon, afin d'en tirer des conclusions et de les transposer à l'ensemble de la population, qu'on appelle population "mère". En statique inférentielle, on part du principe qu'il est rarement possible d'observer l'entité de la population étudiée, à cause de critères contraignants : taille de la population, temps contraint, coût de l'étude, accès limité aux données, etc. L'objectif est donc d'obtenir un échantillon représentatif, autrement dit assez fidèle à l'ensemble de la population pour permettre une généralisation suffisamment robuste. Un échantillon suffisamment bien construit peut fournir des résultats significatifs, remplaçant une hypothétique étude trop technique à appliquer sur l'ensemble de la population statistique. Il existe plusieurs méthodes d'échantillonnage. Celles fondées sur l'aléatoire sont les plus solides, donnant à chaque individu une probabilité similaire d'être sélectionné. On peut notamment citer le tirage aléatoire simple, avec ou sans remise. À l'inverse, les méthodes non-aléatoires (choix raisonné, quotas, opportunité) sont plus faciles à mettre en place, mais introduisent davantage de biais. Le choix entre ces différentes méthodes dépend à la fois des contraintes pratiques, des objectifs de l'étude et du degré de précision ciblé. Un estimateur peut se définir comme une formule statistique dont le calcul repose sur les données d'un échantillon. Son but est d'approcher un paramètre inconnu de la population, comme une moyenne, une variance ou une proportion. Il s'apparente donc comme une variable aléatoire, sa valeur dépend de l'échantillon. L'estimation correspond donc à la valeur numérique concrète obtenue suite à l'application d'un estimateur sur un échantillon donné. Autrement dit, l'estimateur est l'outil théorique, l'estimation en est le résultat pratique.

L'intervalle de fluctuation est utilisé lorsque le paramètre de la population est supposément connu. Il vérifie la validité et comptabilité des valeurs observées dans un échantillon avec l'hypothèse de départ, en tenant compte de l'aléa d'échantillonnage. À l'inverse, l'intervalle de confiance sert à estimer un paramètre inconnu. Il donne une plage de valeurs dans laquelle le paramètre a une forte probabilité de se situer, avec un certain niveau de confiance. C'est un outil essentiel de l'estimation statistique.

En théorie de l'estimation, un biais apparaît suite à la surestimation ou sous-estimation de la valeur réelle d'un paramètre étudié par un estimateur. On dit qu'il est sans biais lorsque son espérance mathématique est égale à la véritable valeur de ce même paramètre. Ainsi, il correspond à une erreur systémique. Lorsqu'une statistique est calculée à partir de l'ensemble de la population, on parle alors de statistique exhaustive. Le big data a pour particularité de parfois donner l'impression d'une utilisation des populations au complet. En réalité, certaines de ces sources de données ne le sont pas, dépendant en partie de la méthodologie de collecte, des plateformes ou encore du comportement des individus. Elles peuvent donc être vues comme des échantillons de tailles significatives, mais comme intégratrices de la totalité de la population. Plusieurs enjeux sont inhérents au choix d'un estimateur, car ils conditionnent plus particulièrement les résultats finaux. Il doit être légèrement biaisé, précis (avec une faible variance) et stable lorsque la taille de l'échantillon augmente. De même manière, il existe plusieurs méthodes pour estimer un paramètre. On peut notamment citer : la méthode des moments, la méthode des moindres carrés, la méthode du maximum de vraisemblance ou encore certaines méthodes de rééchantillonnage comme le *bootstrap*.

Ce choix dépend de la loi supposée liée aux données, de la taille de l'échantillon, des propriétés recherchées pour l'estimateur ou des objectifs plus généraux de l'étude.

