MASTER EN SCIENCES MATHÉMATIQUES

Mobilité et réseaux sociaux : analyse empirique et modélisation

LUCAS, Pauline

Award date:
2013

MASTER EN MATHÉMATIQUES

Mobilité et réseaux sociaux : analyse empirique et modélisation

Pauline Lucas

2013

Mobilité et réseaux sociaux : Analyse empirique et modélisation

Pauline Lucas

Master 2 en Sciences Mathématiques

Année académique 2011-2013

Avant toute chose, je tiens à remercier l'ensemble des professeurs et assistants du département de mathématiques qui, durant ces 5 années, nous ont assuré une formation de qualité, en particulier pour cette dernière ligne droite.

Un merci tout particulier à mon promoteur, Monsieur Lambiotte, pour ses conseils avisés, sa disponibilité et l'aide précieuse apportée au cours de ces 2 années.

Enfin, je remercie toutes les personnes de mon entourage, famille et amis, pour leur soutien, leur encouragement et leur patience lors de la rédaction de ce mémoire.

Mobilité et réseaux sociaux : Analyse empirique et modélisation

Résumé :

Grâce au développement et à l'utilisation courante de la géolocalisation, de nombreux systèmes (parmi lesquels, les réseaux sociaux) permettent d'obtenir des données sur les déplacements des personnes. Au fil du temps, toutes ces technologies ont été utilisées par les chercheurs afin d'élaborer différents modèles destinés à prédire la mobilité humaine et d'en tirer des propriétés universelles. Deux types de modèles ont été créés à cet effet. Les modèles de gravité, largement utilisés depuis de nombreuses années pour prédire le taux de voyageurs en théorie du transport, ont mis en avant l'effet de la distance physique sur la mobilité. Néanmoins, du fait de leur non-universalité (les paramètres variant en fonction de la région), il a récemment été proposé de les remplacer par les modèles de radiation, basés essentiellement sur la densité de points entre une origine et une destination. Notre travail étudie les relations entre ces deux types de modèles : d'abord d'un point de vue théorique et ensuite à travers des expériences numériques, sur des données artificielles et des données de terrain - les flux de navetteurs aux Etats-Unis en l'an 2000 - . Nos analyses confirment un léger avantage pour le modèle de radiation qui fournit, sans paramétrage, une bonne première estimation des flux observés plus précisément dans l'état de New-York.

Abstract :

Thanks to the development and widespread use of geolocation, many devices (*e.g.* social networks) provide data on human trajectories. Over time, these technologies have been used by researchers in order to develop different models for human mobility and to uncover universal properties. Thus, two types of models have been created. Gravity models, widely used for many years to predict the number of travelers in transport theory, emphasise the effect of physical distance on mobility. However, because they are not universal (parameters dependent on region), it has recently been proposed to replace them by radiation models, mainly based on the density of points between an origin and a destination. Our work deals with the relationships between these two kinds of models : first from a theoretical point of view and then through numerical experiments, on artificial and real data - commuting flows in the United States in 2000 - . Our analyses confirm a slight advantage for the radiation model that provides, parameter-free, good first estimates of the flows observed, more specifically in the state of New-York.

Table des matières

Introduction

De tout temps, l'homme a entrepris de se déplacer et de voyager. Outre les flux migratoires motivés par des raisons de survie, les déplacements se sont faits de plus en plus loin pour des raisons économiques, militaires et religieuses. Au 13^{e} siècle, Marco Polo ouvre la voie en s'aventurant à partir de Venise jusqu'en Chine, à travers la Route de la Soie. Au 15^{e} siècle, Christophe Colomb est le premier de l'histoire moderne à traverser l'océan Atlantique et à établir une route entre le continent américain et l'Europe. Magellan, quant à lui, réalise la première circumnavigation, achevée en 1522. Au fil du temps, parallèlement, les moyens de transport n'ont cessé de se développer pour en arriver, en fonction des innovations techniques, à une explosion des réseaux routiers, ferroviaires, aériens et maritimes. Mais depuis quelques décennies, la mobilité moderne se révèle massive et complexe - nous pouvons aller où bon nous semble à travers toute la Terre, en seulement un jour ou deux - . Notre monde actuel se définit ainsi en mouvement : chaque année, plus de trois milliards d'êtres humains entreprennent des voyages en avion ; à plus petite échelle, des centaines de millions de personnes se rendent quotidiennement au travail en empruntant les axes routiers ou les transports publics, utilisés souvent à leur capacité maximale. En ce sens, tous ces mouvements contribuent à modéliser la structure des zones urbaines et la connectivité des modes de transport et des marchés de l'emploi.

Comprendre la mobilité des hommes est une démarche intéressante en soi, mais elle est également d'une importance primordiale car tous ces mouvements de population ont un impact indéniable sur un plan plus large, notamment comme facteur de propagation des maladies infectieuses. Notre travail consistera donc à développer différents modèles mathématiques et à étudier si les déplacements présentent des propriétés universelles, s'il existe des principes gouvernant l'évolution des réseaux ... Enfin, nous essayerons de déterminer quels sont les facteurs à prendre en compte lors de l'élaboration de ces modèles.

A l'heure actuelle où le développement de la géolocalisation se trouve régulièrement à la une de l'actualité, les données sur la mobilité de millions de personnes à travers le monde sont générées directement ou indirectement par les technologies modernes comme le GSM, le GPS, l'outil de réseau social FOURSQUARE combinés avec les sites internet

qui localisent les personnes en un lieu déterminé. Ces nouvelles technologies permettent d'identifier le moment et le lieu où nous nous déplaçons avec une précision spatiotemporelle jamais égalée auparavant.

A partir de ces observations empiriques, de nombreux modèles ont été développés pour la mobilité humaine. L'un des premiers fut construit sur base du jeu "*Where's George*". Dirk Brockmann a modélisé les données obtenues en suivant les trajectoires des billets de dollars à l'effigie du président américain Georges Washington, révélant une loi de puissance énoncée en annexe. La trajectoire de ces billets rappelle les marches aléatoires connues telles les *Lévy flights*. De la même façon, les données obtenues à partir des GSM indiquent également que les distributions du temps passé à un endroit précis et de la taille du saut vers le prochain endroit caractérisant les trajectoires humaines ont des distributions avec longue traine. Ces observations, développées dans l'article de [Song *et al.*] tendent à indiquer que les trajectoires sont mieux décrites par les *Lévy flights* ou par un modèle CTRW (*Continuous-Time Random-Walk*). Cependant, ce type de modèle souffre de son caractère purement aléatoire, contraire aux régularités de la mobilité humaine, tel que le fait d'aller et venir quotidiennement à son travail. Les trajectoires humaines suivent plusieurs lois d'échelle hautement reproductibles que l'on se doit d'incorporer dans un modèle fidèle.

Une autre question qui agite les débats est celle de déterminer quelles variables freinent la mobilité humaine. D'une part, les modèles de gravité sont largement utilisés pour prédire le nombre de voyageurs se déplaçant d'un endroit à un autre, et les mouvements de population. Ces modèles mettent en avant l'effet de la distance physique sur la mobilité. Néanmoins, ils laissent paraître quelques lacunes, notamment en terme de propriétés universelles. D'autre part, les modèles de radiation conçus par Simini se basent essentiellement sur la densité de points entre origine et destination. Ces modèles ont l'avantage de présenter des profils universels : où que l'on soit dans le monde, le même type de courbe est observé. Cependant, ici aussi, quelques inconvénients sont relevés.

Notre but sera de développer ces différents modèles mathématiques et de comparer leurs propriétés à celles observées de manière empirique. Nous essayerons d'établir quel modèle correspond le plus à nos données en dépit, parfois, de quelques inconvénients. Enfin, nous essayerons de trouver un modèle universel, qui puisse s'appliquer à tous les cas.

Notre tâche s'est articulée en trois étapes : une première partie d'étude plus théorique où, après avoir dressé un aperçu global des différents modèles à travers les âges, nous nous intéresserons plus particulièrement au tout récent modèle de radiation. Suivra ensuite une partie plus pratique et numérique, où nous tenterons d'établir des comparaisons avec le modèle de gravité dans différents cas : après avoir supposé que les villes étudiées se situent toutes sur une seule droite, les deux modèles seront expérimentés avec un jeu de données réelles après avoir suggéré une partie en deux dimensions. Nous terminerons, enfin, par une conclusion finale, en tentant de dégager certaines pistes de réflexion pour l'avenir.

Partie I

Partie théorique

CHAPITRE 1

Les différents modèles au fil du temps

Le XX^{e} siècle restera assurément dans les mémoires comme le siècle des révolutions industrielles et technologiques. Parmi celles-ci, nous retiendrons tout particulièrement les moyens de transport, de communication, ... et une mobilité quasi universelle. Avec les derniers outils en matière de télécommunication électronique, la communication s'avère plus que jamais un vecteur de connaissance, pour les scientifiques.

Ainsi, depuis quelques décennies, grâce au développement et à l'utilisation courante de la géolocalisation, de nombreux appareils (GSM, GPS, réseaux sociaux, ...) permettent d'obtenir des données sur les déplacements de personnes. Au fil du temps, toutes ces technologies vont être utilisées par les chercheurs afin d'élaborer différents modèles destinés à prédire la mobilité humaine. Comment en sommes-nous arrivés là ?
Dans les deux sections qui suivent, nous présenterons brièvement certains de ces modèles portant sur la mobilité humaine, les uns basés sur la distance, les autres sur le rang.

1.1 Modèles basés sur la distance

Mis à part une légère amorce au XVIII^{e} siècle avec les travaux de G. Monge, c'est au cours des 80 dernières années que de nombreux processus basés sur la loi universelle de la gravitation découverte par Newton ont été conçus, développés et utilisés pour prédire la mobilité humaine : un des premiers chercheurs à avoir relié la population et le transport est Reilly ([Bamis]). En 1931, alors qu'il étudie le commerce de détail, il constate que plus une agglomération est importante et concentrée, plus son attraction est grande. De fait, les flux entre deux villes sont directement proportionnels au volume de leur population

respective, et inversément proportionnels au carré de leur distance. En 1946, Zipf explique la migration par le principe du moindre effort ([Bamis]) : les flux de voyageurs entre 2 villes sont directement proportionnels au produit de leur population, et inversément proportionnels à leur distance.
Depuis quelques années, l'afflux de données empiriques a motivé le développement de modèles de plus en plus précis.

Afin de mieux comprendre le concept complexe de la mobilité humaine, en 2006, [Brockmann 2] utilise le jeu *"Where's George ?"*, qui consiste à répertorier sur Internet les billets à l'effigie du président américain G. Washington. Le principe se révèle simple : si vous êtes en possession d'un billet d'un dollar, vous l'enregistrez sur le site du jeu en encodant votre code postal (voire le pays, depuis quelque temps) et le numéro de série du billet. Ensuite, celui-ci est remis en circulation et réencodé tour à tour par ses différents utilisateurs. Afin de soutenir cette démarche, une petite phrase d'encouragement figure parfois sur le billet :

> "See where I've been.
> Track where I go next !
> `www.wheresgeorge.com`"

Ainsi, nous pouvons suivre la trajectoire des différents billets de banque à travers les Etats-Unis.

A partir de toutes ces données, D. Brockmann entendait définir les propriétés essentielles de la mobilité aux Etats-Unis : il est arrivé à la conclusion que la probabilité, notée $P(r)$, qu'un billet parcoure une distance r sur une courte période de temps suit une loi de puissance :

$$P(r) \sim r^{-\mu} \qquad \text{où } \mu = 1.6.$$

La trajectoire de ces billets n'est pas sans rappeler les marches aléatoires connues, telles les *"Lévy flights"*.
Il faut cependant garder en mémoire que la trajectoire d'**un seul** billet de banque reflète en fait **plusieurs** parties de trajectoires de personnes diverses (étant donné que le billet circule de portefeuille en portefeuille). De plus, ces données ont seulement été utilisées pour les Etats-Unis. Que se passerait-t-il à l'extérieur ? Pourrait-on déduire les mêmes observations ?

Brockmann et Theis ont suggéré une autre piste afin précisément d'analyser la mobilité humaine hors USA. Ils ont utilisé les données d'un autre jeu, cette fois international (ce jeu est disponible dans 222 pays) : le *"géocaching"* qui consiste en fait en une chasse au trésor. Les joueurs dissimulent une boite appelée "cache" ou "géocache" à différents endroits et publient ensuite les coordonnées du système de positionnement par satellite (GPS) sur le site. Dans chaque balise se trouvent une notice expliquant le principe du *"géocaching"*, un

journal de visite, un stylo, et quelques trésors - souvent des bibelots sans valeur - . Ensuite, les géo-chercheurs essayent de trouver la cachette et peuvent prendre un des trésors, à condition d'en déposer un de même valeur. Comme pour les billets de banque, si vous découvrez par hasard une de ces "caches", une notice explicative vous invite à la laisser à sa place et, pourquoi pas, à participer au jeu.

Pour ces deux chercheurs, les trajectoires de ces différentes balises fournissent une bonne topographie de la mobilité humaine en Europe. Certaines similitudes se dégagent avec les Etats-Unis.

Deux ans après le modèle des billets de banque, Barabási va plus loin et utilise les GSM afin d'analyser les trajectoires des individus grâce à la localisation des antennes-relais connectées aux portables. A chaque coup de fil passé entre 2 personnes, la position de la station de base est encodée. Dès lors, la précision de la position de l'utilisateur dépendra de l'emplacement de cette tour : dans les zones urbaines - où le nombre d'antennes est plus important - , la position sera "faussée" de quelques centaines de mètres ; dans les zones plus rurales - où elles sont plus dispersées - , la marge d'erreur pourra atteindre quelques kilomètres. Barabási tire les mêmes observations : les données suivent une loi de puissance. A l'inverse du modèle précédent toutefois, il s'agit ici de mobilité individuelle : chaque personne possède généralement son propre GSM.

Autre principe : [Song *et al.*] découvrent que les distributions du temps passé à un point précis et de la taille du saut vers le prochain point caractérisant les trajectoires humaines, suivent également des lois de puissance. Ces observations tendent à indiquer que les trajectoires sont mieux décrites par les *"Lévy flights"* ou par un modèle CTRW (*Continuous-Time Random-Walk*). Néanmoins, ce type de modèle souffre de son caractère purement aléatoire, contrairement aux régularités de la mobilité humaine, telles les navettes quotidiennes vers le lieu de travail. Afin de remédier à ce problème, Song *et al.* ont pris en considération plusieurs contraintes :

- Deux mécanismes propres à la mobilité humaine :
  - L'exploration de nouveaux endroits décroît avec le temps : de fait, les gens ont tendance à connaitre une majorité des endroits à proximité de leur domicile ou de leur lieu de travail ;
  - Retour préférentiel : les individus ont plutôt tendance à retourner dans des endroits familiers, à savoir, par exemple, leur domicile ou leur lieu de travail.
- La vie humaine est caractérisée par des périodes : chaque semaine comprend 7 jours, eux-mêmes divisés en 24 heures. De plus, les personnes se déplacent moins fréquemment la nuit (elles sont censées dormir), et beaucoup plus en début et fin de journée (quand elles partent au travail et en reviennent).
- Certaines corrélations dans la mobilité spatiale doivent être prises en compte : si une ville se situe entre son point de départ et sa destination, l'individu passe nécessairement par cette ville.

