Robin Pisot M1GéoSuds

Analyses de données : rapport d'activité.

Parcours débutant.

Je vous informe que je souffre de troubles dys, alors même si j'ai chassé autant que possible les fautes, je vous présente mes excuses si certaines devait m'avoir échappé.

Séance 2 :

Questions de cours :

1. La Géographie se sert des statistiques pour traiter l'importante masse de données que son étude produit. Cependant, la Géographie ne considère les statistiques seulement ou presque comme un outil. On pourrait parler de « schismogenèse » telle que définie par Grégory Bateson, ainsi paradoxalement la Géographie se renforce dans son statut de science humaine et sociale en se dissociant de son outil : les statistiques. Cette dissociation produit une conception de la Géographie comme séparée des statistiques, comme la main ne se confond pas avec l'outil qu'elle tient, alors qu'en pratique, Géographie et statistiques se mêlent dans l'étude géographique.

2. La question de savoir si le hasard existe ou non en Géographie est une mauvaise question, de même pour la question de savoir si le déterminisme existe ou non en Géographie. D'abord parce que la définition du hasard varie en fonction de la tradition philosophique et/ou épistémologique dans laquelle on s'inscrit, et ensuite car l'étude géographique ne prévoit rien, et si dans une certaine mesure elle cherche à dégager les causes à l'origine de tel ou tel phénomène géographique, il est toujours vis-à-vis des études de l'œcoumène une opacité des causes liées aux pratiques, aux volontés et aux représentations de l'espace par l'Homme. La Géographie peut bien s'appuyer sur d'autres disciplines des sciences humaines et sociales, en premier lieu l'Anthropologie et la Psychologie, pour dégager les causes à l'origine des pratiques humaines qui modifient les objets des études géographiques. Cependant, d'un point de vue statistique, le hasard n'existe que dans la mesure où certaines causes ne sont pas connues, mais toute variation numéraire, mathématique ou physique ayant une cause, le hasard au sens strict n'existe pas.

3. On considère deux types d'informations géographiques. Le premier ce sont les données que vous qualifiez d'« attributaires », c'est-à-dire qui désignent l'ensemble des données caractéristiques géographiques physiques, humaines ou environnementales d'un territoire. Cette catégorie regroupe autant des données quantitatives (températures, densité de population, indice de fécondité, taux en

tout genre) que qualitative (toponymie, caractéristiques pédologiques ou météorologiques, typologie socio-économique). Le second type de d'information géographique désigne les données morphologiques, géométriques, regroupant ainsi les formes, les superficies et surfaces, les coordonnées les contours, les lignes de niveaux, les réseaux … Ces deux types d'informations géographiques sont les piliers des SIG : ArcGis Pro fonctionne avec des couches morphologiques sur lesquelles on représente avec des figurés cartographiques des données attributaires contenues dans la table éponyme.

4. L'analyse de données en géographie a 3 besoins principaux. D'abord, c'est tautologique mais il y a un besoin de données, ou de métadonnées pour en extraire des données. Cela dit, elle a ensuite besoin de trier les types de données, ce qui passe par la nécessité de les nommer, d'où l'usage crucial de la nomenclature statistique pour connaître le type d'objet statistique, chaque type nécessitant un traitement différent adapté. Enfin, il faut évidemment traiter ces données avec des outils statistiques pour pouvoir en extraire des informations qui peuvent être appréhendées visuellement (dans un SIG, par exemple) ou intellectuellement par un résumé ou un résultat statistique (moyenne, médiane, quartiles, etc.) qui permet de définir une caractéristique de l'ensemble de la population traitée.

5. Alors, si je puis me permettre, la première décrit, la seconde explique… Pour développer un peu, il s'agit de deux types de statistiques aux usages différents et transparents par leurs noms. La première, la plus commune, décrit des populations à l'aide des moyennes, médianes, quartiles… des outils statistiques communs pour classer ou ordonner une population, ce qui peut se matérialiser par des représentations graphiques. La seconde explicite une variable, en faisant apparaître une variable cachée ou corrélée, à faire apparaître les relations entre différentes variables, parfois de nature différente, en utilisant des outils discriminants ou les régressions statistiques. La statistique explicative a une dimension plus prévisionnelle qui permet, *ceteris paribus*, de montrer une dynamique des données quantitatives.

6. On connait trois principaux types de représentations des données en géographie, que l'on choisit en fonction du type de données et du besoin. Le tableau d'abord permet de visualiser l'ensemble des données, classées, ordonnées, et en chiffres ce qui permet la précision dans la lecture. Les graphiques ensuite : diagrammes, histogrammes, nuages de points, courbes, « boites à moustaches », moins précis qu'un tableau mais qui permettent souvent mieux d'appréhender les dynamiques et les évolutions en fonction des différentes variables. Enfin, outil du géographe par excellence, la carte, qui permet la représentation spatiale. On peut lui

superposer des graphiques pour cumuler les types de données représentées sur un même support.

7. Je crains de me répéter avec la question 5. Il y a deux types de méthodes d'analyse de données statistique. La première est la statistique descriptive, la seconde est la statistique explicative. Alors, issue de cette dernière on a bien une méthode de statistique prévisionnelle, mais cela en est une branche. (cf. Question 5)

8. Définitions

    a. Une population statistique c'est l'ensemble des individus statistiques, des unités étudiées. C'est le groupe étudié qui, par l'accumulation des données individuelles, donne à voir des dynamiques ou des caractéristiques soit commune soit propre au groupe.
    b. L'individu statistique c'est l'unité à laquelle est attribuée une ou plusieurs données. Cette unité prend une importance lorsqu'elle est analysée au sein d'un groupe d'autres individus statistiques, c'est-à-dire une population statistique.
    c. Un caractère statistique c'est un attribut pour un individu, c'est ce qu'on mesure statistiquement de la population statistique, de l'ensemble des individus : l'âge, le revenu, le sexe, la taille, la couleur des yeux … Le type de caractère se divise en deux catégories : qualitative et quantitative, qui se divisent elles-mêmes en deux sous-ensembles
        i. Caractères qualitatifs :
            1. Nominaux : désigne les caractères qu'on ne peut pas ordonner dans un ordre de valeur. C'est le cas des noms par exemple.
            2. Ordinaux : Désigne les caractères que l'on peut ordonner selon un ordre de valeur non-chiffré. Par exemple du plus au moins bien, du plus ou moins régulier (rarement, parfois, souvent, toujours : sont des caractères qualitatifs ordinaux du plus au moins régulier).
        ii. Caractères quantitatifs :
            1. Discrets : désigne une valeur en nombre entier, comme un nombre d'enfants dans une famille.
            2. Continus : désigne une valeur qui peut être dans un intervalle infini et en nombre décimal comme une température ou une taille.
    d. Une modalité statistique c'est une valeur possible pour un caractère. Un sexe (je parle bien du sexe et non du genre, ce qui implique des considérations politiques non pertinentes ici) ne pourra pas être qualifié

comme jaune ou rugueux, ses modalités ne peuvent être que : Féminin / Masculin / Autre.

e. Pour ce qui est de la hiérarchie, vous considérez que les caractères quantitatifs sont plus utiles que les caractères qualitatifs puisque l'on peut en extraire plus de données chiffrées. Je m'inscris en désaccord avec le cours, ce sont deux types de caractères aux usages différents, le fait qu'on puisse faire plus de manipulations statistiques avec l'une ne réduit en rien l'importance de l'autre. Il n'y a selon moi pas de hiérarchie à faire entre des objets qui ne me paraissent pas comparables.

9. Alors, si par « amplitude » vous entendez l' « étendue » (déjà dans ce cas pourquoi ne pas utiliser le terme consacré d'étendue ?), on la mesure en soustrayant la valeur minimale d'une donnée dans une population de la valeur maximale dans la même population pour le même caractère. Dans l'ensemble 5,8,12,3,10, on soustrait 3 à 12. Alors l'étendue est de 9, c'est-à-dire le nombre d'unité qui sépare la valeur minimale de la valeur maximale. Si l'amplitude ne désigne pas l'étendue je ne vois pas à quoi vous faîtes référence.
   La densité se mesure en fonction de l'unité choisie. Ainsi on mesure une densité en divisant un effectif par une valeur d'une unité de référence (les km2 par exemple).