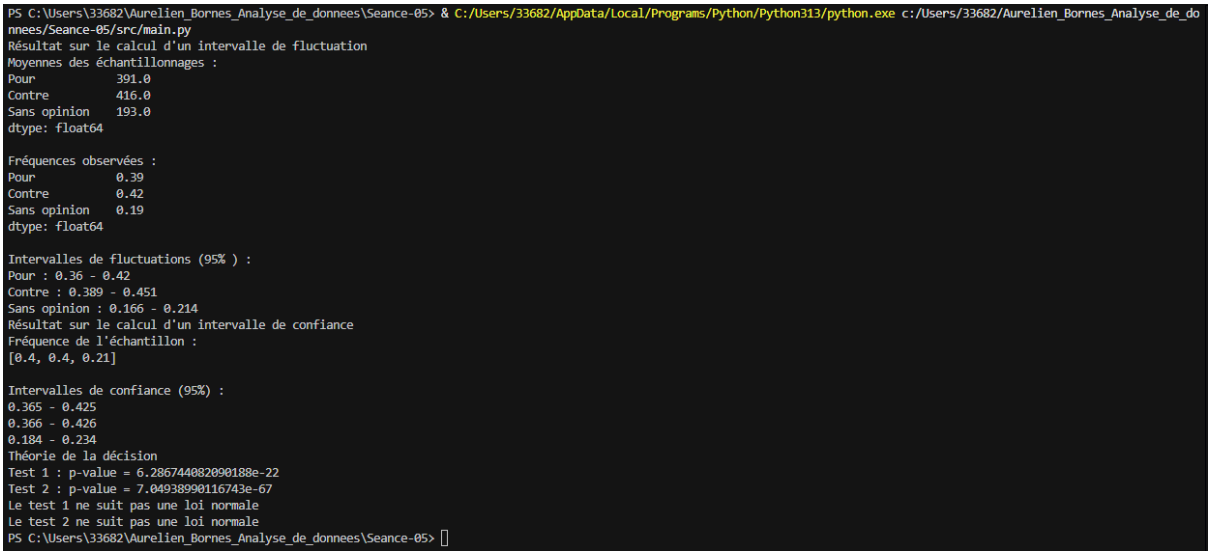
Les tests statistiques servent à prendre une décision à partir des données, en évaluant la compatibilité des observations avec une hypothèse donnée. Ils sont notamment utilisés pour la comparaison entre groupes, le test d'une relation ou la vérification d'une hypothèse. On distingue en particulier les tests "paramétriques", reposant sur des hypothèses fortes sur la distribution des données, et les tests "non paramétriques", plus souples mais parfois moins puissants. La construction d'un test statistique consiste à formuler une hypothèse nulle, choisir un seuil de risque, calculer une statistique et décider, à partir d'une *p-value* ou d'une région critique, si cette hypothèse peut être rejetée.

La statistique inférentielle fait l'objet de plusieurs critiques, notamment en raison de sa dépendance à ces hypothèses, des risques de biais liés à l'échantillonnage et de l'interprétation pouvant être parfois abusive des différents résultats. Les tests statistiques ne fournissent jamais de certitudes absolues, mais seulement des conclusions probabilistes. Malgré ces limites, il n'en demeure pas moins que la statistique inférentielle reste indispensable pour l'analyse de données, notamment dans la comparaison des différentes situations et la production de connaissances à partir de sources de données partielles.

Résultats :

Les résultats montrent une répartition relativement équilibrée des opinions au sein de l'échantillon. La modalité "contre" est légèrement majoritaire (42 %), suivie par la modalité "pour" (39 %), tandis que la catégorie "sans opinion" reste minoritaire (19 %). Les intervalles de fluctuation à 95 % indiquent que les fréquences observées sont compatibles avec un échantillonnage aléatoire, témoignant de la stabilité des résultats. Les intervalles de confiance à 95 %, relativement étroits, montrent une bonne précision de ces estimations. Le recouvrement important des intervalles associés aux modalités "pour" et "contre" ne permet pas de conclure à une différence statistiquement significative entre ces deux proportions au seuil de 5 %. En revanche, la proportion de "sans opinion" est nettement inférieure et bien distincte des deux autres modalités. Les tests de décision conduisent au rejet des hypothèses nulles, mais les tests de normalité indiquent que les distributions ne suivent pas une loi normale. Par conséquent, l'interprétation doit privilégier une approche fondée sur les proportions et les intervalles de confiance plutôt que sur des tests paramétriques classiques.

En conclusion, l'opinion apparaît partagée entre les positions "pour" et "contre", sans réelle domination statistique de l'une sur l'autre, tandis que l'absence d'opinion reste marginale.