Ces caractéristiques propres à l'homme ne sont pas prises en compte, par exemple, pour les 2 modèles précédents.

Tous ces modèles dépendent principalement d'une variable - la distance - et sont donc communément appelés modèles de gravité. Néanmoins, un des principaux inconvénients de ce type de modèle réside en sa non-universalité. De fait, le point de départ va avoir un impact sur le déplacement. Les paramètres définissent des modèles différents en fonction de l'endroit où l'on se trouve sur la planète.

1.2 Modèles basés sur le rang

Au vu du principal inconvénient des modèles de gravité, les chercheurs refusent l'idée que la mobilité puisse être définie à partir de la distance physique. Ils ont donc tenté de trouver une nouvelle variable, laissant dès lors libre cours à de nouveaux modèles, comme le modèle de [Noulas *et al.*] : la probabilité de se déplacer d'une place à une autre est inversément proportionnelle à une puissance de leur rang, à savoir le nombre d'opportunités (ou d'endroits) entre ces 2 villes. A partir de là, ces chercheurs ont pu relever une universalité surprenante, malgré les différences culturelles, organisationnelles et nationales, grâce aux données collectées via l'application FOURSQUARE. Ce réseau permet en fait à l'utilisateur d'indiquer exactement où il se trouve via des *"check-ins"*.

Ce type de modèle se révèle similaire au modèle IO (*Intervening Opportunities*)[1] développé pour la première fois en 1940 par [Stouffer] : les individus ne voyagent plus au hasard, ils ont un objectif qu'ils veulent satisfaire. Malheureusement, les paramètres définissant ce modèle sont de nouveau non universels et il est difficile de les calibrer correctement. Dès lors, aucune forme fonctionnelle dépendant du rang n'est proposée.

Poursuivant sur cette lancée, [Simini *et al.*] se sont attachés à déterminer d'emblée la forme fonctionnelle et ont proposé en 2011 un nouveau modèle, qui est en fait un cas particulier du modèle IO. Celui-ci sera largement développé dans les chapitres suivants : les individus choisissent de préférence la meilleure opportunité (de travail par exemple), la plus proche de chez eux.

1. Le principe est relativement simple : le fait de se déplacer ne dépend pas de la distance mais de l'accessibilité des opportunités à pouvoir satisfaire à l'objectif du trajet. Le nombre de personnes effectuant une certaine distance est directement proportionnel au nombre d'opportunités à la destination et inversément proportionnel au nombre de possibilités intermédiaires.

CHAPITRE 2

Le modèle de radiation

Au vu de la complexité de la mobilité humaine, de la migration ou de la communication entre différentes régions, il pourrait à première vue s'avérer impossible de capturer les bases de données nécessaires à l'objet de notre étude et donc d'en tirer des modèles, comme souligné dans [Brockmann 3]. Néanmoins, le développement de la géolocalisation a conduit à l'élaboration de différents modèles, jusqu'à celui relativement simple défini par Simini *et al.*, concordant très bien avec les données empiriques.
Le modèle de radiation ainsi conçu s'appuie essentiellement sur la densité de points entre l'origine et la destination. Cet outil, développé pour la mobilité humaine et la migration, a l'avantage de présenter des profils universels : où que l'on soit dans le monde, le même type de courbe est observé. L'hypothèse de base est que les personnes recherchent l'opportunité maximale la plus proche. Cependant, quelques inconvénients sont mis en évidence.

Au cours de ce chapitre, nous décrirons le principe du modèle de radiation, et le développerons mathématiquement afin d'en tirer des propriétés. Nous tenterons également d'établir la relation entre ce modèle et la loi de gravité, pour conclure avec une amélioration du modèle.
Ce chapitre se base essentiellement sur l'article de Simini *et al.* ([Simini *et al.*]).

2.1 Mise en contexte

La loi de gravité est largement utilisée pour prédire les taux de voyageurs, les mouvements de population ou la diffusion d'épidémies. Ainsi, le nombre d'individus, T_{ij}, se déplaçant entre les endroits i et j est proportionnel à une puissance de la population d'ori-

gine (n_i) et à une puissance de celle de destination (n_j), et inversément proportionnel à une fonction f dépendant de la distance entre ces deux points, d_{ij}. Dès lors,

$$T_{ij} = \frac{n_i^\alpha n_j^\beta}{f(d_{ij})}$$

où α et β sont des paramètres ajustables, et la fonction de dissuasion $f(d_{ij})$ est choisie pour correspondre au mieux aux données.

Malheureusement, ce modèle présente quelques lacunes. Ainsi, par exemple, les paramètres décrivant la loi de gravité varient d'une région à l'autre, tout comme la fonction de dissuasion $f(d_{ij})$ diffère suivant l'endroit où nous nous trouvons sur la planète : ainsi par exemple, $f(d_{ij}) = d_{ij}^\gamma$ est la fonction consacrée pour prédire le modèle de la navette entre les comtés américains, alors que les modèles de navette mondiale entre 29 pays suggèrent plutôt une fonction de dissuasion de type exponentielle, $f(d_{ij}) = e^{d_{ij}c}$. L'universalité tant convoitée est donc remise en question. De plus, pour prédire les valeurs de ces paramètres, il nous faut tenir compte des données antérieures sur la mobilité, ce dont nous ne disposons pas toujours. Enfin, avec ce modèle, le nombre de voyageurs augmente sans limite si la population de destination n_j s'accroît. Ceci semble peu réaliste si nous considérons que le nombre de voyageurs reliant un endroit à un autre ne peut dépasser le nombre maximal de personnes partant de i, à savoir la population d'origine n_i.

Différents modèles ont été élaborés afin d'essayer de remédier à ces failles : les modèles IO (*Intervening Oppurtunities Model*) et RU (*Random Utility Model*) [1]. Néanmoins, il s'avère assez vite qu'ils présentent les mêmes difficultés : pour le modèle IO, les paramètres de base varient et la complexité numérique est importante ; quant au modèle RU, les paramètres sont spécifiques au contexte. Outre le coût élevé de la paramétrisation, ces modèles n'offrent pas une meilleure correspondance aux données empiriques. Dès lors, le modèle de gravité est resté l'outil principalement utilisé dans la modélisation de la mobilité.

En 2011, Filippo Simini, Marta C. González, Amos Maritan et Albert-László Barabási mettent sur pied un modèle plus fiable, qui comble les lacunes du modèle de gravité, tout en assurant une meilleure correspondance aux données empiriques : le modèle de radiation.

2.2 Principe du modèle de radiation

Considérons une origine i émettant un flux de particules indépendantes et identiques. Définissons tout d'abord les processus d'absorption et d'émission en deux phases. A des

1. A chaque destination est assignée une utilité - suivant une fonction logit - . L'individu choisira l'alternative lui proposant l'utilité maximale.

fins explicatives pratiques, nous établirons un parallèle avec le choix d'une activité professionnelle. En quoi ce phénomène constitue-t-il un flux de particules indépendantes et identiques ? Il suffit d'observer, matin et soir, les mouvements des personnes actives résidant et travaillant au sein même des grandes agglomérations urbaines, mais surtout de tous les navetteurs en provenance de la périphérie... avec l'image-type des kilomètres d'embouteillage à l'entrée ou à la sortie de Bruxelles, ou des flux humains dans les grandes gares ou aux stations de métro aux heures de pointe... Même si la navette entre le domicile et le lieu de travail est un processus quotidien, son origine et sa destination sont déterminées par le choix d'un travail pris à long terme. Ce choix consiste également en deux étapes : les personnes se rendent à un nouveau lieu professionnel seulement s'il leur offre une plus grande proximité par rapport à leur domicile, et un meilleur emploi.

Attachons-nous à présent à développer les deux phases.

1. A chaque particule X émise de i, nous associons le nombre $z_X^{(i)}$, seuil d'absorption pour la particule X. Dans le cadre du choix d'un travail, ce nombre représente les bénéfices d'une opportunité d'emploi tels que la rétribution, le nombre d'heures de travail. Nous considérons que la population en i est donnée par le nombre n_i. Nous extrayons n_i nombres aléatoires d'une distribution $p(z)$ préselectionnée du seuil d'absorption. Cette distribution est inconnue. Néanmoins, elle n'interviendra plus par la suite. Nous pourrons ainsi choisir une distribution particulière sans fausser nos résultats.

   Illustrons ce processus via un graphique : supposons que $n_i = 2$, alors nous extrayons deux nombres aléatoires T_1 et T_2 d'une distribution $p(z)$ choisie au hasard.

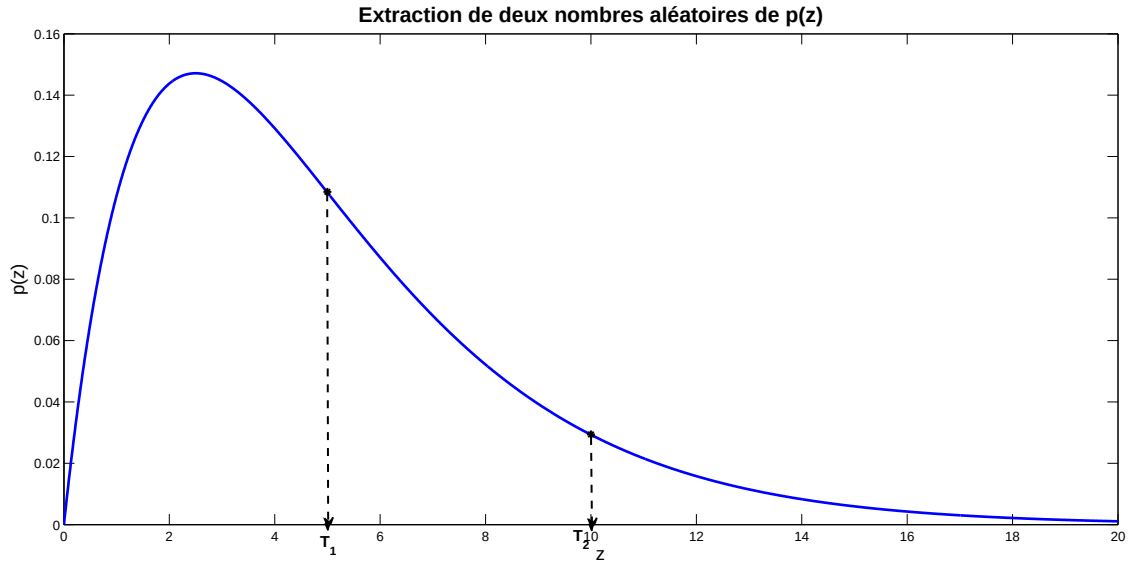


FIGURE 2.1 – Extraction de deux nombres aléatoires d'une distribution

Le seuil d'absorption $z_X^{(i)}$ est défini comme étant le maximum de ces nombres aléatoires : dans notre exemple,

$$z_X^{(i)} = \max_{i=1,2} T_i = T_2.$$

Une particule présentant un grand seuil d'absorption aura donc moins de chance d'être absorbée par un autre endroit. De fait, lorsque nous revenons à notre application pratique, si, au point de départ, nous avons déjà de sérieuses garanties de bénéfices dans une offre d'emploi, pourquoi aller chercher ailleurs ?

De plus, si nous augmentons n_i, il est fort probable que notre seuil $z_X^{(i)}$ s'élève.
Reprenons notre graphique 2.1 en considérant cette fois que la population de départ n_i est passée de 2 à 3. Nous extrayons alors trois nombres aléatoires T_1, T_2 et T_3 de cette distribution $p(z)$.

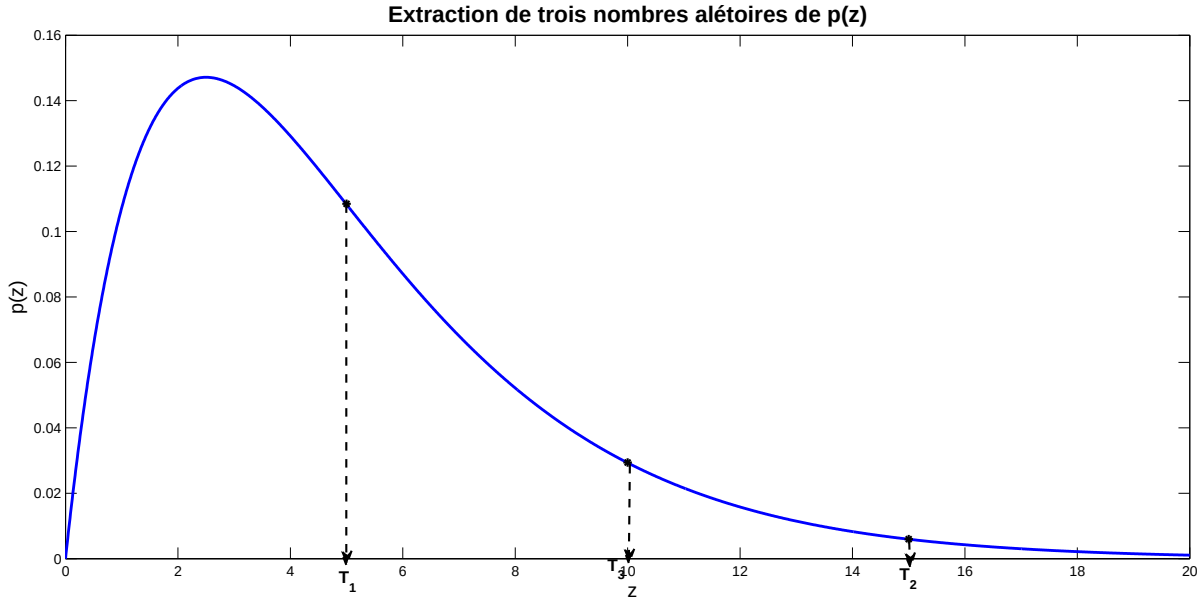


FIGURE 2.2 – Extraction de trois nombres aléatoires d'une distribution

De fait, notre seuil d'absorption $z_X^{(i)}$ a augmenté.
En moyenne, les particules émises d'un endroit à forte densité de population auront un seuil d'absorption plus grand que dans le cas inverse. Dans le cadre de notre exemple pratique, plus notre point d'origine (une ville, ici, en l'occurrence) est peuplé, plus il y aura d'offres d'emploi sur le marché. Dans ce cas, nous avons plus de chance de trouver un emploi nous proposant des bénéfices importants.

2. Nous allons procéder de la même manière pour les villes avoisinantes : chaque endroit j avec une population n_j a une probabilité d'absorber la particule X. Nous définissons donc l'absorbance de l'endroit j pour la particule X, $z_X^{(j)}$, comme étant le maximum des n_j extractions de la distribution $p(z)$.