10. Alors, c'est bien simple, je ne sais pas. De ce que j'ai réussi à comprendre, elles permettent de faire passer une valeur à caractère quantitatif continu à une valeur à caractère quantitatif discret, pour permettre la simplification des données dans le cadre d'une représentation graphique sans corrompre les données. Sinon comment ? pourquoi ?

11. L'effectif, c'est le nombre d'apparition d'une modalité dans une population. Une fréquence se mesure en divisant l'effectif par la population. Une fréquence cumulée se mesure en additionnant les fréquences précédentes au-dessous ou au-dessus d'une certaine valeur. La distribution statistique c'est l'ensemble des effectifs ou fréquences par modalité de caractère.

Code :

Alors, faisons court : je n'ai pas réussi à faire fonctionner python. Alors pourtant vous-même puis Zara m'avez installé la machine mais je n'arrive pas à faire fonctionner le code. Alors, on me l'a offert, ou plutôt j'ai été pris en pitié par des camarades qui m'ont fait la charité de leur temps pour me faire et m'expliquer le code de la séance. Le résultat

est que je n'arrive pas à le reproduire. Alors oui j'ai des graphiques, diagrammes en bâton et histogrammes, mais si j'en crois votre page d'exercice, il n'y a pas à les commenter. Alors, je peux comprendre que ce n'est pas l'objet du cours, que l'objectif est de produire de la donnée et de la représentation, mais produire de la donnée pour produire de la donnée me paraît pour l'instant un peu stérile, surtout dans la mesure où nous ne savons pas d'où extraire les données qui pourrait nous être utiles, et que nous n'apprenons nulle part à commenter les données que nous avons extrait.

Conclusion :

Pour l'essentiel je pense cumuler toutes mes considérations dans la conclusion globale. Je note cependant que pour 11 questions j'ai mis 6 heures à y répondre, donc si j'en crois mes tables de multiplication, et si les questions sont en nombre et en difficulté comparable pour les prochaines séances, je mettrais environ 30 heures pour faire les questions, sur un parcours qui en fait 50 selon votre aveu, et cela sans compter le code que, pour l'instant, je suis incapable de faire seul. Puisque le rapport doit se faire par étape, je pense que j'en suis à l'étape du pessimisme.

Séance 3 :

Questions :

1. Je crois que la question est mal posée, au sens où il me semble y avoir là une ambiguïté sémantique. Selon votre cours, le caractère le plus général semble être le caractère quantitatif, ce qui se défend par la variété des outils et des méthodes pour l'analyser. D'un autre côté, le caractère qualitatif est adossé à l'ensemble du langage, et que ce soit dans sa dimension nominale ou ordinale, ayant derrière lui l'ensemble des adjectifs et des toponymes (ou même des noms d'ailleurs), ce qui en fait un caractère très général aussi. Aussi, de ce que je comprends, le caractère quantitatif est plus général par ses dimensions d'analyses, le qualitatif plus général dans la nature des objets qu'il désigne, objets régulièrement inquantifiables.

2. Pour ce qui est de définir des caractères quantitatifs discrets et les caractères quantitatifs continus : cf Séance 1, Q.8, c. Pour ce qui est du pourquoi les distinguer ? Simplement parce qu'ils ne traitent pas des mêmes objets. Les caractères quantitatifs discret par exemple ne peuvent pas traiter d'objet décimaux : pour reprendre l'exemple du nombre d'enfants par famille, il est difficilement envisageable pour une famille (= un individu statistique) d'avoir 0.5 enfants, au sens d'avoir un demi-enfant. Il y a des découpages qui ne sont pas

possibles, d'où le caractère discret pour différencier les objets indivisibles de ceux qui sont divisibles et donc relatifs au caractère continu.

3. Encore une fois, il existe différents types de moyennes pour différents types d'objets et pour différents types d'usages. La moyenne arithmétique est utilisée quand toutes les valeurs ont la même importance. La moyenne pondérée est utilisée quand certaines données comptent plus que d'autres, d'où l'usage des coefficients. La moyenne géométrique sert elle plutôt à faire la moyenne de taux d'évolutions. La moyenne harmonique est elle plutôt utilisée pour les vitesses et les débits.
   Le calcul d'une médiane permet de distribuer les individus statistiques en deux parties égales avec autant de valeurs en-dessous et au-dessus de la médiane. Elle est certes insensible aux valeurs externes, à la différence d'une moyenne, mais puisqu'elle n'est pas une moyenne, elle n'a pas le même usage. On compare les deux puisqu'elles cherchent à trouver un milieu à un ensemble statistique, mais la moyenne cumule les individus statistiques (et leurs valeurs), la médiane sépare, et en cela elle est bien plus un outil de tri.
   Le mode c'est la valeur dont l'effectif est le plus élevé. Donc pour calculer un mode, on fait l'effectif de chaque valeur, on sélectionne le plus élevé, et sa valeur correspond au mode. Cela n'a d'intérêt que lorsque l'on remarque une valeur semblant extrêmement dominante dans la population statistique.

4. La médiale, si l'on veut c'est la médiane des pourcentages : elle partage une série de données ordonnées en deux parties égales : 50 % des valeurs sont inférieures ou égales, et 50 % sont supérieures ou égales. Ainsi, elle a un intérêt comparable à celui de la médiane, c'est-à-dire une fonction de tri en divisant les valeurs en 2, ici en 2 parts égales.
   A mes yeux l'indice de Gini est plus intéressant. Cet indice est l'outil le plus commun pour mesurer les inégalités de revenus et de richesses. Compris entre 0 (égalité parfaite) et 1 (concentration totale), l'indice de Gini est un nombre décimal qui, plus il est élevé, plus il révèle une situation d'inégalité de répartition. On représente souvent l'indice de Gini avec la courbe de Lorenz : plus la courbe de Lorenz (qui, si j'ai bien saisi, correspond peu ou prou à la médiale) s'éloigne de la bissectrice qui correspond à la moyenne, plus les inégalités sont fortes. Alors, je sors un peu de la question mais je tiens à souligner le fait que cet indice souligne un type d'inégalité qui peut être facilement chiffré, c'est-à-dire les inégalités de revenus et de richesses. Cependant, un grand nombre d'inégalités échappent à la dimension financière, en particulier d'un point de vue social et géographique, et l'indice de Gini ne peut pas mesurer toutes les formes d'inégalités. C'est néanmoins un outil fort utile.

5. Paramètres de dispersion
    - On calcule la variance plutôt que les écarts à la moyenne parce que la somme des écarts est toujours nulle. La variance utilise les écarts au carré, ce qui permet de mesurer la dispersion réelle des données. Cela donne une idée plus précise de la distance entre les valeurs et leur moyenne.
    - L'écart-type en est la racine pour une interprétation plus simple. En somme c'est le retour à l'ordre de grandeur des autres valeurs. En revenant à la même unité que les données on s'assure une facilité d'interprétation.
    - L'étendue permet de se représenter la dispersion de la population, en cela, et même si elle manque de précision car reposant sur les extrêmes valeurs minimales et maximales, elle permet d'avoir une première vue sur l'extension maximale des valeurs dans une population.
    - Les quantiles sont des valeurs qui partagent une population en parts égales, c'est un outil de tri des individus statistiques. La médiane est un quantile qui sépare la population en deux. On utilise aussi souvent en statistique les quartiles, c'est-à-dire une division de la population en quatre parts égales.
    - Une boite à dispersion, aussi appelée boite à moustache permet de cumuler les outils statistiques évoqués ci-dessus dans une même représentation graphique. Ainsi on y représente l'étendue, la médiane, les quartiles 2 et 3, c'est-à-dire le deuxième et le troisième quart de l'étendue, qui forment ensemble un rectangle avec la médiane pour centre. La boite a pour avantage de révéler la dispersion plus précise des valeurs, et révèle aussi la symétrie, mais plus souvent la dissymétrie de la dispersion plus ou moins orientée vers la valeur minimale ou maximale.
6. Paramètres de forme
    - Les moments centrés sont les moments de l'analyse statistique qui se font autour de la moyenne, les variances et les asymétries au premier chef. Les moments absolus ce sont les moments liés à une valeur, mais sans liens directs avec la moyenne. C'est l'équivalent des moments centrés mais plus autonome des valeurs minimales et maximales. On utilise les moments pour décrire la distribution de la population.
    - L'avantage de la symétrie statistique c'est qu'elle facilite la lecture et l'analyse en faisant coïncider moyennes et médianes par exemple, ou en fiabilisant l'usage de l'écart-type. Pour vérifier la symétrie il existe un coefficient qui, si j'ai bien saisi après y avoir passé ma matinée, plus il est proche de 0 plus il révèle une symétrie de la distribution. Si le coefficient est supérieur à 0, on parle d'asymétrie à droite, c'est-à-dire vers les valeurs

maximales, à l'inverse, le coefficient inférieur à 0 traduit une asymétrie à gauche, c'est-à-dire vers les valeurs minimales.