Séance 6 - La statistique d’ordre des variables qualitatives

Une statique ordinale est une statistique s’appliquant à des variables qualitatives ordonnées, pour lesquelles les modalités peuvent être classées selon un ordre logique ou “naturel”, sans que l’on puisse quantifier précisément les écarts entre elles. Elle se distingue ainsi de la statistique nominale, qui concerne les variables qualitatives sans ordre particulier, par exemple des types de sols. Les statistiques ordinales peuvent reposer sur des rangs, des classements, ou encore des positions relatives plutôt que sur des valeurs numériques absolues. Elles sont particulièrement importantes en géographie, car de nombreux phénomènes spatiaux peuvent se retranscrire au travers de ce type de hiérarchisation. On peut notamment penser à des pays classés selon leur niveau de développement (IDH, PIB, etc.). On comprend dès lors, que le recours à ce genre de statistiques peut mettre en évidence le positionnement relatif des objets et entités géographiques les uns par rapport aux autres, et ce, malgré une certaine inexactitude.

Dans les classifications statistiques, l'ordre à privilégier est l'ordre croissant, aussi appelé ordre naturel. Il facilite l'interprétation des données, la comparaison entre objets et l'identification des valeurs extrêmes, qu’il s’agisse des plus faibles ou des plus élevées. Cet ordre permet également de repérer des valeurs aberrantes et d’étudier les minimums ainsi que les maximums d’une série statistique. Il existe toutefois des exceptions en géographie, et notamment dans les études urbaines, où le recours à la loi rang-taille peut privilégier l’utilisation d’un ordre décroissant afin de mieux rendre compte des différentes spécificités inhérentes aux phénomènes hiérarchiques. La corrélation des rangs vise à mesurer le lien statistique entre deux variables ordinales en comparant les différents rangs attribués aux objets dans deux classements différents. Elle permet de savoir si, globalement, un objet est bien classé selon tel ou tel critère. La concordance de classements, quant à elle, s’interroge davantage sur la cohérence globale entre plusieurs classements. Elle identifie dans quelle mesure différents classements peuvent aboutir à des ordres similaires, en comptant les paires tant concordantes que discordantes. Alors que la corrélation des rangs compare deux variables, la concordance peut être étendue à plusieurs classements et cherche à évaluer leur concordance globale.

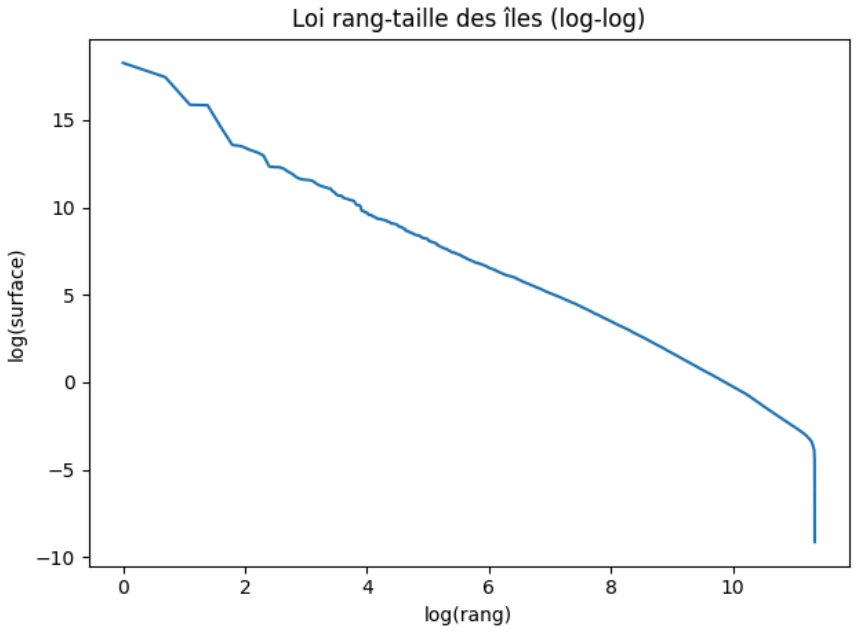
Le test de Spearman repose sur le calcul d’un coefficient de corrélation appliqué aux rangs. Il mesure la robustesse et le sens de la relation monotone entre deux classements. Ce test est simple à mettre en œuvre et se rapproche du coefficient de corrélation linéaire, supposant en revanche l’absence de rangs *ex æquo* ou nécessitant des corrections lorsqu’ils existent. Le test de Kendall, en revanche, repose sur le dénombrement des paires concordantes et discordantes entre deux classements. Souvent considéré comme plus robuste, notamment lorsque les effectifs sont plus faibles ou les rangs *ex æquo* plus fréquents, il peut être généralisé à plusieurs classements. Cela constitue un avantage important dans l'analyse géographique multi-critère.