Le principe du modèle de radiation réside dans le fait que la particule est absorbée par l'endroit le plus proche avec l'absorbance supérieure au seuil d'absorption. En terme de sélection d'un travail, l'individu choisit l'emploi le plus proche de chez lui avec le plus grand bénéfice z. Il est en effet admis de manière incontestable que les travailleurs ne recherchent pas forcément la meilleure opportunité, mais plutôt la destination la plus proche avec un seuil d'absorbance plus grand. En répétant ce processus pour toutes les particules émises, nous obtenons les différents flux à travers le pays.

Nous allons à présent développer le modèle de radiation et établir diverses équations. Supposons que la personne X se trouvant en i choisisse l'endroit j au vu des plus grands bénéfices engendrés.

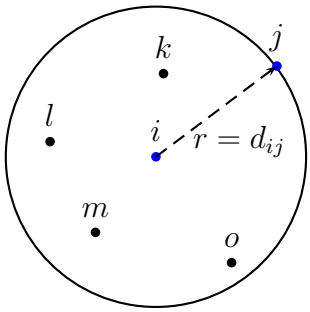


FIGURE 2.3 – Définition du rang

Elle a dès lors refusé toutes les propositions représentées par les points k, l, m, o, dans le cercle de rayon $r = d_{ij}$ centré en l'origine i, étant donné que les bénéfices y afférents sont inférieurs à celui proposé au point de départ i, $z_X^{(i)}$.
Nommons s_{ij} la population totale établie dans ce cercle, à l'exception des populations en i, n_i, et en j, n_j. Ce paramètre sera également appelé rang. Nous aurons dès lors, dans le cas présent :

$$s_{ij} = n_k + n_l + n_m + n_o.$$

La probabilité qu'une particule en i avec une population n_i soit absorbée dans l'endroit j avec une population n_j, car l'absorbance en ce point est supérieure à toutes les autres et au seuil d'absorption z, peut donc s'écrire comme :

$$P(1 \mid n_i, n_j, s_{ij}) = \int_0^\infty dz\, P_{n_i}(z) P_{s_{ij}}(< z) P_{n_j}(> z) \tag{2.1}$$

où $P_{n_i}(z)$ est la densité de probabilité que la valeur maximale extraite de la distribution $p(z)$ après n_i essais vaille z.
Revoyons les notions de fonction de répartition et de fonction de densité.

Définition 1.

La **fonction de répartition** F d'une variable aléatoire Z sur $\mathbb{R}$ est la fonction suivante :

$$F_Z(z) = P(Z < z)$$

où le terme de droite représente la probabilité que la variable aléatoire Z prenne une valeur inférieure à un nombre z.

Définition 2.

La fonction dérivée f de la fonction de répartition F est dite **fonction de densité de probabilité** de Z et vérifie les relations :

$$\forall z \in \mathbb{R} : \quad f(z) = F'(z) \ \text{ et } \ F(z) = P(Z < z) = \int_{-\infty}^{z} f(t)dt.$$

Définissons la probabilité que les n_i nombres extraits de $p(z)$ soient tous inférieurs à z, $P_{n_i}(< z) = p(< z)^{n_i}$. Il s'agit ici de la fonction de répartition.

Nous déduisons ainsi la fonction de densité :

$$\begin{aligned} P_{n_i}(z) &= \frac{dP_{n_i}(< z)}{dz} \\ &= n_i \, p(< z)^{n_i - 1} \, \frac{dp(< z)}{dz}. \end{aligned}$$

De même, nous pouvons écrire : $P_{s_{ij}}(< z) = p(< z)^{s_{ij}}$.

Définissons enfin la probabilité que, parmi n_j nombres extraits de $p(z)$, un au moins soit supérieur à z :

$$P_{n_j}(> z) = 1 - P_{n_j}(< z) = 1 - p(< z)^{n_j}.$$

A partir de ces différentes formules, nous allons pouvoir évaluer l'intégrale (2.1) :

$$\begin{aligned} P(1 \mid n_i, n_j, s_{ij}) &= \int_0^\infty dz\, P_{n_i}(z) P_{s_{ij}}(< z) P_{n_j}(> z) \\ &= n_i \int_0^\infty dz\, \frac{dp(< z)}{dz}\, \left[p(< z)^{n_i+s_{ij}-1} - p(< z)^{n_i+s_{ij}+n_j-1}\right] \\ &= n_i \left[\frac{1}{n_i + s_{ij}} - \frac{1}{n_i + n_j + s_{ij}}\right]. \end{aligned}$$

Nous voyons ainsi que la probabilité pour une particule de se déplacer de i à j peut se réécrire comme :

$$P(1 \mid n_i, n_j, s_{ij}) = p_{ij} = \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})}. \tag{2.2}$$

D'emblée, nous remarquons que cette équation est indépendante de la distribution $p(z)$ et ne dépend d'aucun paramètre. De plus, elle reste inchangée si nous multiplions les différentes populations par le même facteur, ce qui établit l'invariance d'échelle.

Démontrons à présent une proposition importante en statistiques afin de nous assurer de l'obtention d'une probabilité.

Propriété 1.

La distribution p_{ij} est normalisée, c'est-à-dire

$$\sum_{j \neq i} p_{ij} = 1.$$

Démonstration. Considérons chaque cercle centré en i de rayon d_{ij} avec $j \geq 1$; chaque cercle est ordonné de manière croissante. Représentons graphiquement la situation où la personne sise en i se trouve, par exemple, entourée de 4 villes voisines situées respectivement en $j = 1$, $j = 2$, $j = 3$ et $j = 4$.

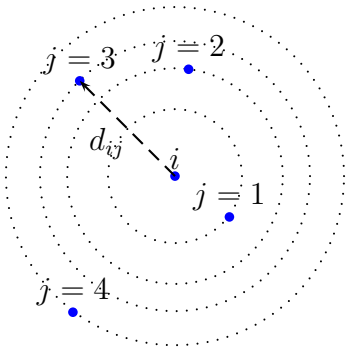


FIGURE 2.4 – Définition du rang

Si nous cherchons à déterminer le j^{e} cercle le plus proche de i, nous voyons que la ville localisée en $j = 1$ se trouve la plus proche de l'origine. A l'inverse, la ville située en $j = 4$ apparait la plus éloignée.

Regardons ce que devient le rang s_{ij} en faisant varier le paramètre j :

$$\begin{array}{lll} s_{i1} & = 0 & \text{(pas de ville entre } i \text{ et } j) \\ s_{i2} & = n_1 & \text{(la seule ville entre } i \text{ et } j = 2 \text{ est } j = 1) \\ s_{i3} & = n_1 + n_2 & \text{(2 villes se situent entre } i \text{ et } j = 3 : j = 1 \text{ et } j = 2) \\ \vdots & & \end{array}$$

Définissons ensuite s'_{ij} comme la population totale établie dans le cercle centré en i et de rayon d_{ij} (excepté la population en j, n_j).
Nous obtenons donc :

$$\begin{array}{lll} s'_{i1} & = n_i & \text{(la seule ville entre } i \text{ et } j \text{ est } i) \\ s'_{i2} & = n_i + n_1 & \text{(2 villes se situent entre } i \text{ et } j = 2 : i \text{ et } j = 1) \\ s'_{i3} & = n_i + n_1 + n_2 & \text{(3 villes se situent entre } i \text{ et } j = 3 : i,\ j = 1 \text{ et } j = 2) \\ \vdots & & \end{array}$$

Nous pouvons ainsi écrire : $s'_{ij} = n_i + s_{ij}$ et $s'_{ij+1} = s'_{ij} + n_j$.

Evaluons

$$\sum_{j \neq i} p_{ij} = n_i \sum_{j=1}^{\infty} \frac{1}{n_i + s_{ij}} - \frac{1}{n_i + n_j + s_{ij}}.$$

Or,

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{1}{n_i + s_{ij}} - \frac{1}{n_i + n_j + s_{ij}} &= \sum_{j=1}^{\infty} \left[\frac{1}{s'_{ij}} - \frac{1}{n_j + s'_{ij}} \right] \\
&= \sum_{j=1}^{\infty} \frac{1}{s'_{ij}} - \sum_{j=1}^{\infty} \frac{1}{s'_{ij+1}} \\
&= \sum_{j=1}^{\infty} \frac{1}{s'_{ij}} - \sum_{j=2}^{\infty} \frac{1}{s'_{ij}} \\
&= \frac{1}{s'_{i1}} \\
&= \frac{1}{n_i}.
\end{aligned}$$

Nous en tirons :

$$\sum_{j \neq i} p_{ij} = n_i \frac{1}{n_i} = 1.$$

□

Etant donné leur indépendance, le nombre moyen attendu de particules émises de i vers j équivaut à :

$$T_{ij} = T_i \, p_{ij} = T_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} \tag{2.3}$$

où $T_i = \sum_{j \neq i} T_{ij}$ est le nombre total de particules émises de i, soit, dans notre cas pratique, le nombre de navetteurs partant de i.
Cette formule constitue l'équation fondamentale du modèle de radiation.

Le modèle de radiation présente également des profils universels : où que l'on soit dans le monde, le même type de courbes empiriques est observé. Les prédictions du modèle de radiation sont en accord avec celles-ci, validant ainsi le modèle. Contrairement au modèle de gravité où la distance domine - signifiant que plus le travail est éloigné, moins il offre d'attrait - , nous observons certaines modifications dans le modèle de radiation : la distance géographique séparant les grandes villes s'estompe au profit d'un rapprochement, vu qu'elles présentent un marché de l'emploi beaucoup plus vaste. De plus, le modèle de radiation ne nécessite aucun paramètre, contrairement au modèle de gravité et à ses trois paramètres, α, β et γ. Même à défaut de données, il est en effet possible de prédire les déplacements et schémas de transport, étant donné que la densité de population est connue

à travers le monde. Enfin, grâce à différents jeux de données sur la mobilité, le transport et la migration, les auteurs Simini *et al.* ont pu constater que les prédictions théoriques de ce modèle surpassent largement le modèle de gravité, vu leur meilleure correspondance aux données. Pour démontrer la généralité du modèle, les auteurs ont utilisé quatre jeux de données différents présentant une description quantitative exacte de la mobilité et du transport sur une large échelle temporelle (allant de la mobilité horaire et des navettes quotidiennes aux migrations annuelles) et sur base de divers processus (navettes, commerce, mobilité et appels téléphoniques) étudiés à travers différents continents.
Autre avantage révélé : ce modèle nous permet également de découvrir une invariance d'échelle dans les modèles de voyage via la proposition suivante.

Propriété 2.

La probabilité d'un déplacement à partir d'un point d'origine i avec une population n_i jusqu'à une destination j se situant au-delà du cercle de rayon $d_{ij}(s)$ centré à l'origine peut se réécrire comme :

$$p_{s_{ij}}(\geq s_{ij} \mid n_i) = \frac{1}{1 + \frac{s_{ij}}{n_i}}. \tag{2.4}$$

Démonstration. Nous allons réutiliser les définitions utilisées dans la proposition 1, à savoir

$$s'_{ij} = n_i + s_{ij} \text{ et } s_{ij}.$$

Calculons donc $p_{s_{ij}}(\geq s_{ij} \mid n_i)$. Comme nous travaillons au-delà du cercle de rayon d_{ij}, avec j comme délimitation de notre cercle, nous considérons que l'indice de sommation k débute à j sans aucune limite :

$$\begin{aligned}
p_{s_{ij}}(\geq s_{ij} \mid n_i) &= n_i \sum_{k \geq j} \left[\frac{1}{n_i + s_{ik}} - \frac{1}{n_i + n_k + s_{ik}} \right] \\
&= n_i \sum_{k \geq j} \left[\frac{1}{s'_{ik}} - \frac{1}{s'_{ik+1}} \right] \\
&= n_i \sum_{k \geq j} \frac{1}{s'_{ik}} - n_i \sum_{k \geq j+1} \frac{1}{s'_{ik}} \\
&= n_i \frac{1}{s'_{ij}} \\
&= n_i \frac{1}{n_i + s_{ij}}.
\end{aligned}$$

Nous revenons ainsi à la formule (2.4). □

Nous pouvons constater que cette formule est invariante sous la transformation $n_i \to \lambda n_i$ ou $s_{ij} \to \lambda s_{ij}$, ce qui n'est pas vérifié par le modèle de gravité. Cette équation aura son importance car elle permettra d'améliorer la correspondance du modèle aux données empiriques grâce à l'ajout d'une nouvelle variable. Ce sujet sera traité dans la section *Amélioration du modèle*.

2.3 Cas de la distribution uniforme

Comparons le modèle de gravité et celui de radiation. La différence clé réside en l'absence de la variable d_{ij} (distance entre i et j) et l'introduction d'une nouvelle variable s_{ij} (population totale établie dans le cercle de rayon d_{ij} centré à l'origine i). Dans le cadre d'une densité de population uniforme, nous avons le pouvoir de transformer le modèle de radiation en un modèle de gravité via un changement de variable.

Si nous considérons une population uniforme, la population d'origine équivaut à la population de destination, c'est-à-dire $n_i = n_j$, et nous pouvons réécrire s_{ij} comme

$$s_{ij}(d_{ij}) = n_i \pi d_{ij}^2.$$

Le nombre de voyageurs s'estime donc :

$$\begin{aligned} T(n_i, n_j, s_{ij}) &= n_i \, P(1 \mid n_i, n_j, s_{ij}) \\ &= \frac{n_i^2 \, n_j}{(n_i + s_{ij}) \, (n_i + n_j + s_{ij})} \\ &= \frac{n_i^3}{(n_i + s_{ij}) \, (2n_i + s_{ij})} \\ &= \frac{n_i}{(1 + \pi d_{ij}^2) \, (2 + \pi d_{ij}^2)} \\ &\approx \frac{n_i}{d_{ij}^4} \end{aligned}$$

Nous obtenons ainsi une loi de gravité avec $f(d_{ij}) = d_{ij}^{\gamma}$ où $\gamma = 4$ et $\alpha + \beta = 1$.
Néanmoins, l'hypothèse de distribution de population uniforme est irréaliste, par essence.

2.4 Limites asymptotiques

Attachons-nous à présent aux limites asymptotiques des différents modèles. Considérons d'abord le modèle de gravité et le cas où la population d'origine n_i augmente très fort par rapport à la population de destination n_j et à la variable d_{ij} :

$$\lim_{n_i \to \infty} T_{ij}^G = \lim_{n_i \to \infty} \frac{n_i^\alpha \, n_j^\beta}{f(d_{ij})} = \infty.$$

Le modèle de gravité prédit que le nombre de voyages T_{ij} diverge à l'infini, ce qui n'est pas très réaliste. Dans le cadre d'une recherche d'emploi par exemple, le point de destination ne peut offrir du travail à un nombre illimité de personnes.