Code :

A nouveau rien ne semble fonctionner quand j'essaie de coder quelque chose. Alors une nouvelle fois j'ai reçu le soutien des mes camarades, et, pour une fois, il me semble avoir réussi à avoir quelques effets.

Conclusion :

J'en reviens à ma conclusion de la séance précédente, je réserve l'essentiel pour la conclusion finale et globale, en espérant trouver un peu plus de sens dans les séances suivantes. M'échappent toujours le sens et la finalité des exercices, sans doute en grande partie par l'hermétisme du codage, et dans une certaine mesure par le caractère abstrait du cours Dans la mesure où nous avons établi les bases, cela devrait s'éclairer. Je m'alarme tout de même de vous avoir entendu dire hier à Clignancourt à un camarade en difficulté sur la question 3 de cette séance, comme je l'étais moi-même, que « Tout est dans le cours ! » et qui sonnait beaucoup comme un « Débrouillez-vous ! ». J'y reviendrai sans doute en conclusion mais j'en ai été refroidi pour vous demander de l'aide sur la même question.

Séance 4 :

Questions :

1. J'ai le sentiment de répéter d'une séance à l'autre, déjà à la Séance 2, puis à la Séance 3, puis à celle-ci. Donc cf : Séance 3 Q2 & Séance 1, Q.8, c. Néanmoins, je reprends. Le choix d'une distribution statistique à variables discrètes ou continues est corrélé au type d'objet que l'on souhaite traiter. Ainsi, pour tous les objets dénombrables dont l'unité est indivisible (je rappelle comme indiqué plus haut qu'il n'est pas question de compter des demi-enfants par famille) et qui ne peuvent en aucun cas être représentés par un nombre à décimales, on appliquera une distribution statistique à variables discrètes. Dans une certaine mesure on préfèrera cette distribution pour les représentations graphiques puisqu'elle est plus adaptée et plus facile à mettre en place. A l'inverse, tous les objets indénombrables, ou dont la mesure est continue (distance, température, indice…), donc tout ce qui est mesure ou intervalle, on préfèrera une distribution statistique à variables continues. Alors bien sûr c'est de manière générale. Le choix dépend à la fois de la nature des données, de leur forme, et du contexte d'étude.
2. Il y a pléthore de lois statistiques très utilisées en géographie. Parmi les plus utilisées j'en citerai 3, qui me paraissent les plus importantes, et que je peux le mieux expliquer. Je laisse donc de côté les lois Poisson et Pareto, que je sais

importantes mais que je ne saurais expliquer, et, du peu que j'en ai compris, qui me paraissent moins cruciales que les trois lois suivantes.

a. Loi de Gauss : ou loi normale est une loi de distribution statistique à variables continues, on la représente généralement avec une courbe en forme de cloche puisque la loi décrit des phénomènes symétriques autour d'une moyenne. On l'utilise pour représenter la distribution de variations démographiques, de températures, d'altitudes. On la voit régulièrement aussi pour illustrer la distribution dans la population générale du taux de Q.I.
b. Loi Log-normale : autre loi de de distribution statistique à variables continues avec des valeurs toujours positives, mais contrairement à la loi de Gauss, ici asymétrique vers la droite. Ainsi, la moyenne y est inférieure à la médiane, ce qui explique cette asymétrie. On l'utilise pour illustrer des inégalités spatiales ou des surfaces par exemple.
c. Loi de Bernoulli : loi de distribution statistique à variables discrètes la plus utilisée, cette loi permet de modéliser une situation qui ne peut avoir que deux états, l'un excluant l'autre : soit 1, soit dans le cas contraire 0. Donc la loi de Bernoulli permet de comptabiliser la présence ou l'absence de quelque chose, toujours selon une modalité duelle : oui/non, présence/absence, équipe 1/équipe 2 …

Code :

Vous verrez sans doute que j'ai essayé à nouveau de faire des choses, mais encore sans succès. Je n'ai plus le cœur à demander encore le soutien de mes camarades. De toute façon l'objet du code ici de de mettre en action des lois statistiques que je n'arrive pas à comprendre.

Conclusion :

Je ne suis qu'à la moitié des séances que je dois mener pour effectuer le parcours débutant et pourtant je commence à éprouver une certaine lassitude. L'Analyse de données me prend un temps fou, pour des résultats nuls, au sens de ni de bonne qualité, ni utiles. Et encore, quand j'arrive à produire des résultats. J'ai le sentiment d'évoluer dans les étapes du deuil : je suis passé du déni à la colère. Colère parce que je me frustre de ne voir rien avancer, en particulier le code, si bien que, là où j'en suis-je pense à abandonner purement le codage pour me concentrer sur les questions, même si je n'en comprends pas grand-chose de plus, et lorsqu'enfin j'ai le sentiment d'avoir saisi quelque chose, ça m'est inutile. Ce n'est pas directement en lien avec la séance mais je commence à m'alarmer de voir mes camarades qui soit abandonnent, soit se refusent à commencer le dossier. Il me semble que mon pessimisme initial était fondé, je ne doute plus que ma conclusion finale se fasse l'écho de ma déception. Note du 16/12 : C'est le cas.

Séance 5 :

Questions :

1. L'échantillonnage c'est sélectionner et analyser une partie (l'échantillon) de la population statistique sinon impossible à traiter dans son ensemble. Cette pratique permet un considérable gain de temps et de ressources, et est absolument incontournable pour les sondages et enquêtes de terrain d'ampleur. Alors tout l'enjeu est de construire un échantillon représentatif de la population générale. On dispose pour cela de deux méthodes : le hasard et le choix. Le hasard est la plus simple : soit on tire au hasard la population de l'échantillon, soit on prend un individu tous les X individus, ce qu'on appelle l'échantillonnage systématique. Alors, si on est un peu rigoureux, il vaut mieux vérifier par quelques calculs la représentativité de l'échantillon, le hasard comme on l'a dit à la Séance 2, n'est pas franchement l'ami des statistiques. Sinon, autre méthode pour s'assure de la représentativité de l'échantillon : le choix. Cette méthode implique un tri préalable et relatif de la population, soit avec un échantillonnage stratifié, c'est-à-dire en prélevant dans chaque strate de la population, strates préalablement définies, soit avec une méthode d'échantillonnage par quotas, c'est-à-dire respectant certaines proportions de caractères prédéfinis. Le choix de la méthode est relatif aux conditions de production de l'étude : taille de la population, rigueur de l'étude, financements, temps disponible… Je note aussi la méthode Monte-Carlo, mais je ne la comprends pas, donc, sachez que je connais son existence, mais que j'ai abandonné, à force d'échecs, l'idée d'en comprendre le sens.
2. Un estimateur c'est un calcul sur un échantillon dans l'objectif de tirer des conclusions sur la population totale. L'estimation c'est le résultat de ce calcul. En d'autres termes : le calcul de la moyenne d'un échantillon est un estimateur pour la population totale, la moyenne de ce même échantillon est une estimation pour cette même population totale.
3. Les intervalles de fluctuations et de confiances désignent en vérité des marges d'erreur. Cependant, la première, l'intervalle de fluctuation, mesure un écart entre l'échantillon et une proportion connue. En cela elle révèle les limites des échantillons qui « synecdocquisent » une population statistique, puisque par la nature même de la réduction, l'échantillon ne peut pas être complètement représentatif. En revanche, les intervalles de confiances désignent elles l'écart entre l'échantillon et une proportion inconnue. On estime donc une proportion probable, et à partir de cette proportion imaginaire, souvent floue (« entre 10 et 15% »), et on déduit un intervalle de confiance. Alors cela peut être fait de manière plus ou moins rigoureuse, les intervalles de confiance sont d'ailleurs au cœur des processus de « redressements » des sondages effectués par les instituts qui les

mènent. Néanmoins, vu depuis la rigueur mathématique, on ne m'ôtera pas de l'idée que c'est une statistique du « doigt mouillé ».