Le coefficient de Goodman-Kruskal sert à mesurer l’association entre deux variables ordinales en comparant le nombre de paires concordantes et discordantes, tout en donnant une mesure synthétique de la force et du sens même de l’association. Il est particulièrement utile lorsque l’on travaille sur des classements ou des variables qualitatives ordonnées. Le coefficient de Yule est un cas particulier du coefficient de Goodman-Kruskal, appliqué

spécifiquement aux tableaux de contingences “2 x 2”. Il mesure notamment l’association entre deux variables binaires et est souvent interprété en lien avec le “rapport de cotes”. C’est un bon outil pour l’analyse de relations entre variables qualitatives, lorsque les hypothèses des statistiques paramétriques classiques ne sont pas satisfaites.

Résultats :

L’analyse des données issues des fichiers met en évidence de fortes disparités entre les pays. Les statistiques descriptives calculées avec Python montrent que les variables étudiées ne sont pas réparties de manière homogène. Certaines valeurs extrêmes - pays très peuplés, très vastes, très petits États insulaires - influencent fortement la moyenne. Le script révèle également des relations intéressantes entre certaines variables. Par exemple, la comparaison entre la surface des pays et leur population met en évidence l’absence d’une corrélation simple : des pays très vastes peuvent avoir une faible population, tandis que certains pays de petite taille sont fortement peuplés.



Le graphique produit par le script Python montre une décroissance rapide des surfaces en fonction du rang. Les premières valeurs correspondent à des îles de très grande surface, tandis que la majorité des observations présente des surfaces beaucoup plus faibles. Cette forte dissymétrie explique pourquoi la moyenne calculée par le script est élevée par rapport à la médiane, cette dernière étant plus représentative de la tendance centrale des données. La représentation en échelle logarithmique (log-log) permet de transformer cette relation en une presque linéaire. Cette linéarisation suggère que la distribution des surfaces suit une loi de type puissance, caractéristique des phénomènes géographiques hiérarchisés. Les résultats du script montrent ainsi que le passage en log-log est indispensable pour interpréter correctement la structure des données.

Réflexion personnelle sur l'exercice

Tout d'abord, il me paraît nécessaire de préciser que mon parcours académique, initialement orienté vers des disciplines davantage "lettrées", à savoir le droit puis la géographie et enfin la géopolitique, ne m'a pas permis de bénéficier d'une formation préalable en statistique ni en rédaction de script, et notamment en langage Python. Cette absence de socle technique a constitué pour moi une difficulté majeure dans l'appropriation des outils mobilisés au cours de cet enseignement. L'apprentissage du script représente, pour moi, un changement de paradigme intellectuel important, dans la mesure où il suppose non seulement la maîtrise de nouvelles notions, mais également l'intégration d'une logique procédurale qui m'était jusqu'alors peu familière. À ce titre, la compréhension globale de la structure des scripts, de leur enchaînement logique et des interactions entre les différentes fonctions s'est révélée particulièrement exigeante, et n'a pu être que partiellement assimilée dans le temps imparti.

Néanmoins, malgré ces difficultés initiales, j'ai tenté de répondre de mon mieux à vos attentes. Ces efforts, bien que parfois infructueux ou incomplets, m'ont permis de dépasser une appréhension première et de prendre conscience du caractère non fataliste de cette situation. Progressivement, et au fil des essais, j'ai commencé à identifier les étapes essentielles du raisonnement à adopter, ainsi que les bonnes pratiques nécessaires à une utilisation plus efficace et plus autonome des outils d'analyse de données. Cette montée en compétences, bien qu'encore limitée, m'a permis d'entrevoir les potentialités offertes par ces méthodes, et notamment sur le plan analytique.