Passons ensuite au modèle de radiation. Supposons que la population d'origine n_i se développe très fort, par rapport à la population de destination n_j et à la variable s_{ij} :

$$\lim_{n_i \to \infty} T_{ij}^R = \lim_{n_i \to \infty} \frac{n_i^2 \, n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} = n_j + O\left(\frac{1}{n_i^2}\right).$$

Effectuons le même raisonnement en considérant, cette fois, la population de destination n_j qui s'accroît :

$$\lim_{n_j \to \infty} T_{ij}^R = \lim_{n_j \to \infty} \frac{n_i^2 n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} = \frac{n_i^2}{n_i + s_{ij}} + O\left(\frac{1}{n_j}\right) \leq n_i.$$

Il en ressort que le modèle de radiation offre une approximation plus réaliste du modèle observé, étant donné que le nombre de voyages sature à la population d'origine n_i ou à la population de destination n_j.

2.5 Relation entre le modèle de gravité et le modèle de radiation

Recherchons la relation existant entre le modèle de gravité et celui de radiation. Suite aux observations empiriques réalisées par Simini *et al.*, nous pouvons conclure que $T_{ij}^{data} \simeq T_{ij}^{R}$ étant donné que le modèle de radiation ressemble fortement aux données observées. Nous allons tenter de déterminer dans quelles conditions l'erreur entre les deux modèles est la moindre possible.

Définition 3.

La fonction d'**erreur** se définit comme étant la déviation moyenne au carré entre le modèle de gravité et celui de radiation :

$$E = \frac{1}{Nbr_p} \sum_{\{i,j:i\neq j\}} \left[\ln \frac{C\ n_i^{\alpha}\ n_j^{\beta}}{d_{ij}^{\gamma}} \ - \ \ln \frac{n_i^2\ n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} \right]^2 \tag{2.5}$$

où Nbr_p représente le nombre total de paires considérées.

Déterminons ensuite la population locale moyenne.

Définition 4.

La **population locale moyenne** se définit comme le ratio entre la population totale et le nombre d'endroits considérés, N_{ij}, dans le cercle de rayon d_{ij} centré en i :

$$\bar{m} = \frac{n_i + n_j + s_{ij}}{N_{ij}}. \tag{2.6}$$

Si nous écrivons les variables n_i et n_j en fonction des déviations de la population locale moyenne $\bar{m}$, nous obtenons :

$$\begin{array}{rcccl} n_i & = & n_i + \bar{m} - \bar{m} & = & \bar{m}(1+\delta_i) \\ n_j & = & n_j + \bar{m} - \bar{m} & = & \bar{m}(1+\delta_j) \end{array}$$

où $\delta_i = \frac{n_i - \bar{m}}{\bar{m}}$ et $\delta_j = \frac{n_j - \bar{m}}{\bar{m}}$ sont les déviations des populations par rapport à la moyenne locale. Grâce à ces différentes formules (*cf.* le calcul en annexe), nous pouvons donc réécrire l'erreur (2.5) comme :

$$E = \frac{1}{N} \sum_{\{i,j:i\neq j\}} \left[\ln \frac{\bar{m}^{\alpha+\beta}\ (1+\delta_i)^{\alpha}\ (1+\delta_j)^{\beta}}{\rho^2\ d_{ij}^{\gamma}} \ - \ \ln \frac{\bar{m}\ (1+\delta_i)^2\ (1+\delta_j)}{N_{ij}\ (N_{ij} - 1 - \delta_j)} \right]^2$$

où $\rho^2 = \frac{1}{C}$.

L'erreur E atteint son minimum lorsque chaque élément de la somme est nul ou proche de 0, c'est-à-dire :

$$(\alpha+\beta-1)\ln\bar{m}+(\alpha-2)\ln(1+\delta_i)-ln(\rho^2 d_{ij}^{\gamma})+(\beta-1)\ln(1+\delta_j)+\ln(N_{ij})+\ln(N_{ij}-1-\delta_j)=0$$

pour tout $\{i,j : i \neq j\}$.

Nous pouvons récrire cette équation sous le système suivant pour tout $\{i,j : i \neq j\}$:

$$\Leftrightarrow \begin{cases} (\alpha+\beta-1)\ln(\bar{m}) = 0 \\ (\alpha-2)\ln(1+\delta_i)+(\beta-1)\ln(1+\delta_j) = 0 \\ \ln\dfrac{N_{ij}}{\rho d_{ij}^{\gamma/2}} + \ln\dfrac{(N_{ij}-1-\delta_j)}{\rho d_{ij}^{\gamma/2}} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \alpha+\beta = 1 & (2.7) \\ (\alpha-2)\delta_i+(\beta-1)\delta_j = 0 & (2.8) \\ \rho d_{ij}^{\frac{\gamma}{2}} \approx N_{ij} & (2.9) \end{cases}$$

Considérons avec attention les équations (2.7), (2.8) et (2.9), et tentons d'en dégager un sens.
La première équation nous permet de retrouver les conditions obtenues si nous travaillons dans le cadre d'une population uniforme.
L'équation (2.8) est vérifiée si $\alpha = 2$ et $\beta = 1$. Malheureusement, nous sommes alors en contradiction avec l'équation (2.7). Si les déviations des populations à partir de la moyenne locale sont petites (autrement dit δ_i, $\delta_j << 1$), la contribution à E de (2.8) est négligeable, contrairement à $\ln(\bar{m})$, le terme le plus grand si $\bar{m} > 1$.
Enfin, la troisième équation est vérifiée pour $\gamma = 4$ si les endroits sont régulièrement espacés ou d'aires égales. Ainsi, le paramètre ρ peut être défini comme la densité globale des endroits :

$$\rho = \frac{\text{nombre total d'endroits}}{\text{aire du pays}}.$$

Cependant, lorsque ce n'est pas le cas, la densité des deux points i et j sera différente, ce qui entraîne une impossibilité à trouver une densité constante à travers le pays. Dès lors, la correspondance du modèle aux données empiriques est compromise.

En conclusion, quand les endroits présentent une aire plus ou moins similaire et quand les déviations locales sont petites, ces différentes équations nous permettent d'établir que $\alpha+\beta = 1$ et $\gamma = 4$. Au mieux ces équations seront satisfaites, au mieux la courbe de la

loi de gravité conviendra à nos données empiriques.

En guise d'illustration, considérons les Etats-Unis en fonction de ces deux caractéristiques : population égale et aire identique.

Endroits avec population égale

Nous supposons ici que les districts ont une population quasi identique. Cette subdivision ne satisfait pas (2.9), qui requiert des endroits régulièrement espacés. En effet, la densité est plus grande dans les régions hautement peuplées où la zone moyenne des districts est aussi très restreinte, contrairement aux districts des régions nettement moins habitées. Dans ce cas, $\alpha + \beta = 1$ et $\gamma = 4$ ne sont pas satisfaites.

Endroits avec aire identique

Pour ce faire, construisons une grille composée de carrés identiques et regroupons tous les comtés dont les centroïdes géométriques se trouvent dans le même carré. Nous obtenons une répartition similaire à celle trouvée sur [Massilia].

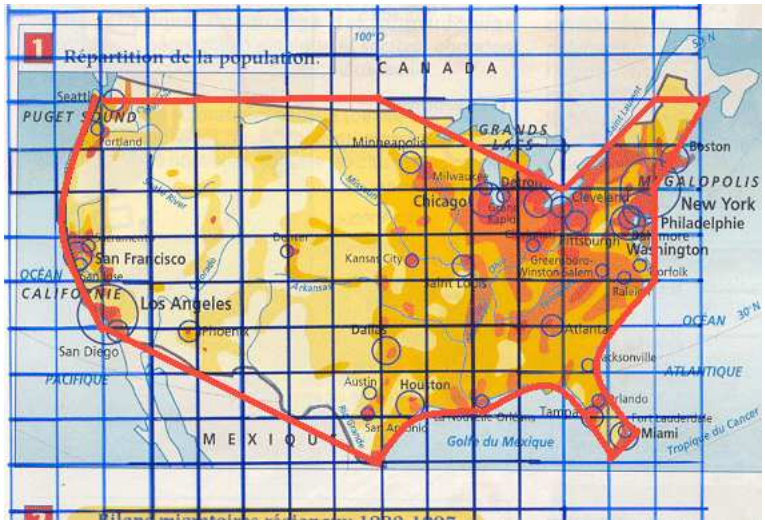


FIGURE 2.5 – Grille associée aux Etats-Unis

Cette subdivision permet d'obtenir des endroits d'aire similaire, condition nécessaire pour que l'équation (2.9) soit vérifiée. Nous obtenons ainsi une marge d'erreur plus petite que dans le cas précédent. En conclusion, la loi de gravité s'applique mieux quand les endroits sont d'aires égales et équidistants ; et aussi longtemps qu'il y a une uniformité locale dans la distribution de population.

2.6 Amélioration du modèle

Dans le modèle de radiation décrit précédemment, nous allons désormais introduire un terme qui prendra en compte l'avantage supplémentaire de se trouver en l'origine - comme se trouver dans sa ville d'origine pour rechercher un emploi par exemple - .

Nous considérons en effet que les individus connaissent mieux l'endroit dont ils sont originaires : ils y ont établi un cercle plus large de connaissances, un réseau social et ainsi, ont accès à plus d'opportunités d'emploi que dans un endroit non familier - où les seules offres d'emploi accessibles sont constituées des annonces parues dans les journaux locaux, *etc* ...
Il est très facile d'introduire cet avantage dans le modèle de radiation et de l'implémenter. Alors que pour les autres endroits, le nombre d'offres d'emploi est proportionnel à chaque population respective, le nombre d'opportunités d'emploi, pour l'endroit d'origine, est proportionnel à la population $n_i + \varepsilon$, correspondant à l'addition efficace de ε personnes. Il y a donc une forte chance que le seuil d'absorption $z_X^{(i)}$ augmente.
Notre équation (2.2) devient donc

$$P(1 \mid n_i, n_j, s_{ij}, \varepsilon) = \frac{(n_i + \varepsilon)\, n_j}{((n_i + \varepsilon) + s_{ij})\, ((n_i + \varepsilon) + n_j + s_{ij})}.$$

Nous pouvons dès lors réécrire notre probabilité (2.4) de nous rendre vers une destination au-delà du cercle de rayon $d_{ij}(s_{ij})$ centré à l'origine comme :

$$p_{s_{ij}}(\geq s_{ij} \mid n_i, \varepsilon) = \frac{n_i + \varepsilon}{(n_i + \varepsilon) + s_{ij}}.$$

En conclusion, le modèle de radiation permet d'affiner l'exactitude des prédictions en la matière, désormais nettement plus concordantes avec les données empiriques.
Ce modèle ouvre une nouvelle voie d'exploration, plus fiable et performante, dans la compréhension des phénomènes de mobilité en terme de mouvements de population à plus ou moins longue distance, de géographie urbaine, d'épidémiologie, de flux économiques ...

Partie II

Partie numérique

Les différents programmes, créés avec le logiciel MATLAB et répertoriés en annexe, ont pour objectif de comparer les deux modèles étudiés dans le présent mémoire :

- le modèle de gravité - largement utilisé au fil des ans - ;
- le modèle de radiation - élaboré en 2011 - .

Cette partie s'articule en 3 chapitres :

- La comparaison des modèles à une dimension ;
- Le passage en deux dimensions ;
- La comparaison des modèles avec un jeu de données réelles.

CHAPITRE 3

Comparaison des modèles à une dimension

Dans cette première application directe, nous avons décidé de travailler à une seule dimension, à savoir que nos villes se situent toutes sur une ligne droite. Néanmoins, comme le montrent les différents schémas développés ci-après, il existe deux chemins possibles :

- soit nous empruntons le chemin traditionnel direct, en ligne droite ;
- soit nous empruntons un autre chemin, en respectant dès lors une condition périodique (en formant une boucle).

Dans un premier temps, nous supposerons que toutes nos villes sont peuplées d'un seul habitant et étudierons 2 cas d'application : nous nous intéresserons tout d'abord à une population homogène uniformément située à égale distance, ensuite à une population plus hétérogène, composée de 2 groupes distincts séparés par une distance quelconque - . Dans un second temps, nous admettrons comme hypothèse que toutes nos villes comptent un nombre fini d'habitants, par exemple 10 habitants.

Etablissons au passage un index des quelques notations utilisées dans ce rapport :

N	:	nombre de villes considérées ;
D	:	distance entre les 2 extrémités ;
x_i	:	une ville i quelconque ;
n_i	:	population de la ville i ;
d_{ij}	:	distance entre les villes d'origine, i, et de destination, j ;
p_i	:	position de la ville i ;
r_{ij}	:	différence d'ordre entre les villes d'origine, i, et de destination, j ;
o_i	:	ordre de la ville i ;
s_{ij}	:	population établie entre les villes i et j - également appelée rang - ;
T_{ij}^{G}	:	flux attendu de voyageurs se rendant de l'endroit i à l'endroit j pour le modèle de gravité ;
T_{ij}^{R}	:	flux attendu de voyageurs se rendant de l'endroit i à l'endroit j pour le modèle de radiation.

3.1 Première application : Villes placées de manière homogène

En premier lieu, nous avons considéré que chaque ville, x_i, peuplée d'une personne ($n_i = 1\ \forall i$), se situe à une distance d'une unité de ses voisines directes. Si nous prenons un échantillon de 10 villes ($N = 10$ dans ce cas-ci), nous pouvons schématiser cette hypothèse de la manière suivante :

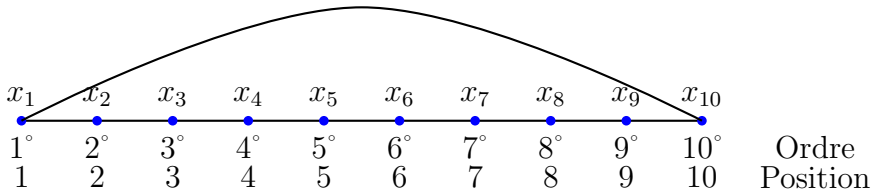


FIGURE 3.1 – Représentation schématique avec 10 villes ($N = 10$)

A chaque ville x_i est donc attribué un ordre o_i - à savoir la place de la ville i par rapport aux autres - , et une position p_i, comme indiqués sur le schéma :

$$\begin{aligned} o_i &= i; \\ p_i &= i; \end{aligned}$$

où $i = 1, \ldots, N$.

Rappelons quelques définitions théoriques utilisées dans le programme.

Définition 5.

Dans le cas homogène, la **distance séparant 2 extrémités**, D, est définie comme :

$$D = N - 1.$$

Néanmoins, en considérant la condition périodique, un tour complet de x_1 à x_1 correspond à une distance de N unités.

Deux possibilités se présentent au voyageur pour se rendre d'un endroit i à un endroit j :
- il peut soit utiliser le chemin traditionnel direct (en allant de 1 à 6 par exemple) ;
- soit décrire une boucle (en passant par 10 pour revenir ensuite à 6).

Au final, d'un point de vue pragmatique, l'utilisateur choisira le chemin le plus court. Nous obtenons ainsi la définition suivante :

Définition 6.