4. Un biais pour un estimateur c'est l'écart entre le résultat obtenu et la proportion réelle, et donc attendue. Si l'estimateur a tendance à donner des valeurs trop importantes, on parle de biais positif de l'estimateur, tandis que dans le cas contraire où l'estimateur à tendance à donner des valeurs trop petites on parle de biais négatif.
5. Les statistiques traitant d'une population totale sont dites « exhaustives ». Cela dit, l'émergence des Big Data, c'est-à-dire des données massives, qui, allant avec une augmentation de la puissance de calcul des ordinateurs, permettent de traiter exhaustivement de très larges populations. Alors l'usage des Big Data complexifie grandement la statistique, et leurs traitements sera inenvisageable sans ordinateurs, si bien que c'est un traitement couteux, en tout cas pour l'instant nettement plus que l'échantillonnage, même si cela a l'avantage de l'exhaustivité échappent ainsi aux questions de représentativités.
6. Un estimateur, comme tous les outils, a une efficacité qui dépend de son usage. Donc, en fonction du type de données que l'on étudie, l'estimateur doit réduire et le biais, et la variance. Ces deux réductions sont souvent difficilement conciliables, alors en fonction de l'objet étudié et de la situation, il faut adapter le compromis entre les deux. Cela dit, il doit aussi ne pas être perméables aux valeurs extrêmes et aberrantes. L'enjeu du choix de l'estimateur se fonde sur le compromis entre sa fiabilité et sa robustesse.
7. On compte plusieurs méthodes d'estimation d'un paramètre : méthode bayésienne, méthode par vraisemblance, méthode par moments … En fonction des données et de la nature des paramètres, la méthode à utiliser diffère. Néanmoins, il y a deux méthodes, les plus employées qui permettent de défricher un peu le terrain : l'estimation ponctuelle qui s'appuie sur une valeur unique (cela peut être la moyenne), et l'estimation par intervalle, plus générale que la première, et qui compte sur la probabilité d'inclure le paramètre recherché. Cette dernière méthode s'appuie sur l'intervalle de confiance, c'est-à-dire sur la mesure d'un écart avec une valeur probable attendue.
8. Un test statistique c'est un calcul de vérification d'une valeur vis-à-vis de la population statistique. Selon le type de variable et l'objectif, on utilise différents tests : pour les variables qualitatives, on retrouve le χ^2 (adéquation ou indépendance) et le test exact de Fisher ; pour les variables quantitatives, les plus courants sont le t de Student (comparaison de moyennes), l'ANOVA (plus de deux groupes) et les tests non paramétriques comme Mann-Whitney ou Wilcoxon. Des tests de corrélation, comme Pearson ou Spearman, servent à étudier l'association entre deux variables.

   La création d'un test statistique suit toujours la même logique : définir les hypothèses H_0 et H_1, choisir une statistique de test adaptée, déterminer le niveau

de signification (α), calculer la statistique à partir des données, comparer avec la valeur critique et prendre une décision. Le choix du test dépend du type de variable, de la taille et distribution de l'échantillon et de l'objectif du test.

9. Généralement je ne pense pas grand-chose de la statistique, alors de la statistique inférentielle… Encore une fois, il s'agit d'un outil. Le reproche fait à la statistique inférentielle d'imprécision et de probabilisme n'a pas lieu d'être, aucun outil ne peut faire de la divination. Quand il faut traiter avec des données incomplètes, et qu'il faut avoir une dynamique prévisionnelle à partir des statistiques, alors la statistique inférentielle est incontournable. Sinon on emploie une méthode statistique adaptée, plus rigoureuse.

Code :

Je vous l'épargne.

Conclusion :

J'avance dans mes étapes du deuil : j'en suis à l'acceptation. Cette acceptation est en grande partie due au fait que j'ai réalisé que j'avais un mauvais état d'esprit : je m'étais habitué ces quatre dernières années à envisager le travail en études supérieures comme devant « produire de la pensée ». Or je réalise que la pensée n'a pas sa place ici, il me faut produire du code, et, à défaut, reproduire du cours. J'ai fait la connaissance lors de ce cours avec la statistique du « à peu près », c'est-à-dire que comme à mo, habitude j'en comprends un mot sur quatre, mais en plus ce n'est pas fiable ! J'essaye d'en rire, ça soulage. Et je suis aussi dans l'acceptation de ma propre incompétence pour pouvoir produire du code. J'ai tenté quelques lignes, et puis finalement, n'arrivant comme d'habitude à rien j'ai baissé les bras. Alors je vais comme à la catastrophe, avec le sourire. J'ai des camarades qui comptent faire les 5 séances le vendredi 19 même, et je leur souhaite d'avance bien du courage tout en espérant qu'ils soient et des génies et des foudres-de-guerres pour abattre tant de travail en si peu de temps. D'un autre côté je prends avec philosophie, c'est-à-dire un stoïcisme heureux, ce que je lis sur le coefficient de notation des séances. Ainsi, ces coefficients augmentent avec les séances et la difficulté, et cela à la fois pour les questions de cours et pour le code ? Quid de ceux qui, vaincus auront abandonné ? J'ai moi-même abandonné le code. En vérité ce sont surtout les questions qui m'interrogent, il y en a nettement moins que sur les premières séances mais elles valent largement plus, si bien que je m'oblige à une exhaustivité forcée, alors qu'en fait je ne pense plus bien comprendre ce que les questions demandent, ou si par miracle je comprends quelque chose, je ne vois pas l'intérêt de développer plus (cf estimateur et estimation). Tant pis pour nous je suppose ? « C'est ainsi. » disait les stoïciens. C'est ainsi… L'échec est plus doux quand on l'accepte. Mais ne faisons pas attendre le désastre : il me faut y courir. Je consacrerais mon dimanche, c'est-à-dire demain, à la séance 6. Au point où j'en suis dans ce chemin de croix, autant aller au bout de ce qui m'est possible.

Séance 6 :

Questions :

1. Les statistiques ordinales concernent des variables qualitatives dont les modalités peuvent être classées selon un ordre, sans que les distances entre elles soient nécessairement égales. Elles permettent d'analyser l'ordre et la répartition à l'aide de mesures comme la médiane, les quartiles ou le mode, ainsi que les fréquences et pourcentages, mais ne permettent pas de calculer des moyennes ou écarts-types fiables. Ces statistiques sont utilisées pour les classements, échelles de satisfaction ou niveaux d'éducation par exemple. Les variables utilisées sont de nature qualitative.
   Encore une fois, la question de l'opposition n'a pas de sens, en particulier lorsque l'on pense à un outil. Il est absurde de penser que le tournevis s'oppose au marteau. De même, si je vois bien où vous voulez m'emmener avec votre question, il n'y a pas d'opposition entre les deux types de statistiques qualitatives que sont les statistiques ordinales et les statistiques qualitatives nominales. Les premières sont d'un point de vue statistique plus fécondes que les secondes, mais il y a une différence de nature qui ne permet en rien de les placer sur un rapport commun pour les opposer.
   La question de la hiérarchie spatiale se pose en fonction des échelles ordinales construites (niveaux de richesses en strates, types d'éruptions volcaniques, échelle de Richter…). Les strates des échelles ordinales ont le bon goût de se laisser aisément représenter en cartographie.
2. L'ordre à privilégier dans les classifications est l'ordre qui semble le plus lisible à l'imagination humaine. La norme est de classer par ordre croissant de caractéristiques, ce qui permet d'illustrer une dynamique d'évolution. Encore faut-il ensuite la démontrer.
3. La corrélation des rangs mesure le lien entre deux variables ordonnées : elle indique si les rangs élevés d'une variable correspondent à des rangs élevés (ou faibles) de l'autre. La concordance des classements, elle, mesure le degré d'accord entre plusieurs classements portant sur les mêmes objets. En résumé, la corrélation des rangs analyse une relation, tandis que la concordance des classements évalue une cohérence.
4. Le test de Spearman mesure l'intensité et le sens d'une relation ordonnée entre deux variables : lorsque les valeurs de l'une augmentent, celles de l'autre tendent à augmenter ou diminuer de façon cohérente. Il est simple à mettre en œuvre et bien adapté aux échantillons de taille moyenne ou grande.
   Le test de Kendall, lui, s'appuie sur la comparaison des rangs pris deux par deux pour évaluer le niveau d'accord entre deux classements. Il est plus robuste lorsque l'échantillon est petit ou lorsque plusieurs valeurs ont le même rang, et son interprétation est plus intuitive en termes de cohérence entre classements.