Par ailleurs, l'importance du travail collectif s'est révélée déterminante dans ce processus d'apprentissage. La réflexion en groupe, fondée sur l'échange des difficultés rencontrées et des solutions envisagées, a constitué un appui méthodologique précieux. Elle a été complétée par le recours à diverses ressources externes (plateformes en ligne, forums spécialisés) ainsi que les outils d'intelligence artificielle. Pour un novice en programmation, ces supports apparaissent ainsi comme indispensables dans l'appropriation progressive des savoirs techniques.

Toutefois, la rigueur méthodologique et la technicité inhérentes à l'analyse de données et à la programmation requièrent un investissement temporel conséquent et continu. Or, le temps dont j'ai disposé au cours de ce semestre s'est avéré limité. Cette contrainte s'explique en partie par la charge de travail importante liée au master Géopolitique-Geoint, dont les exigences académiques sont particulièrement élevées, mais également par mes engagements militaires, au sein desquels j'exerce des responsabilités d'encadrement impliquant une charge professionnelle supplémentaire. La conciliation de ces différentes obligations a nécessairement restreint le temps que j'ai pu consacrer à ce travail.

C'est ainsi avec un certain regret que je constate, au moment de la remise de ce travail, que des délais plus larges m'auraient permis de produire un rendu plus abouti, tant sur le plan méthodologique que sur le plan analytique. Néanmoins, cette expérience demeure particulièrement formatrice, dans la mesure où elle m'a permis de prendre conscience des exigences propres aux méthodes quantitatives, tout en constituant une base solide pour un approfondissement ultérieur de ces compétences.

Réflexion personnelle sur les sciences des données et les humanités numériques

---

Les sciences des données et la dimension numérique occupent aujourd'hui une place de plus en plus centrale dans l'évolution des pratiques dans le champ du renseignement géospatial (Geoint). Dans un environnement numérique saturé en données et dans un contexte géopolitique et géostratégique d'affrontements hybrides entre puissances, les sciences des données peuvent apporter une capacité accrue à traiter des volumes importants d'informations multi-sources et multi-capteurs : images aériennes (satellites, radar, drônes, etc.), photographies, bases de données économiques, flux issus des réseaux sociaux (vidéos notamment), etc.

La cartographie, ainsi que l'analyse et la fusion de ces données montrent que la valeur stratégique ne réside pas uniquement dans l'accès à ces données, mais dans la capacité à les structurer, à les croiser et à en extraire des informations pertinentes. Couplé avec une automatisation des processus, la maîtrise des possibilités offertes par les sciences des données permettront de prendre l'ascendant sur nos compétiteurs. Dans une perspective plus Geoint, ces compétences permettent de fluidifier et d'améliorer nos analyses à destination des grands décideurs, militaires comme politiques.

Cependant, il faut également souligner les limites d'une approche strictement techniciste. Les humanités numériques jouent ici un rôle fondamental en rappelant que les données ne sont jamais neutres. Leur production, leur sélection et leur visualisation sont le résultat de choix, voire de biais, méthodologiques et culturels, voire politiques. Les travaux mobilisant des textes, récits historiques, discours politiques ou des représentations cartographiques montrent que l'analyse géopolitique ne peut simplement se réduire à des modèles statistiques ou à des algorithmes. Elle nécessite une contextualisation, une compréhension des enjeux, des échelles, des temporalités, des jeux des acteurs et des rapports de pouvoir. Et c'est là qu'interviennent les compétences d'étudiants formés en sciences humaines, et notamment en géographie.

Enfin, cette réflexion met en avant l'évolution du rôle d'analyse en géopolitique ou en renseignement. Il est clair que l'avenir de ma discipline tend vers une articulation plus poussée entre sciences humaines et sciences des données, notamment dans un objectif de former des analystes dotés d'une posture réflexive et analytique hybride. On ne peut plus se passer des opportunités offertes par les sciences des données. Cette volonté d'une maîtrise des outils de données ne vise pas à remplacer l'analyse humaine, mais à l'enrichir. L'enjeu est de former des profils capables d'utiliser cette masse de données, tout en conservant une capacité de recul critique, d'interprétation et d'analyse en géopolitique. Les sciences des données et les humanités numériques apparaissent ainsi non comme des champs opposés, mais comme deux dimensions complémentaires d'une même démarche : comprendre la complexité du monde contemporain à partir des informations numériques qu'il produit, sans perdre de vue les dimensions humaines, politiques et spatio-temporelles qui le structurent.