La **distance** et la **différence d'ordre** séparant une ville i d'une ville j peuvent être définies par les formules suivantes :

$$d_{ij} = \min\{p_j - p_i; D - p_j + p_i\};$$

$$r_{ij} = \min\{o_j - o_i; D - o_j + o_i\};$$

$\forall i, j = 1, \ldots, N$.

Remarque : Pour plus de facilité, au lieu de dire différence d'ordre, nous utiliserons l'appellation ordre.

Nous obtenons ainsi 2 matrices : la matrice des distances, d, et la matrice des ordres, r, de dimension $N \times N$. Nous observons que les 2 matrices seront identiques, étant donné que chaque distance égale chaque ordre.

Cette observation établie, nous pouvons définir, dans le cas où $N = 10$ villes, les matrices de la manière suivante :

$$d = r = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 4 & 3 & 2 & 1 \\ 1 & 0 & 1 & 2 & 3 & 4 & 5 & 4 & 3 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 4 & 3 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 & 4 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 \\ 4 & 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 \\ 3 & 4 & 5 & 4 & 3 & 2 & 1 & 0 & 1 & 2 \\ 2 & 3 & 4 & 5 & 4 & 3 & 2 & 1 & 0 & 1 \\ 1 & 2 & 3 & 4 & 5 & 4 & 3 & 2 & 1 & 0 \end{pmatrix}.$$

De façon évidente, nous obtenons une diagonale nulle (la distance ou l'ordre d'une ville avec elle-même vaut bien 0). De même, la matrice est symétrique, car considérer la distance (respectivement l'ordre) entre une ville i et une ville j revient au même que de considérer la distance (respectivement l'ordre) entre cette ville j et cette ville i.

Distinguons ensuite la population établie entre chaque paire de villes ($\forall i, j = 1, \ldots, N$) suivant le modèle de radiation et le modèle de gravité, également appelé rang :

Définition 7.

En ce qui concerne le modèle de radiation, le **rang** entre les villes i et j, s_{ij}^R, symbolise la population totale comprise entre ces 2 villes. Etant donné que chaque ville est peuplée d'une seule personne, il peut être défini de la manière suivante :

$$s_{ij}^R = r_{ij} - 1.$$

Dans ce cas bien particulier, nous pouvons définir la **population établie**, pour le modèle de gravité, comme :

$$s_{ij}^G = s_{ij}^R.$$

Dans notre programme expérimental, nous pouvons ensuite reprendre les formules établies dans l'article de [Simini *et al.*] :

$$\begin{aligned} T_{ij}^G &= \frac{n_i^\alpha n_j^\beta}{f(d_{ij})}; \\ T_{ij}^R &= \frac{n_i^2 n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})}. \end{aligned}$$

En remplaçant directement le poids de chaque ville par 1, nous obtenons la définition suivante :

Définition 8.

Le **flux obtenu de voyageurs** se rendant d'un endroit i à un endroit j pour les modèles de gravité, T_{ij}^G, et de radiation, T_{ij}^R, s'écrit respectivement sous la forme suivante :

$$\begin{aligned} T_{ij}^{\mathrm{G}} &= \frac{1}{(1+s_{ij}^G)(2+s_{ij}^G)}; \\ T_{ij}^{\mathrm{R}} &= \frac{1}{(1+s_{ij}^R)(2+s_{ij}^R)}. \end{aligned}$$

Afin que nos résultats soient similaires, nous avons décidé de considérer comme fonction de dissuasion du modèle de gravité, le dénominateur de la fonction T_{ij}^{R}. Nous avons également fait le choix de prendre le même numérateur, à savoir $n_i^2 n_j$.

Définissons à présent les 2 nouveaux vecteurs suivants : $T_G(dst)$ et $Nbr_G(dst)$.

Définition 9.

Pour le modèle de gravité, $T_G(dst)$ - soit la **somme des flux** des paires de villes séparées par une distance dst - et $Nbr_G(dst)$ - soit le **nombre de paires de villes** séparées par une distance dst - peuvent s'exprimer comme :

$$\begin{aligned} T_G(dst) &= \sum_{\{i,j:d_{ij}=dst\}} T_{ij}^{\mathrm{G}}; \\ Nbr_G(dst) &= \sum_{\{i,j:d_{ij}=dst\}} 1; \end{aligned}$$

pour toute distance dst, comprise entre le minimum et le maximum de la matrice des distances d.

Sur base de ces deux formules, nous pouvons enfin établir, pour le modèle de gravité, le vecteur suivant :

Définition 10.

Le **flux moyen attendu de personnes** ralliant 2 villes séparées par une distance dst, $T_{moyen_G}(dst)$ se définit, pour toute distance dst, comme

$$T_{moyen_{G(dst)}} = \frac{T_G(dst)}{Nbr_G(dst)}.$$

Nous pouvons faire de même pour le modèle de radiation, excepté que nous remplaçons la distance dst par le rang rg. Nous obtenons ainsi $T_{moyen_R}(rg)$, où rg prend des valeurs comprises entre le minimum et le maximum de la matrice des rangs s.

Au final, en considérant 100 villes placées de manière homogène, nous obtenons le graphique suivant des flux moyens attendus en fonction de la distance dst ou du rang rg, pour les modèles de gravité et de radiation :

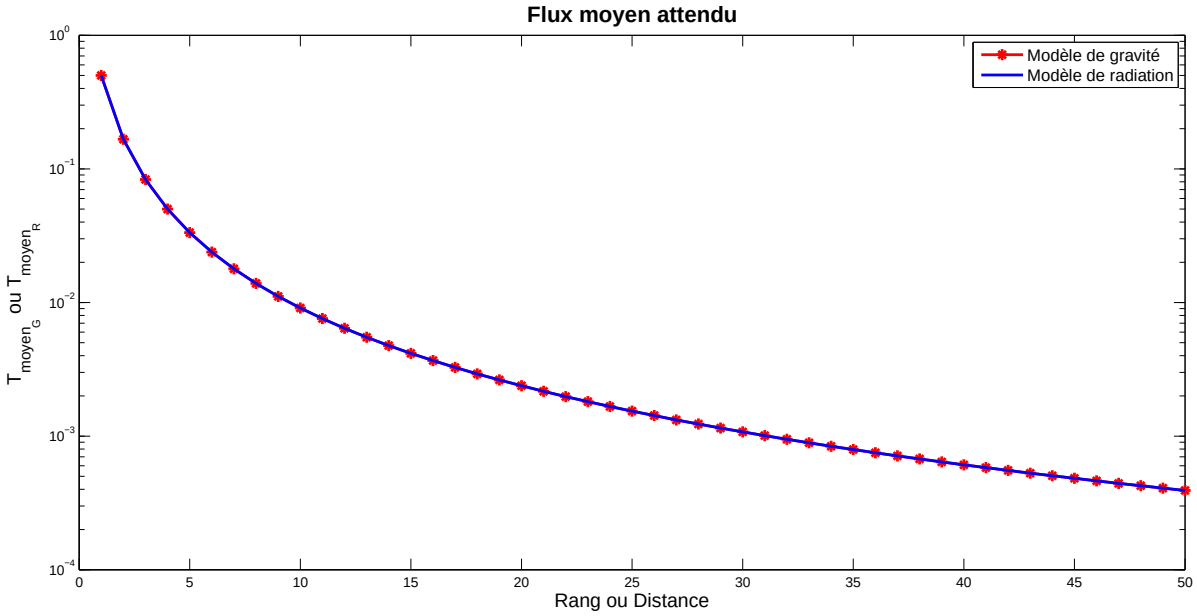


FIGURE 3.2 – Flux moyen attendu - $N = 100$

Nous pouvons remarquer que les 2 graphes se superposent. Les deux modèles sont donc équivalents, ce qui était déjà apparu dans la partie théorique du présent mémoire.
De plus, le flux moyen diminue lorsque le rang ou la distance augmentent, constatation qui aurait pu être faite lors de la définition des flux par les modèles de Simini et de gravité.

Développons une seconde méthode pour calculer et représenter graphiquement les flux moyens attendus en fonction du rang : pour ce faire, il suffit de calculer le flux moyen attendu pour le modèle de radiation à partir de celui obtenu pour le modèle de gravité.

Avant toute chose, rappelons la définition de probabilité conditionnelle :

Définition 11.

Soit $B \subset \Omega$ un événement tel que $P(B) > 0$. Pour tout $A \subset \Omega$, la **probabilité conditionnelle** de A sachant que la condition B est remplie, est définie par :

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Nous considérons donc $P(rg \mid dst)$.
Pour ce faire, nous avons calculé les 2 probabilités $P(dst)$ et $P(rg \cap dst)$ de la manière suivante :

Définition 12.

La **probabilité que la distance soit égale à** dst entre 2 villes, $P(dst)$, se présente comme le nombre de paires de villes séparées par une distance dst divisé par le nombre total de paires de villes :

$$P(dst) = \frac{Nbr_G(dst)}{Nbr_p} \quad \forall\, dst;$$

où $Nbr_p = C_N^2$ symbolise le nombre total de paires de villes et $Nbr_G(dst)$ le nombre de paires de villes séparées par une distance dst (*cf.* Définition 5).

Définition 13.

La **probabilité qu'une paire de villes soit séparée par une distance** dst **et par un rang** rg se présente comme le rapport entre le nombre de paires de villes ayant ces caractéristiques (que nous noterons K) et le nombre total de paires de villes :

$$P(rg \cap dst) = \frac{K}{Nbr_p} \;\forall\, dst, rg.$$

Enfin, nous pouvons relier le flux moyen pour le modèle de radiation à cette probabilité $P(rg \mid dst)$ et au flux moyen du modèle de gravité :

$$T_1^R(rg) = \sum_{dst} P(rg|dst) T_{ij}^G(dst)$$

$\forall rg$ compris entre le minimum et le maximum de la matrice des rangs s et $\forall i, j : d_{ij} = dst$. Nous obtenons donc un nouveau vecteur T_1^R.

Comparons ce T_1^R obtenu, représenté en rouge, au T_{moyen_R}, représenté en bleu, pour 100 villes placées de manière homogène :

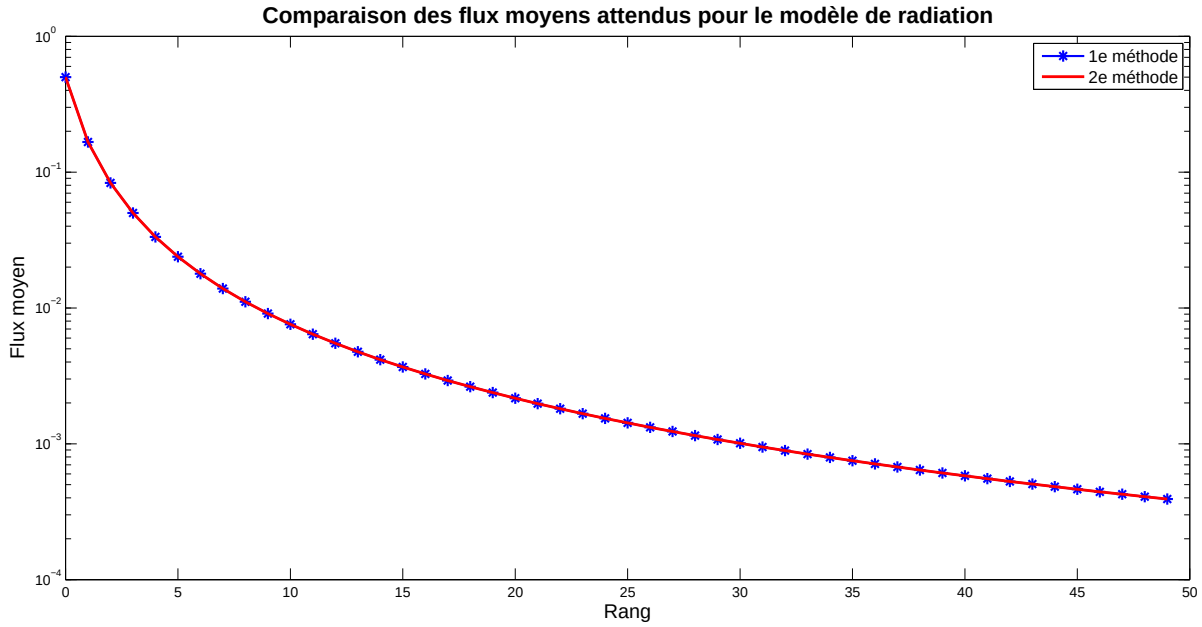


FIGURE 3.3 – Comparaison des 2 méthodes - $N = 100$

De nouveau, nous constatons que, dans le cas d'une répartition homogène, les 2 méthodes sont équivalentes et donnent le même taux moyen pour chaque rang considéré. Cette fois encore, plus le rang augmente entre 2 villes, plus le flux de voyageurs diminue.

3.2 Deuxième application : Villes placées en 2 groupes distincts

Pour cette deuxième application, nous avons de nouveau considéré que chaque ville, x_i, est peuplée d'une seule personne ($n_i = 1\ \forall i$). Néanmoins, nous séparons cette fois les villes en 2 groupes : chaque groupe se constitue d'un ensemble de villes, toutes distantes

à une unité de la (les) ville(s) voisine(s). L'écart séparant les 2 groupes, soit le nombre l, peut être considéré comme le nombre de villes manquant dans la suite entre les extrémités de chaque groupe. Représentons cette situation où $N = 5$ et $l = 3$:

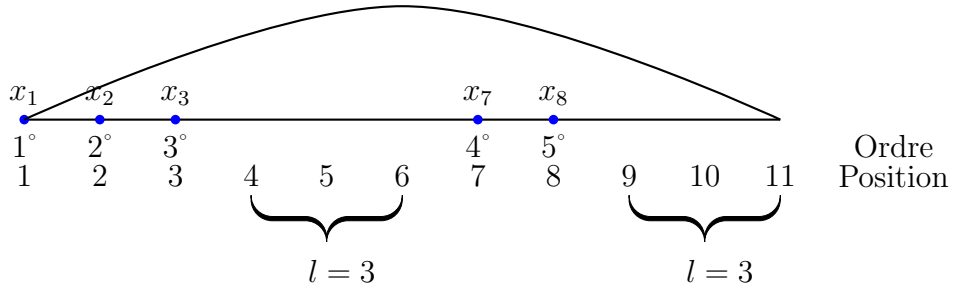


FIGURE 3.4 – Représentation pour $N = 5$ avec $l = 3$

Il parait évident que si nous donnons au nombre l la valeur 0, nous retombons sur le cas précédent.
Reprenons les formules utilisées antérieurement ; et adaptons-les.

Définition 14.

Dans le cas de 2 groupes distincts, séparés par le paramètre l, la **distance** D **séparant les 2 extrémités** du segment est définie comme

$$D = N - 1 + 2 \times l.$$

Néanmoins, en considérant la condition périodique, un tour complet de x_1 à x_1 correspond à une distance de $N + 2 \times l$ unités.