5. Les coefficients de Yule et de Goodman-Kruskal servent à mesurer l'intensité de l'association entre des variables qualitatives. Le coefficient de Yule, le plus facile à appréhender, s'applique aux tableaux simples à deux modalités, tandis que les coefficients de Goodman-Kruskal évaluent dans quelle mesure une variable qualitative permet de mieux expliquer ou prédire une autre. Et dans ce second cas, si j'en comprends l'idée je n'arrive pas à comprendre comment cale fonctionne.

Code :

Non.

Conclusion :

Cela m'aura pris 2 jours. Donc j'ai pris le temps de la lecture, j'ai essayé de répondre aux questions, mais je suis incapable de comprendre l'intérêt des objets évoqués entre la question 3 et 5. Ces objets ne semblent exister que dans une dimension métaphysique qui leur est propre, et je me sens incapable de les faire fonctionner. Je ne sais plus si le parcours débutant implique la séance 7, mais, même si c'est le cas, je vais m'arrêter là, il me semble avoir eu ma dose de statistique. J'ai relu mes conclusions des séances précédentes, où je plaisantais sur les étapes du deuil et sur la colère. Là, finissant les questions de la séance 6, la colère n'a rien d'une plaisanterie. Je regrette mon ton badin dans la conclusion de la séance précédente, j'ai pensé à le modifier mais je le crois témoin d'une évolution qu'il vous faudra saisir dans ma conclusion globale. La conclusion de la séance précédente c'est l'expression pudique d'un désespéré en quête de sens. Me sentant bouillir en écrivant ces lignes, je vais laisser quelques jours pour faire redescendre la pression, puis j'expliciterai dans une conclusion finale le bilan que je tire de cette expérience. Je sais que beaucoup de mes camarades veulent éluder cette conclusion que vous nous avez demandé, par volonté de s'épargner du temps, de vous épargner des critiques négatives donc désagréables. Pour ma part, je suis peiné de voir dans quel état ce cours me laisse, et je pense qu'il vous faut savoir ce qui n'a manifestement ni fonctionné ni convenu. Une réflexion à chaud, qui ne vaut que ce que la colère peut nous faire dire : je suis fasciné que ce cours puisse rendre plus malheureux à la fin qu'on y était en le commençant.

Je rajoute un mot le 19/12 lors de mon ultime relecture, il y a tout de même au bout la fierté de rendre quelque chose. Quelque chose d'incomplet et de honteux certes, mais quelque chose tout de même. La colère passée, je me suis interrogé sur la pertinence de laisser mes conclusions de séances qui sont complétement des manifestations de subjectivité. Cependant je porte un regard sur ce rapport comme une forme de journal de bord, l'évolution de la psyché de l'étudiant, ici la mienne, est je crois révélatrice du rapport entretenu avec la matière.

Conclusion critique :

Lors de la première séance, vous nous avez demandé de rédiger une critique du parcours et de la pédagogie proposée sans toutefois vous « descendre », selon votre expression. Je ne vous cache pas que j'ai, pour ma part, un bilan très négatif de cette expérience, que je vais préciser ensuite. Je tiens cependant à vous dire que mon expression reflète le plus justement ma pensée et n'est en aucune manière une critique vengeresse ou malveillante ; la virulence que vous pourriez trouver en mon propos est véritablement l'expression pesée, réfléchie de mon expérience de l'Analyse de données. La colère partie, seuls les faits comptent, et je me limiterai à cela dans cette conclusion. Je tiens aussi à souligner dans ce paragraphe liminaire que ma critique sanctionne seulement une expérience individuelle d'un parcours pédagogique, la mienne, vous n'êtes évidemment en rien personnellement visé. Je n'ai rien à gagner à faire votre procès, je n'ai pas le besoin d'une jouissance cathartique dans une violence verbale. Le sens de ma démarche est d'exposer mon point de vue à un enseignant, vous, qui l'a demandé ; et je crois vous le devoir, même si la critique sera sans doute, comme toute critique négative, désagréable.
Cela dit, je vous fais part de ma frustration, de ma colère et de mon dépit vis-à-vis de la formation. Et, malgré un parcours scolaire tumultueux, je crains avoir à vous dire que ce fut sans doute une des pires expériences académiques. Il me semble que, à une seule exception que j'exposerai à la toute fin de cette conclusion, toutes les dimensions de l'Analyse de données m'ont posé problème. Ainsi, je vais essayer de disséquer mon expérience le plus précisément possible, étape par étape, pour que ma critique mise en perspective avec celles de mes camarades puisse vous être utile. Il ne s'agit pas dans ma critique de garder l'objectif en éludant la part subjective de mon expérience, il vous faut je crois saisir **le tout** pour comprendre tous les éléments de ma déception, et ainsi pouvoir peut-être améliorer certains aspects du parcours pour l'année prochaine.

Le premier problème est d'ordre épistémologique. Il y a dans la première question de la première séance à rendre (Séance 2) un parti pris vis-à-vis de la relation entre Géographie et statistiques qui me paraît théoriquement inopérant. Pour répondre à cette première question, il faut s'appuyer sur votre document qui lie les deux disciplines, si bien que je pense pouvoir dire sans caricaturer votre position que vous avancez le fait que la statistique serait presque un pilier la Géographie, et que finalement les géographes se serait trop concentrés sur son aspect « qualitatif ». Ainsi, dès la première question vous cadrez la perception que nous, étudiants, devrions avoir de la statistique dans la Géographie, or c'est une prémisse orientée, sans doute par votre expérience professionnelle, ce qui se comprend tout à fait, mais orientée tout de même. La Géographie, certes, ne se limite pas à la Géographie « qualitative », mais c'est **déjà** cela, c'est **au moins** cela, c'est d'ailleurs sans doute **d'abord** cela. Depuis les logographes