A chaque ville x_i sont attribués un ordre, o_i, et une position p_i, comme sur le schéma ci-dessus. En appliquant les mêmes formules pour obtenir les matrices de distance d et d'ordre r, nous obtenons, pour 5 villes réparties en 2 groupes distants de $l = 3$:

$$d = \begin{pmatrix} 0 & 1 & 2 & 5 & 4 \\ 1 & 0 & 1 & 5 & 5 \\ 2 & 1 & 0 & 4 & 5 \\ 5 & 5 & 4 & 0 & 1 \\ 4 & 5 & 5 & 1 & 0 \end{pmatrix} \quad \text{et } r = \begin{pmatrix} 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 \end{pmatrix}$$

Cette fois-ci, nous observons que les matrices ne sont pas identiques, étant donné que nous avons 2 groupes de villes. Néanmoins, elles restent symétriques et nulles sur la diagonale. De plus, remarquons que la distance peut prendre plus de valeurs distinctes par rapport à

la différence d'ordre : alors que deux villes peuvent être distantes jusqu'à 5 unités maximum, la différence d'ordre maximale s'élève à 2 ; ce qui implique qu'une seule ville est comprise entre l'origine et la destination.

Poursuivons à présent notre raisonnement en distinguant la population établie (ou le rang) entre chaque paire de villes ($\forall i, j = 1, \ldots, N$), suivant les modèles de radiation et de gravité :

Définition 15.

Pour le modèle de radiation, le **rang** entre la ville i et la ville j pour le modèle de radiation, s_{ij}^R, symbolise la population totale comprise entre ces 2 villes. Comme chaque ville est constituée d'une seule personne, nous pouvons le déterminer de la sorte :

$$s_{ij}^R = r_{ij} - 1.$$

Par contre, dans le cas du modèle de gravité, s_{ij}^G sera défini par :

$$s_{ij}^G = \begin{cases} s_{ij}^R & \text{si } d_{ij} = r_{ij} = s_{ij}^R + 1; \\ s_{ij}^R + l & \text{sinon.} \end{cases}$$

En procédant de manière similaire au cas homogène, en considérant 100 villes séparées en 2 groupes distincts, avec $l = 50$, nous pouvons représenter les flux moyens attendus en fonction de la distance dst ou du rang rg, pour les modèles de radiation et de gravité :

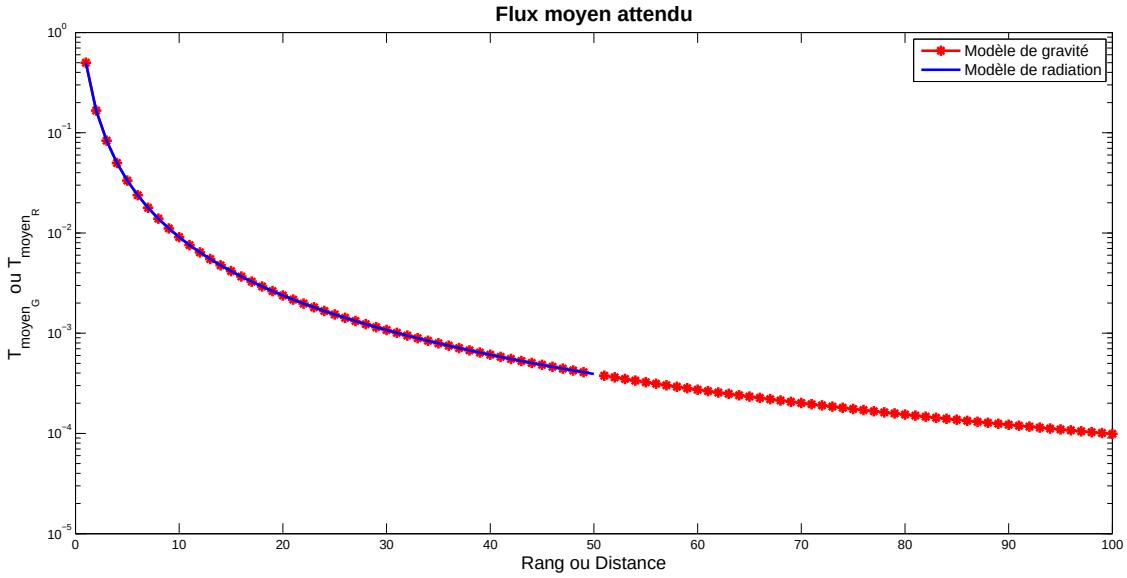


FIGURE 3.5 – Flux moyen attendu - $N = 100$ et $l = 50$

Nous obtenons ici aussi une superposition des 2 graphes - ce qui nous semble logique étant donné que nous avons défini de la même manière les modèles de gravité et de radiation - . Néanmoins, alors que le rang vaut au maximum 50, la distance peut aller jusqu'à 100. De plus, nous observons une discontinuité en 50, pour le modèle de gravité : en effet, aucune paire de villes n'est distante de 50 unités.
Nous remarquons également la décroissance de notre fonction : plus le rang ou la distance augmentent, plus le flux diminue.

Dressons le même graphique en réduisant cette fois le paramètre l à 30 - au lieu de 50 - :

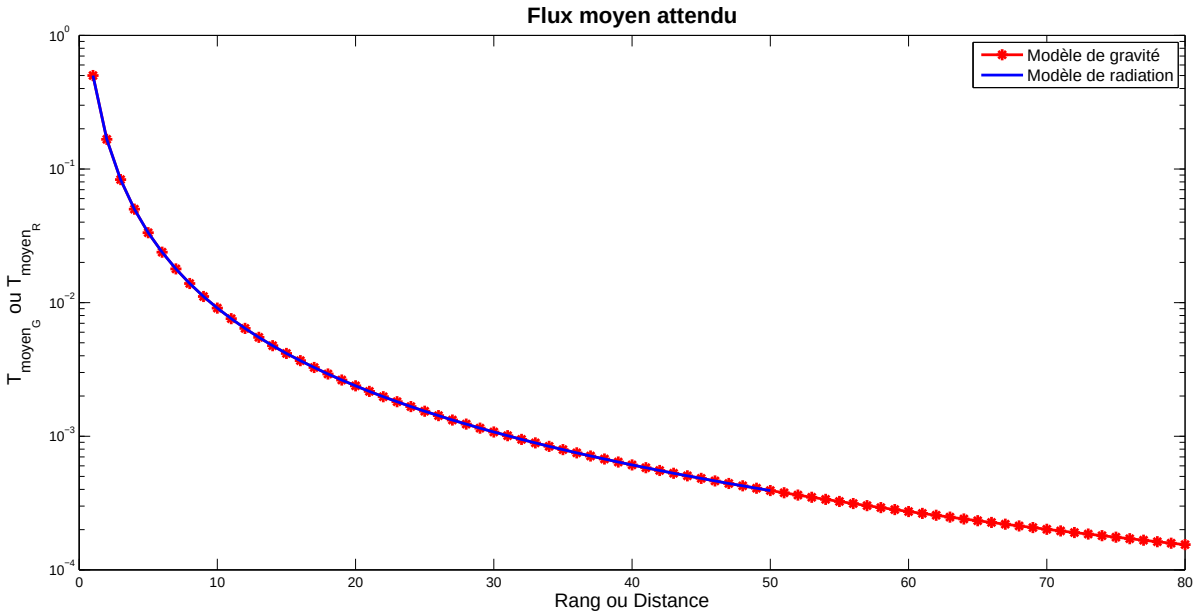


FIGURE 3.6 – Flux moyen attendu - $N = 100$ et $l = 30$

De nouveau nos courbes se superposent jusqu'à un rang ou une distance de 50. Le flux moyen du modèle de gravité, quant à lui, continue et est défini jusqu'à un écart maximal de 80 unités entre 2 villes.
Comme précédemment, le flux diminue au fur et à mesure que le rang ou la distance augmentent.

Passons à présent à la seconde méthode, en posant $P(rg \mid dst) = 0$ si $P(dst) = 0$, étant donné que nous ne pouvons diviser par 0.

Comparons le T_1^R obtenu - représenté en rouge - au T_{moyen_R} - représenté en bleu -, pour 100 villes réparties en 2 groupes espacés de 50 unités.

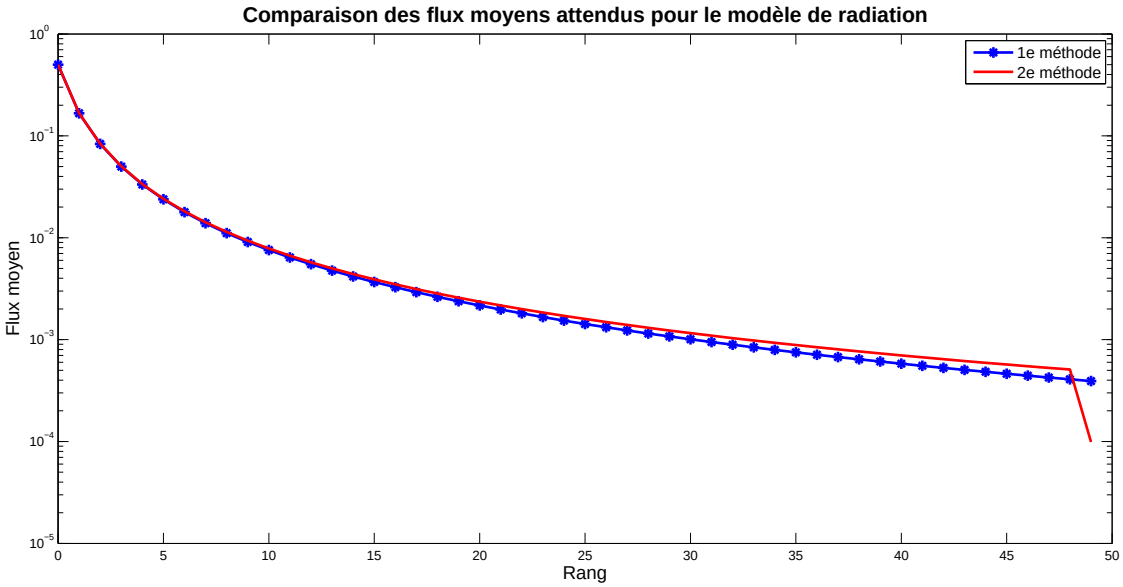


FIGURE 3.7 – Flux moyen attendu - $N = 100$ et $l = 50$

Lorsque $l = 50$, nous observons une superposition des 2 courbes jusqu'à une différence de rang de 10 ; à partir de ce point, elles se distinguent peu à peu.

Si nous rapprochons les 2 groupes de villes, en réduisant l à 30, nous obtenons le graphique suivant :

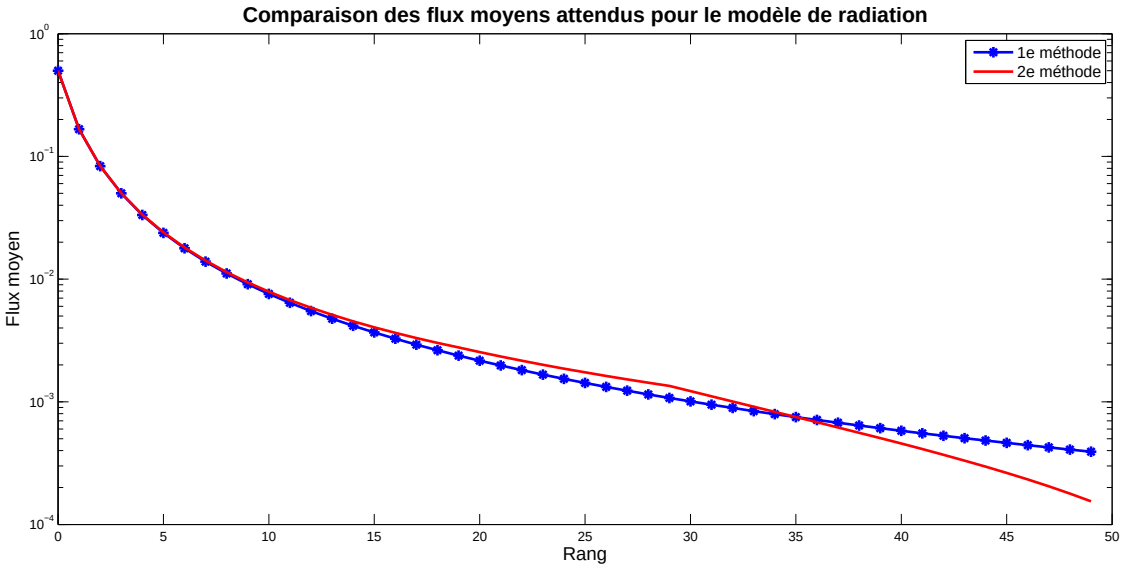


FIGURE 3.8 – Flux moyen attendu - $N = 100$ et $l = 30$

De nouveau, nos graphes se superposent jusqu'à un rang de 10 environ, avant que la courbe de la deuxième méthode se différencie.

3.3 Généralisation

Supposons à présent que chacune des villes considérées compte un nombre fini d'habitants, par exemple 10 habitants, à savoir $n_i = 10 \ \ \forall i = 1, \ldots, N$.
Toutes nos définitions restent d'application, excepté aux quelques endroits où intervient le nombre d'habitants par ville - comme, par exemple, la définition du rang - .

Cependant, une remarque générale pour l'ensemble de cette partie s'impose : étant donné que le nombre d'habitants par ville est supérieur à 1, le rang va fortement se distinguer de la distance. Dès lors, nous ne pourrons pas représenter sur un même graphique les flux obtenus avec les deux modèles. Pour pouvoir les comparer, nous représenterons donc le flux prédit par le modèle de radiation en fonction du flux prédit par le modèle de gravité. Ainsi, si les points sont alignés sur une même droite (à savoir la première bissectrice), cela signifiera que les deux modèles sont identiques. Attention, toutefois, seule une partie du flux du modèle de gravité sera représentée sur ce type de graphique, étant donné qu'il possède plus de valeurs.

3.3.1 Villes placées de manière homogène

Admettons comme hypothèse que toutes les villes sont distantes d'une unité de leur(s) voisine(s) directe(s). Nous constatons une légère modification au niveau du taux de population établie entre chaque paire de villes (rang). En effet, chaque entité comptant 10 personnes, il suffit tout simplement de multiplier par 10 les résultats obtenus dans le premier cas.

Définition 16.

Pour le modèle de radiation, le **rang** entre la ville i et la ville j, s_{ij}^R, symbolise la population totale comprise entre ces 2 villes. Chaque ville comptant 10 personnes, il peut donc être défini de la manière suivante :

$$s_{ij}^R = (r_{ij} - 1) \cdot 10.$$

Pour le modèle de gravité, nous aurons :

$$s_{ij}^G = s_{ij}^R.$$

De même, si nous reprenons la définition du flux obtenu de voyageurs ralliant une ville à une autre dans l'article de [Simini *et al.*], nous obtenons la modification suivante :

Définition 17.

Le **flux de voyageurs** ralliant un endroit i à un endroit j pour les modèles de gravité, T_{ij}^G, et de radiation, T_{ij}^R, se calcule respectivement de la manière suivante :

$$T_{ij}^{\mathrm{G}} = \frac{n_i^2\, n_j}{(n_i + s_{ij}^G)(n_i + n_j + s_{ij}^G)};$$

$$T_{ij}^{\mathrm{R}} = \frac{n_i^2\, n_j}{(n_i + s_{ij}^R)(n_i + n_j + s_{ij}^R)}.$$

Afin que nos résultats soient similaires, nous avons décidé de prendre comme fonction de dissuasion du modèle de gravité le dénominateur de la fonction T_{ij}^{R}. Nous avons également choisi le même numérateur, à savoir $n_i^2 n_j$.

Au final, pour 100 villes placées de façon homogène, nous obtenons les graphiques suivants pour les flux moyens attendus par les modèles de gravité (en fonction de la distance ; à gauche) et de radiation (en fonction du rang ; à droite) :

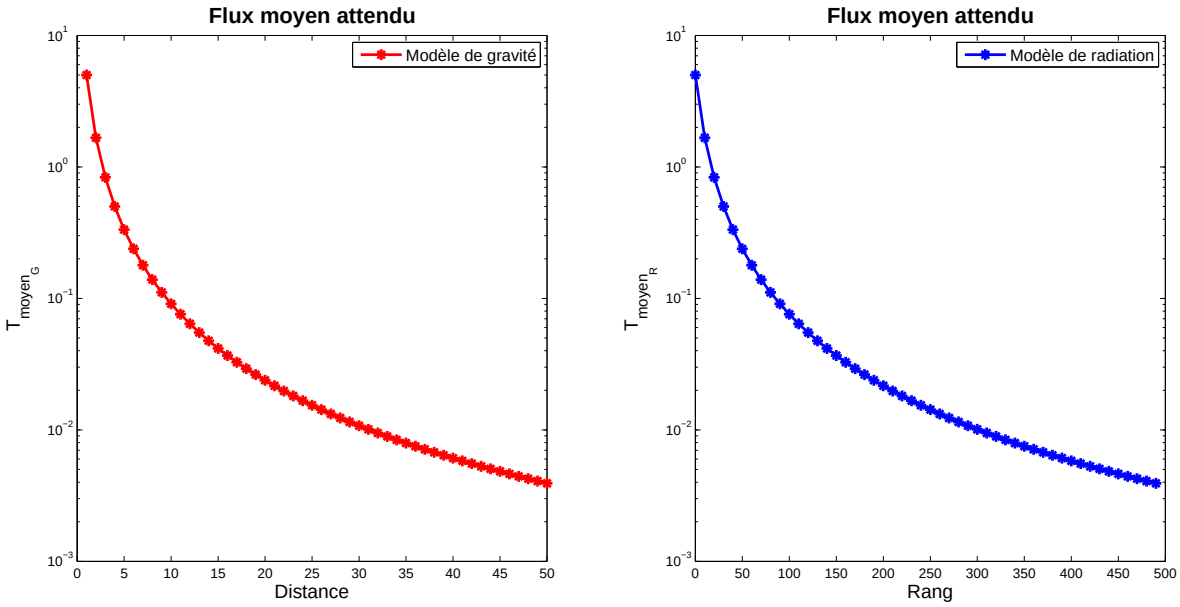


FIGURE 3.9 – Flux moyen attendu - $N = 100$

Alors que la distance peut s'élever jusqu'à 50 unités, le rang quant à lui peut prendre des valeurs jusqu'à dix fois supérieures à ce maximum. Les deux fonctions sont décroissantes, signifiant ainsi un plus faible taux de voyageurs pour une plus grande distance (et respec-

tivement un plus grand rang).
Nous noterons par ailleurs une certaine similitude entre ces 2 graphiques.

Afin de nous aider à vérifier cette hypothèse, représentons le flux prédit par le modèle de radiation en fonction de celui prédit par le modèle de gravité :

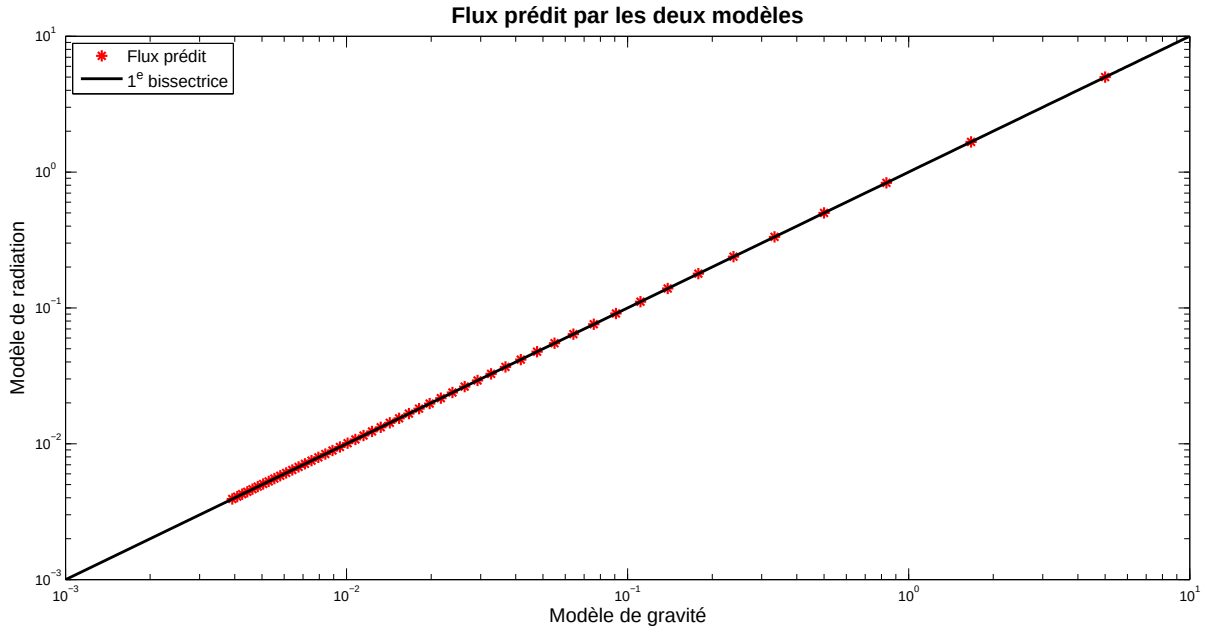


FIGURE 3.10 – Flux prédit par les deux modèles - $N = 100$

La lecture du graphique tend à démontrer que le nombre de navetteurs prédit par le modèle de radiation équivaut à celui du modèle de gravité, étant donné que tous les points sont alignés sur la première bissectrice. Cette observation valide ainsi notre supposition de départ : les deux modèles sont équivalents.

Comparons les flux obtenus ci-dessus pour le modèle de radiation avec le T_1^R admis via le calcul des probabilités conditionnelles.

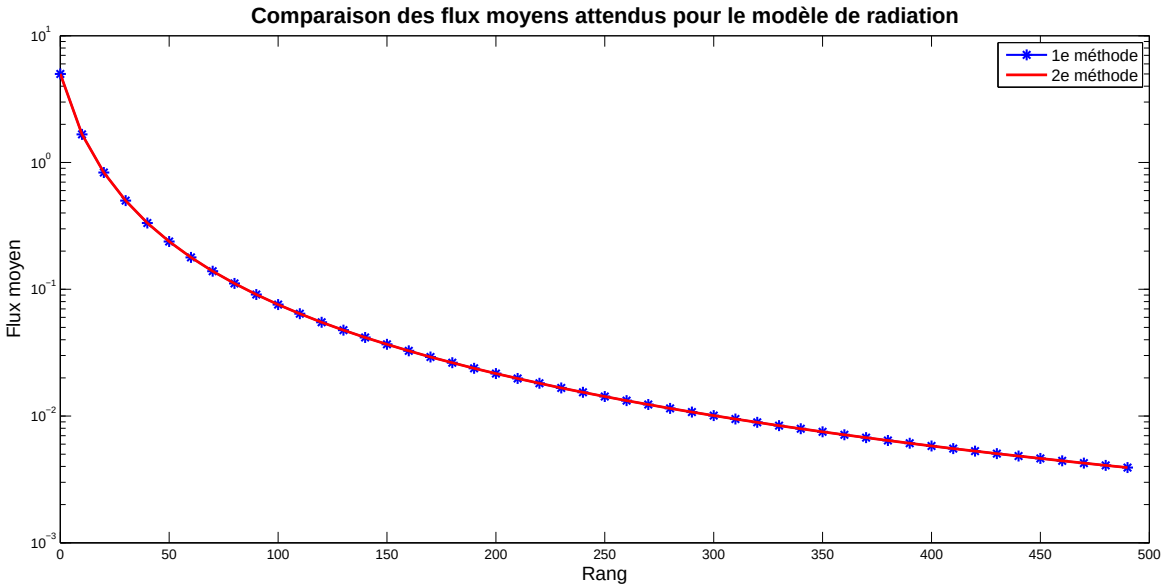


FIGURE 3.11 – Flux moyen attendu - $N = 100$

Dans le cas d'une répartition homogène, les 2 méthodes sont équivalentes et donnent le même taux moyen pour chaque rang considéré. L'allure du graphique est similaire à toutes celles des graphes précédents ; avec un flux décroissant au fur et à mesure que le rang augmente.

3.3.2 Villes placées en 2 groupes distincts

Comme dans le cas précédent, les définitions des rangs vont se trouver également modifiées du fait que nous avons élargi le nombre d'habitants de chaque ville à 10.

Définition 18.

Pour le modèle de radiation, le **rang** entre la ville i et la ville j, s_{ij}^R, symbolise la population totale comprise entre ces 2 villes. Chaque ville comptant 10 personnes, il peut donc être défini de la manière suivante :

$$s_{ij}^R = (r_{ij} - 1) \cdot 10.$$

Par contre, dans le cas du modèle de gravité, nous pouvons déterminer s_{ij}^G de la sorte :

$$s_{ij}^G = \begin{cases} s_{ij}^R & \text{si } d_{ij} = r_{ij} = \frac{s_{ij}^R}{10} + 1; \\ s_{ij}^R + l \cdot 10 & \text{sinon.} \end{cases}$$

Comme l représente le nombre de "villes fantômes" manquant dans la suite, il convient de bien prendre en considération leur taux de population, soit $l \cdot 10$, lorsque nous calculons la population établie entre chaque paire de villes pour le modèle de gravité.

En considérant 100 villes réparties en 2 groupes distincts avec un écart de $l = 50$ unités, représentons les flux moyens attendus en fonction de la distance dst ou du rang rg pour les modèles de gravité (à gauche) et de radiation (à droite) :

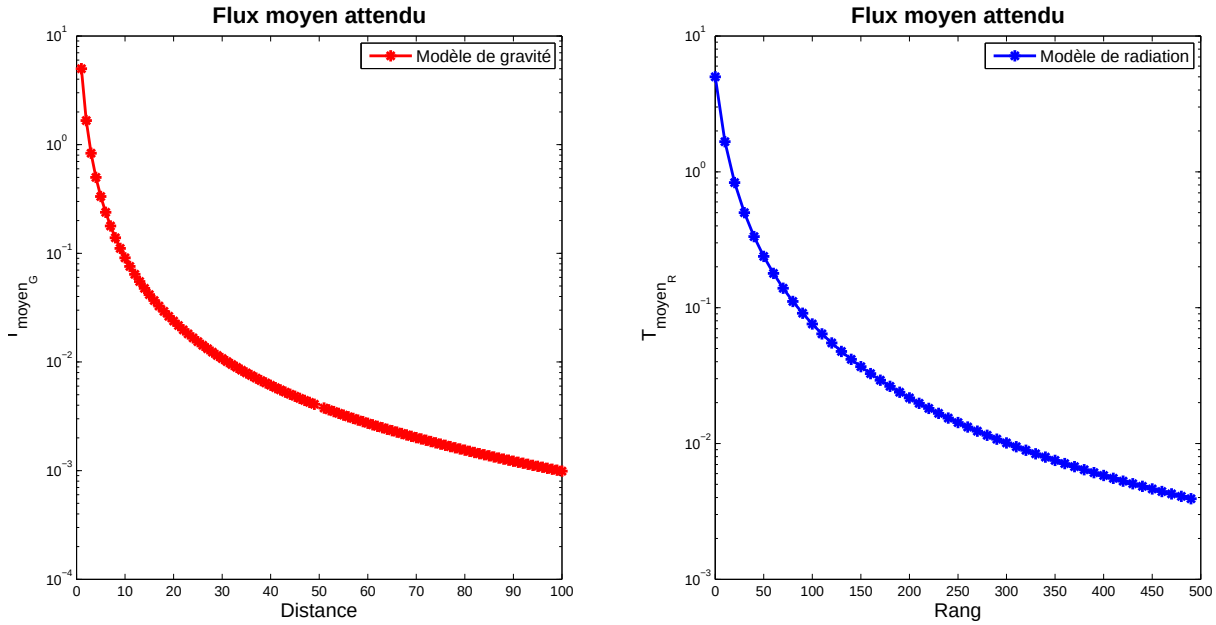


FIGURE 3.12 – Flux moyen attendu - $N = 100$ et $l = 50$

Comme constaté précédemment, alors que le rang atteint une valeur maximale de 500 personnes, la distance quant à elle peut monter jusqu'à 100 unités. De nouveau, il existe un point de discontinuité pour le modèle de gravité (même s'il apparaît peu visible sur ce graphique), lorsque la distance s'élève à 50 unités. De plus, dans les deux cas, le flux décroît toujours au fur et à mesure que la distance ou le rang augmentent.
De nouveau, ces graphes présentent une similitude jusqu'à une distance de 50 unités. De fait, comme nous l'avons expliqué dans la remarque préliminaire de cette section, le modèle de gravité prédit des flux pour plus de valeurs que le modèle de radiation : alors que la distance peut prendre 100 valeurs distinctes, le rang n'en présente que la moitié.