préhérodotiques comme Anaximandre ou Hécatée de Milet, en passant par tous les récits de voyages et d'exploration, et jusqu'à la science universitaire moderne que l'on connait aujourd'hui, la Géographie fut une science humaine et sociale cardinale au même titre que l'Histoire et sous l'égide de la Philosophie, dont le rôle était de comprendre la relation de l'Homme avec l'espace, de répondre à la question sans doute aussi vieille que le langage : « où ? ». La dimension « quantitative » de la Géographie est issue d'un usage d'abord administratif (cf le Domesday Book en Angleterre au XIème siècle), Ératosthène lui-même pense la Géographie comme un moyen d'appréhender l'espace comme un potentiel exploitable, elle en est ainsi une extension annexe, parfois très utile, mais finalement dispensable. Alors entendons nous, elle est effectivement devenue incontournable sur beaucoup de sujets, mais du reste ce n'est qu'une branche d'une science d'abord et avant tout humaine et sociale. Le véritable essor de la « Géographie quantitative » selon Paul Claval n'a lieu que lors de la décennie 1970-1980. Sa survivance, jusque dans votre enseignement, est due à ce que je qualifie comme un complexe dont souffre aussi l'économie qui est celui du rapprochement avec les sciences « exactes », pour « scientiser » ces sciences humaines qui sinon « ne serait pas véritablement des sciences ». La présence de chiffres ne permet en rien une meilleure production de savoirs ou de connaissances du réel, surtout pour caractériser un objet aussi empreint de pratiques et de symboles qu'est le territoire. Sans doute, la confusion vient du fait que la Géographie a « enfanté » des sciences dites « dures », comme al géodésie, la géologie ou l'hydrologie pour ne citer qu'elles. Or, les « enfants » de la Géographie n'ont pas à modifier l'approche que l'on a de la discipline, ce sont des apports, des colorations et non des transformations, il en va de même des statistiques. On dit que les mathématiques sont issues de la Philosophie, il serait absurde de reprocher à la Philosophie de n'être pas une science parce qu'elle manquerait de chiffres ! La Géographie bénéficie toujours des apports des pluridisciplinaires, en particulier venus des sciences « dures », elle bénéficie des apports des statistiques, mais ce sont des apports. Finalement, les statistiques sont utiles en Géographie, mais elles ne sont qu'un outil. Et, puisqu'elles ne sont qu'un outil, tous ne l'utilisent pas de la même manière, tous n'en ont pas le même besoin. Pour les étudiants en Master Geoint ou Enviterr, les statistiques sont peut-être incontournables, pour les étudiants en Master SCT ou GéoSuds, c'est-à-dire plus proches de l'étude des dimensions culturelles et des espaces, la statistique est un outil bien moins utilisé. Au bout du compte, dès la première question du rapport vous souhaitez nous faire admettre l'importance des statistiques en Géographie, importance pourtant toute relative vis-à-vis de la discipline prise dans son ensemble, et relative vis-à-vis de la pratique de la Géographie menée par les étudiants (en master GéoSuds, l'analyse statistique est presque absente, car peu utile). Voilà le fait : les statistiques ne sont en rien un pilier de la Géographie, ni d'un point de vue historique, ni d'un point de vue ontologique, ni d'un point de vue pratique. Ainsi, il est incompréhensible pour nous, étudiants, d'avoir à cumuler le cours d'Analyse de données et le cours de Méthodes quantitatives, quand la réalité de la pratique de la Géographie, en particulier la nôtre, s'appuie finalement si peu

sur la statistique, et que l'usage de la statistique, ainsi que l'importance qu'on lui accorde tiennent bien moins de l'intérêt scientifique que du vernis de la précision mathématique.

Le deuxième problème est de l'ordre de la présentation du cours. Un grand nombre d'étudiants du master sont très distants vis-à-vis des mathématiques et *a fortiori* vis-à-vis du codage. C'est mon cas, je sors quatre années de CPGE AL, c'est-à-dire quatre ans d'études intensives purement littéraires. Il faut bien saisir que, depuis toujours, et surtout lors de mon hypokhâgne, de ma khâgne, de ma khûbe et enfin de ma bicarre, j'ai appris à utiliser deux outils : les lettres et le papier. Avant cette année, je n'avais pas eu le moindre contact avec les mathématiques depuis la classe de seconde au lycée, c'est-à-dire depuis 6 ans. Vous répondrez sans doute que l'Analyse de données c'est plus de l'informatique que des mathématiques. Soit, mes contacts avec l'informatique furent Civilzation 6, Skyrim, que je n'ai pas touché ces quatre dernières années, prépa oblige, et surtout presque exclusivement Word. Donc il faut mesurer le chemin à parcourir pour quelqu'un comme moi, ne serait-ce que pour arriver à la Séance 2. Jusqu'en septembre je n'avais jamais eu l'occasion de rencontrer les statistiques ou l'informatique. Alors la phrase « ouin, ouin, je suis un littéraire, je ne comprends rien aux mathématiques », je vous cite ici, c'est dans le propos liminaire de votre cours sur GitHub, moi non plus je ne veux pas l'entendre. On fait difficilement plus vexant pour quelqu'un qui est si purement littéraire que le codage lui paraît aussi surnaturel que la sorcellerie. Et lorsque l'on souffre de patauger dans une discipline qui nous est en tout point étrangère, jusqu'à la démotivation, au désespoir et enfin au dégoût, et que l'on a consacré quatre ans de sa vie aux Humanités littéraires, un tel commentaire semble profondément insultant, injurieux. Voyez-vous, l'irrespect (parce que, même si ce n'était pas votre intention, cela en est) n'est probablement pas le meilleur moyen d'inciter des étudiants totalement débutants à s'intéresser à votre cours. Là il ne s'agit pas d'une extrême susceptibilité de ma part, c'est qu'il y a ici un propos qui est déplacé, infantilisant, humiliant, et qui s'adresse à au moins une personne qui, initialement de bonne volonté, se retrouve à bûcher sur un cours aussi éloigné de sa formation initiale que peut l'être l'enseignement de l'araméen dans un CAP de métallurgie. Je n'exagère pas tant que vous pouvez le penser, la réflexion mathématique et informatique implique un système de réflexion et d'imagination profondément alien à un système littéraire, c'est à la fois une autre langue, une autre culture, une autre manière de penser et finalement presque un autre monde. La différence de réflexion n'est pas une simple question d'habitude, c'est véritablement un fossé abyssal, fossé dont le comblement aurait sans doute dû se trouver au cœur du cours d'Analyse de données, quand bien même cela aurait été incomplet. Et je ne suis pas un cas isolé, même si les bicarres ne se bousculent pas au portillon, je le reconnais, il y a un grand nombre de masterants qui n'avait jamais eu la moindre expérience ni du codage ni de la statistique. Je pense que vous avez sous-estimé l'ampleur de ce fossé qui sépare une grande partie de vos étudiants des mathématiques et du codage. Parce que, de mon point de vue, le codage tient plus de la magie que des mathématiques : je vois ma

console faire des choses, je ne sais pas ce qu'elle fait, ni d'où cela vient, ni où cela va, ni si cela a fonctionné (je ne suis jamais sûr que mes fichiers arrivent bien mon GitHub puisque j'ai toujours un mal fou à les retrouver, si bien que je ne vérifie même plus). Rien ne semble être la conséquence de quoi que ce soit, autre que l'incantation qu'est pour moi une ligne de code. Encore une fois : je ne suis pas un cas isolé, le contact avec l'informatique, et la reprise du contact avec les mathématiques ont été trop brutaux, et ont donc provoqué un mouvement de réactance, de rejet de la discipline. Il faut aussi saisir que rien n'indiquait lors de l'inscription, dans le cas du master du GéoSuds, que nous, étudiants, aurions à faire de la statistique, l'intitulé « analyse de données » est suffisamment opaque pour que, le jour de la présentation du master, nous soyons mis devant le fait accompli. Enfin, lors de la première séance, vous nous avez accueilli avec une véhémence, une irritation probablement due aux manquements de l'administration en matière d'informatique, mais qui a eu pour effet de nous braquer : nous considérions désormais que le cours d'Analyse de données était le cours où nous étions susceptibles de nous faire engueuler. Alors cela n'a pas eu lieu, et sans doute que probablement cela n'aurait jamais eu lieu, mais cela a teinté chaque cours d'appréhension, ce qui n'est pas un facteur motivant, bien au contraire. Pour ce qui est du livret que vous nous avez distribué à la première séance, c'est très simple : je n'y ai rien compris, d'ailleurs si bien qu'il m'a fallu votre aide pour tout réinstaller. Du reste, cela aurait pu être le manuscrit de Voynich : je n'arrive pas à en faire sens. Cette critique là est moins forte que les précédentes ou que les suivantes, il semble que je sois le seul à être resté perdu, interdit comme une poule devant un couteau, face au manuel. Mais sans doute y-a-t-il peut-être un petit quelque chose à faire pour le rendre un peu plus accessible.

Le troisième problème est de l'ordre de l'organisation du cours. Pour ce qui est de la pédagogie inversée, selon moi elle a été à l'origine du problème qui est devenu peu à peu, autant pour moi que pour mes camarades, un désastre aux airs de catastrophe industrielle. Alors, je vais vous faire une confession, mais sans doute l'aurez-vous vu dans les autres copies : peu ont lu le cours. Et pour être franc c'est totalement mon cas. Alors pour le comment nous avons répondu aux questions, j'y reviendrai aux parties suivantes. Toujours est-il que, puisque nous ne voulions pas prendre du temps personnel pour un cours si éloigné de nous qu'il nous en était répulsif, nous n'avons pas lu le cours. Or, lorsque nous venions en séance, soit nous lisions distraitement un bout du cours de la séance du jour pour faire passer le temps, soit nous nous démêlions avec nos problèmes informatiques, soit nous nous faisions aider par un camarade qui, parce qu'il avait suivi une formation en informatique avant cette année, avait la capacité de produire un peu de chose sur nos machines. Donc, dans mon cas, même si j'ai essayé de lire les quatre premiers cours, il ne m'en reste rien. Je fais le constat de n'avoir strictement rien appris du cours, à la fois parce que je n'y ai rien compris et parce que, peu à peu, j'ai perdu toute envie d'en apprendre quoique ce soit. Alors je lisais, de manière automatique, ou plutôt je regardais les mots… Sans doute, le fait que le cours fasse plusieurs centaines de pages

n'a pas aidé. L'enseignement d'une discipline nouvelle, qui est étrangère par définition aux étudiants implique de ne pas penser qu'ils, par je ne sais quelle curiosité surnaturelle, se lancent avidement dans l'apprentissage de cette discipline sur leur temps libre. Pour employer une pédagogie inversée il faut assurer en amont que les étudiants aiment la discipline, et se soit inscrit dans le parcours d'enseignement pour le recevoir, à cette condition et à cette condition seule, la pédagogie inversée peut être efficace. Or, quand on souhaite enseigner une discipline nouvelle à des étudiants, discipline d'ailleurs probablement impopulaire ou pouvant provoquer une hostilité de principe (ce qui n'était pas mon cas, au contraire j'étais disposé au départ à rencontrer une discipline nouvelle), cela implique ou l'apport de l'apprentissage en cours, sur le mode du CM, liés à des exemples connus ou abordables, ou une maïeutique patiente et individualisée pour chaque étudiant, donc extrêmement difficile à mettre en pratique. Peut-être peut-on imaginer un type d'enseignement intermédiaire avec des cours autour d'exercice de groupes ? S'il fallait refaire le cours, il ne pourrait fonctionner selon moi qu'en CM, avec 1 page maximum de document par cours, sinon, le premier réflexe de beaucoup sera de ne pas lire le cours, sinon l'abandon pur et simple. Autre problème fondamental : les téléchargements et installations. Combien de dossiers, de programmes, de scripteurs, sans compter docker avons-nous dû installer pour faire ce cours ? Je ne sais plus. Concrètement, dossiers mis-à-part, il a fallu se familiariser avec Docker qui, pour résumer à grands traits, n'a pas fonctionné pour la moitié des étudiants, l'autre moitié n'a pas su s'en servir. Puis il a fallu se familiariser avec GitHub pour récupérer les données de cours, et donc rendre ce travail. Or, il y a déjà une structure qui permet cela, qui est commune à tous les étudiants et qui est très nettement plus facile à prendre en main : moodle. Votre mantra étant de nous « faire gagner du temps », tous nous avons pris des heures à installer GitHub (au moins deux lors de la deuxième séance), parfois sans succès, qui pour l'emploi que nous en faisons ne fait rien de plus que moodle, et que personne n'aura à, ni ne voudras réutiliser par la suite. Je m'interroge encore sur l'utilité de GitHub, comme bien d'autre camarades, et il me semble finalement que ce fut un énième téléchargement invasif de la mémoire de mon PC qui crache déjà ses poumons à lancer Excel. Ensuite il a fallu installer et se familiariser avec Discord, réseau social dont on devait « animer le serveur », votre serveur, sans penser que discord était une interface particulière pensée d'abord pour le gaming pas tout à fait adaptée à des néophytes, pour obtenir des points supplémentaires « d'implication » (je vous cite toujours). Là, il y a un problème, c'est que vous avez fait de ce que vous pensiez être une carotte un bâton : poursuivre des points, pour un travail qui nous semble hors de notre sujet d'étude, hors de nos compétences, et qui, nous le pressentions tous, s'en allait à la catastrophe, ne valait pas le coup. De plus, y chercher des points pouvait apparaître comme votre caudataire, dans le serveur sous **votre** regard. Et donc en conséquence, chacun s'empêche de parler sur le serveur, chacun s'observe à qui par un message de trop trahira son avidité du point. Le résultat est qu'il me semble que la participation sur le serveur est un échec, parce qu'elle se fait sous le regard de l'enseignant, et que participer

c'est apparaître aux camarades comme se mettant en scène aux yeux du prof' pour obtenir ses grâces. A la seule exception de notre camarade Zara qui, forte de son expérience en informatique a, très largement soutenu la promotion de notre master GéoSuds, et probablement des autres, jusque sur discord. Il me semble que vous avez indiqué qu'aider ses camarades permettait d'obtenir des points bonus, aucun doute : elle les mérite. Et je veux ici dire mon estime et ma reconnaissance. Sans que cela ne retire rien à l'aide qu'elle nous a, à tous, apporté, il faut dire aussi que c'était probablement la seule à pouvoir le faire. Encore une fois, elle aurait très bien pu ne pas le faire, et nous ayant aidé, elle nous a permis de rendre quelque chose, aussi pitoyable que cela puisse être, mais d'un autre côté, personne d'autre qu'elle dans la promotion n'aurait pu nous aider. C'est d'ailleurs sans doute pour cela que Zara demeure vierge de tout soupçon de courtisanerie : elle n'a pas à chercher des points sur discord puisque l'on lui doit les nôtres. J'insiste : sans Zara, pas de rendu, *a fortiori* le mien. Les quelques autres qui s'en sont sortis, ont soit été formés à quelques rudiments de l'informatique au lycée, soit reçu une aide extérieure, soit se sont appuyés massivement sur l'intelligence artificielle.

Le quatrième et ultime problème est de l'ordre du travail du cours. Soyons francs, quitte à jeter tous les camarades avec moi sous le bus : le rendu du dossier n'est possible qu'après un pillage massif, systématique, industriel des données. Soit refilées de mains en mains par des camarades généreux, soit par l'intermédiaire d'un autre pilleur que soi, c'est-à-dire un logiciel d'intelligence artificielle et, quitte à nommer le coupable : Chat GPT. Tout le monde a utilisé Chat GPT. J'ai utilisé Chat GPT. Et pour être clair, ce n'est pas une utilisation raisonnée, quelques questions çà et là pour éclairer un point. Non, il s'agit de télécharger votre cours, le verser dans la machine et lui demander un résumé, mieux de fournir directement les lignes de code. Si bien que votre travail d'évaluation consistera sans doute surtout à corriger des prompts générés par la machine. Je le déplore, pour vous déjà, parce que rien ne doit être plus désagréable, mais aussi pour nous, puisque finalement nous aurons plus appris à nous reposer sur la machine qu'appris le codage ou les statistiques. Il faut bien saisir là que le recours systématique n'est pas l'expression d'une paresse, c'est une pratique qui répond à des besoins : gagner du temps sur une activité ultra chronophage, s'épargner un enseignement douloureux, et surtout produire de la matière pour remplir le dossier. Il n'est plus question de demander à la machine d'expliquer le cours, il y a un dossier à remplir, alors on remplit avec ce que la machine a saisi du cours qu'on lui a fait ingurgiter. Par honnêteté intellectuelle, j'ai trouvé dans l'ensemble des camarades avec qui j'ai pu discuter du cours, un seul et unique camarade qui m'a indiqué qu'il n'avait pas utilisé une seule fois l'I.A. : il s'agît de Nell Cousin qui, en conséquence, n'a pas fait le code. Jusqu'à aujourd'hui je m'étais toujours opposé par éthique à l'usage de l'intelligence artificielle générative, non par hostilité au progrès mais pour des raisons morales, notamment en ce qui concerne le pillage de la propriété intellectuelle. Mais là j'ai été contraint. Et d'un point de vue personnel, moral, c'est affligeant. Alors, d'une question à l'autre j'ai reformulé, j'ai vérifié, parfois dans un ultime

effort j'ai ouvert le cours et regardé si avec un crtl-F se trouvait quelque chose qui ressemblait vaguement à ce que je bricolais, mais finalement, mon dossier tout entier tient sur la béquille de l'I.A. On dira, lorsque l'on n'est pas mis en situation devant le fait accompli qu'on a toujours le choix, que cela n'est qu'une question de volonté ou de prévision. C'est faux. Même si j'avais consacré 2 heures de mon temps, chaque soir, je n'y aurai rien compris, parce que tout simplement ce n'était pas expliqué, ou du moins que les explications ne m'étaient pas accessibles, et que cela est trop loin de moi. Mais en vérité, la démotivation, le désespoir et le dégoût m'en ont débarrassé avant. Peut-être répondrez-vous que je n'avais, que nous n'avions qu'à vous demander. Or c'est une fable. D'abord, l'incompréhension de la discipline est trop forte, le travail demandé trop important pour que, sans cesse, et je vous assure que cela aurait été **sans-cesse**, nous venions vous demander sur discord de l'aide. Ensuite, vous êtes apparu comme sévère dès le premier cours, ce qui n'a pas aidé à donner envie de demander votre aide. Mais surtout, aucun étudiant ne veut avoir à demander plusieurs fois de l'aide à un professeur, par timidité en partie mais, et en grande partie, aussi par distance avec la matière. À quoi bon s'échiner, même soutenu par l'enseignant, sur une discipline qui paraît ingrate en tout point ? Une dernière chose que l'on m'a fait remarquer aujourd'hui qui cumule les problèmes de discord, de GitHub et du pillage : le lieu de dépôt. Depuis quelques jours, les camarades rendent leurs dossiers GitHub sur le discord du cours, dans un salon public, si bien que tout le monde a accès à leurs codes, à leurs réponses, à leur dossier. J'avoue n'y avoir pas cru, donc je suis allé vérifier, et je peux effectivement ouvrir les copies et les codes de mes camarades sans que rien ne m'en empêche. En d'autres termes, tous les retardataires ont eu la possibilité de piller joyeusement les travaux des camarades les plus assidus, sans être aucunement inquiétés puisqu'il suffit de modifier un peu l'ensemble pour laisser penser que l'on a fournit un travail original. A ma connaissance ce pillage n'a pas eu lieu, au moment où j'écris le 15/12/2025, personne ne m'a indiqué avoir utilisé cette méthode, et si ce passage est toujours présent dans ce texte lorsque j'ai rendu le dossier c'est que cela n'a pas changé. Moi-même je n'ai évidemment pas utilisé cette méthode, dans la mesure où je me targue d'un minimum d'éthique, et aussi, comme vous avez pu le lire, d'un minimum d'honnêteté. J'ajoute le 18/12 que vous avez indiqué comment protéger son GitHub, mais en vérité c'est déjà trop tard. Cependant, le pillage a été possible, si bien que la sécurité des travaux du premier étudiant à avoir rendu le dossier par discord est compromise, et tous les travaux rendus ensuite sont suspects du pillage du premier.

Pour résumer, j'estime que vous nous avez contraint à travailler les statistiques en les travestissant comme plus importantes en Géographie qu'elles ne le sont, que vous avez utilisé une communication rigide voire répulsive à l'égard des étudiants, en particulier pour ceux les plus distants de l'informatique, que vous avez contraint à un grand nombre de téléchargements sur nos ordinateurs personnels (docker, GitHub, discord, Notepad++, et tous les fichiers à télécharger) au point de donner le sentiment

d'un enseignement invasif, que vous nous avez contraint à un travail personnel non proportionné à l'importance réelle des statistiques et du codage en Géographie et en usant d'une pédagogie hautement inadaptée aux nouveaux étudiants, et enfin que vous nous avez contraint à travailler, en grande partie dans l'urgence, à l'aide de pratiques et d'outils dont on aurait voulu se passer.

Je sors de cette expérience à la fois frustré, honteux, et dégoûté. Frustré car je n'aurai rien appris du codage, qui reste pour moi après pourtant un semestre d'étude, une pratique aux propriétés magiques, c'est-à-dire sans relations apparentes au réel. Honteux parce que je n'ai pas trouvé dans cet enseignement le moyen de produire une pensée ou un travail que je serai fier d'avoir accompli, soit parce que j'ai utilisé des outils que j'exècre, soit par le recours intensif à l'aide de mes pairs généreux, qui tient malgré tout au pillage de leur travail. Enfin et surtout, dégoûté par ce qui m'apparaît comme une perte sèche de temps et d'énergie. Il faut bien saisir que ce dégoût va au-delà du dégoût du codage ou de la statistique : c'est un dégoût qui infecte mon rapport au master, que je ne portais déjà pas dans mon cœur, et qui affecte même mon rapport à la Géographie. Je précise : sincèrement, le cours d'Analyse de données, certes parmi d'autres cours mais tout de même au premier chef, m'a fait considérer le plus sérieusement du monde la possibilité d'abandonner la Géographie, et même les études. Et je me répète à nouveau : je ne suis pas un cas isolé. Le cours d'Analyse de données n'est pas le seul facteur de ce dégoût massif, mais c'est l'un des principaux. Finalement, à mon sens, le cours d'Analyse de données n'est pas seulement un échec pédagogique : c'est une faillite morale. L'effet concret du cours c'est le rejet du codage d'abord, de la statistique ensuite, et enfin de la Géographie dans son ensemble, et l'effet de la note sur des travaux qui sont rejetés, bâclés, abandonnés ou non rendus, est de descendre des étudiants qui peuvent être bons par ailleurs mais qui ont été victimes d'un module inadapté, chronophage et douloureux. Il est, de mon point de vue, impossible de concilier le soutien des étudiants avec une telle formation qui soit, en tant de points, si inaccessible. Encore une fois, il n'y a pas eu d'abandon direct, en tout cas pas dans pour moi, il y a eu jusqu'à mi-novembre une vraie tentative de faire quelque chose. Mais l'abandon vient irrémédiablement du dépit, dépit inévitable lorsque l'on constate que tous les efforts déployés le semblent dans le vent. Alors on a baissé les bras, j'ai baissé les bras dans le souffle d'un « à quoi bon ? » qui reconnait un échec retentissant.

Cependant, je veux souligner l'aide que vous avez apporté à tous les étudiants qui vous l'ont demandé en cours ou via discord, et cela n'est pas négligeable. Je tiens donc à vous remercier très sincèrement pour cette aide qui m'a permis, le plus péniblement du monde certes, mais tout de même, de rendre aujourd'hui quelque chose. Alors, après 8 pages de critiques que je reconnais volontiers comme acerbe, ce point positif apparaît, j'en conviens, comme un sparadrap sur une jambe de bois. Pourtant il est décisif, il est même la raison pour laquelle j'ai pris le temps de rédiger dans le détail tous les problèmes que j'ai pu trouver dans l'espoir que, pour les prochaines promotions, vous puissiez adapter votre cours. Si votre aide avait été absente, je me serai contenté de vous

considérer comme malveillant et soit je n'aurai rien rendu, soit quelques questions sur dossier à moitié vide (c'est-à-dire encore moins rempli que celui-ci) que j'aurai conclu avec une critique cinglante qui aurait tenu en une phrase. Néanmoins, puisque j'ai pu mesurer concrètement que vous tenez véritablement à aider vos étudiants, j'ai espoir que vous puissiez prendre ma critique, aussi dure soit-elle, comme constructive.
Vous aurez saisi je pense l'ampleur de mon insatisfaction, et la sévérité de ma critique tient aussi à faire parler malgré eux certains de mes camarades qui n'oseront pas vous dire toute la vérité, tenus à distance par le statut sacré, quasi-mystique de l'enseignant. Je prends le risque de vous déplaire, et ainsi d'immoler la note de mon dossier mais qu'importe : vous saurez une vérité que beaucoup n'oseront jamais vous formuler.
Je tiens à rappeler qu'il s'agît là d'une expérience personnelle, il faut donc la nuancer, la diluer dans l'ensemble des critiques de mes camarades, mais, sans surinterpréter les silences, certaines critiques lacunaires ou absentes cachent des avis parfois bien plus sévères que le mien. Aussi je souhaite que mes camarades aient eu une meilleure expérience, vous épargnant l'inconfort de la critique sèche comme peut l'être la mienne, et que ma critique détaillée puisse vous être utile.

Très respectueusement,

R.Pisot