Afin de tenter de vérifier l'hypothèse de similarité des deux modèles, représentons le flux prédit par le modèle de radiation en fonction de celui prédit par le modèle de gravité (en ne considérant toutefois que les 50 premières valeurs, si nous tenons compte de la remarque précédente) :

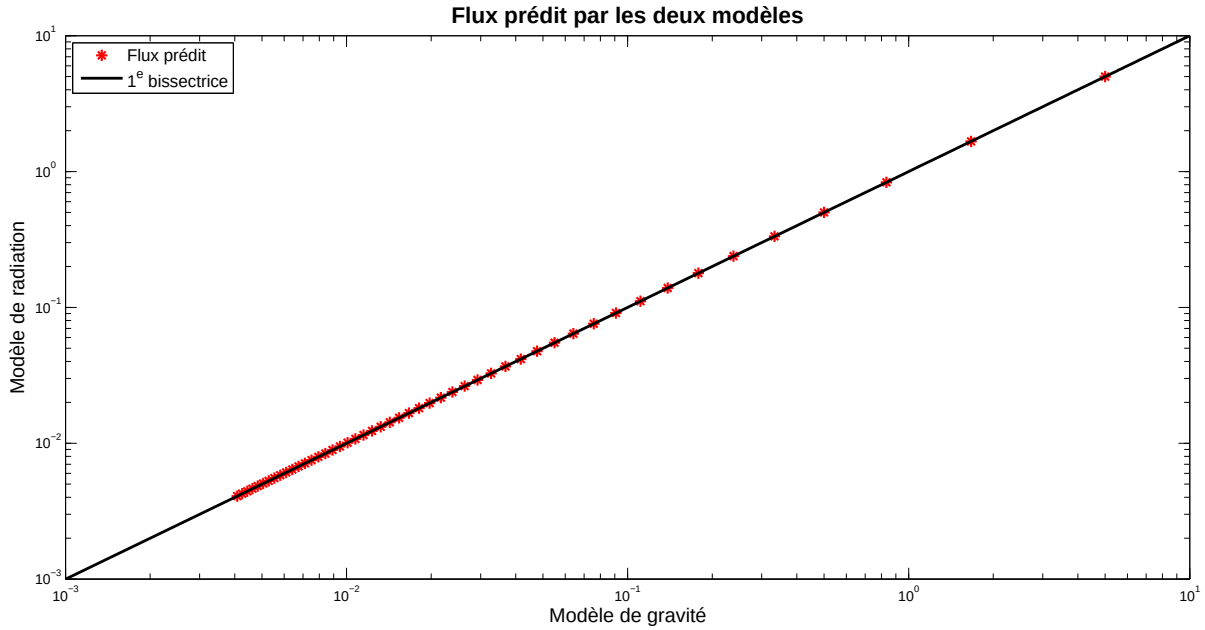


FIGURE 3.13 – Flux prédit par les deux modèles - $N = 100$ et $l = 50$

Ici aussi, nous observons que les différents flux s'alignent sur la première bissectrice. Les deux modèles se révèlent donc équivalents.

Représentons les mêmes graphiques, en réduisant toutefois le paramètre l à 30 unités :

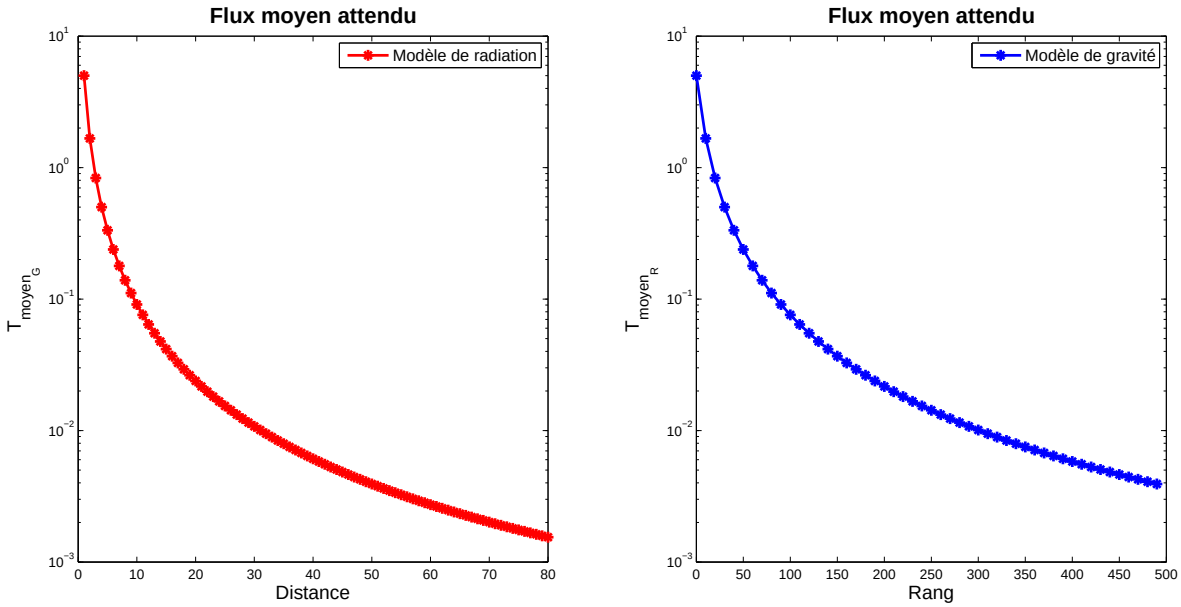


FIGURE 3.14 – Flux moyen attendu - $N = 100$ et $l = 30$

De nouveau, le rang se distingue de la distance : tandis que le premier peut toujours atteindre une valeur maximale de 500 habitants, la seconde s'élève au plus à 80 unités. Nous remarquons également la décroissance du flux lorsque la distance ou le rang augmentent. Comme dans le cas précédent où $l = 50$, ces graphes semblent similaires jusqu'à une distance de 50 unités. De fait, alors que le rang peut toujours prendre 50 valeurs distinctes, la distance peut cette fois en présenter 80.

Afin de nous aider à vérifier cette hypothèse, représentons le flux prédit par le modèle de radiation en fonction de celui prédit par le modèle de gravité (pour les 50 premières valeurs) :

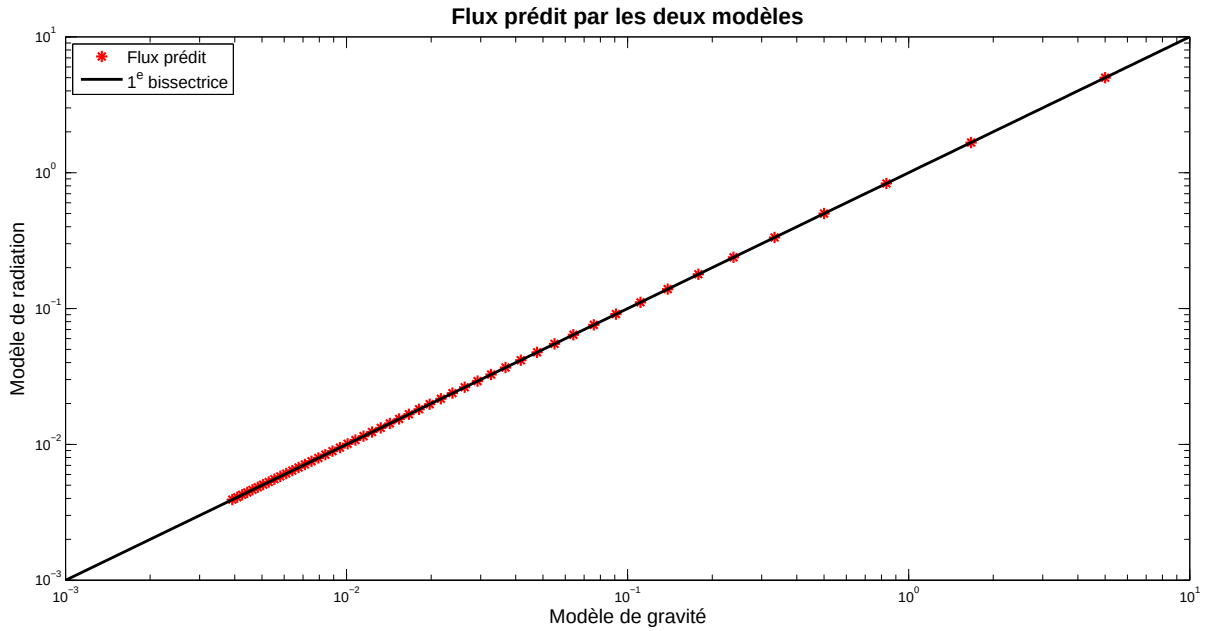


FIGURE 3.15 – Flux prédit par les deux modèles - $N = 100$ et $l = 30$

Les différents flux s'alignent le long de la première bissectrice, signifiant ainsi l'équivalence de nos deux modèles.

Enfin, comparons le flux obtenu ci-dessus par le modèle de radiation (représenté en bleu) avec le T_1^R admis via le calcul des probabilités conditionnelles (représenté en rouge). Dressons tout d'abord le tableau tel que $l = 50$.

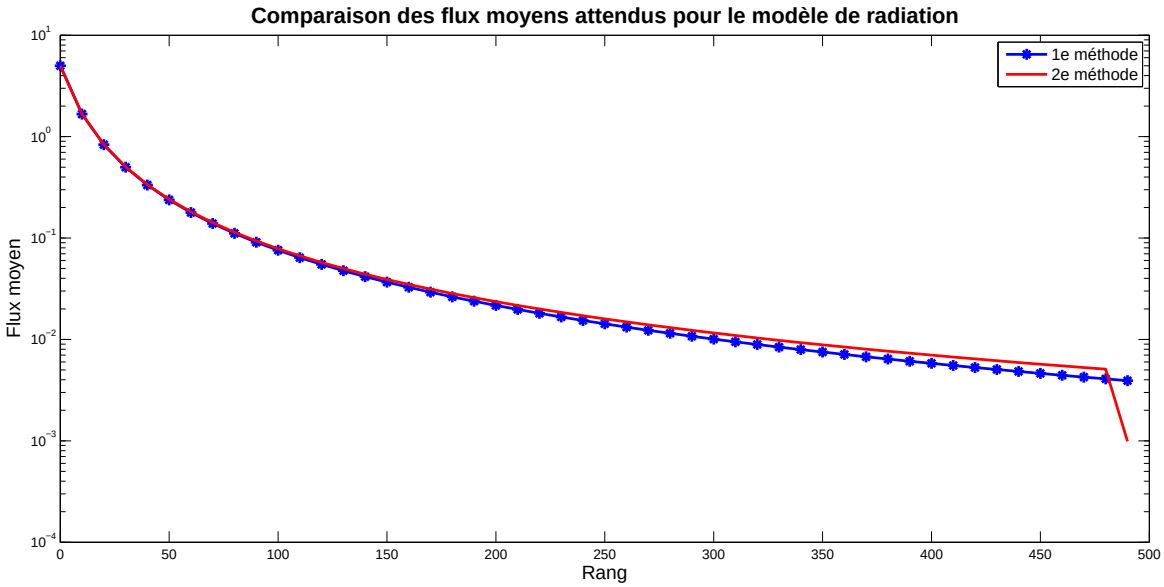


FIGURE 3.16 – Flux moyen attendu - $N = 100$ et $l = 50$

A partir d'un rang proche de 10, nous remarquons une déviation du flux à partir de la méthode utilisant les probabilités (représenté en rouge).
De plus, les flux prédits à partir des deux modèles diminuent au fur et à mesure que le rang augmente.

Si nous modifions le paramètre l à 30, nous obtenons le graphe suivant :

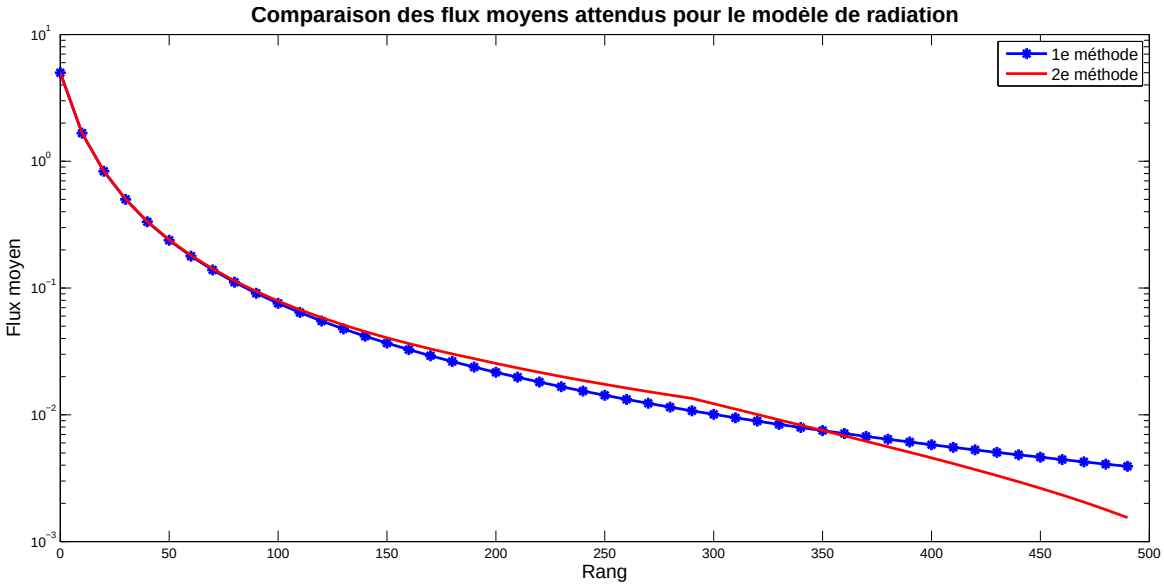


FIGURE 3.17 – Flux moyen attendu - $N = 100$ et $l = 30$

Les deux flux moyens se différencient dès que le rang prend des valeurs supérieures à 10, c'est-à-dire dès qu'il y a une ville d'écart.
Une nouvelle fois, la diminution du flux au fur à mesure de l'augmentation de la distance, se vérifie.

3.4 Discussion

A travers ce chapitre, nous avons essayé de créer un modèle aussi simple que possible, avec des hypothèses fort réductrices, générant ainsi une population homogène, dans le but d'étudier l'effet de l'inhomogénéité sur les modèles de gravité et de radiation.

Dans un premier temps, nous avons considéré une population totalement homogène : même nombre d'habitants dans toutes les villes, placées de surcroît à équidistance. Comme nous avons pu l'établir dans la partie théorique de ce mémoire, le modèle de gravité équivaut au modèle de radiation, quel que soit le nombre d'habitants considéré. De plus, la deuxième méthode utilisée pour prédire le flux en fonction du rang à partir du modèle de gravité et des probabilités conditionnelles, correspond également au modèle de radiation défini. De la sorte, deux méthodes ont pu être trouvées et expérimentées.

Par la suite, après avoir séparé les villes en deux groupes distants d'une entité quelconque (pour nous concentrer cette fois sur des villes sensiblement non homogènes), nous avons pu établir que le modèle de gravité équivaut au modèle de radiation, quelle que soit la population considérée (en gardant toutefois le même nombre d'habitants par ville). Néanmoins, certains points sont absents : par exemple, certaines distances ne sont jamais rencontrées, comme ce fut le cas sur la figure 3.5.
La deuxième méthode, basée sur le modèle de gravité et des probabilités conditionnelles, diffère du modèle de radiation, et prédit des flux différents. Comme nous avons pu le constater, plus le rang augmente, plus la deuxième méthode surestime le nombre de navetteurs ... jusqu'à atteindre une valeur de rang pour laquelle le flux se retrouve sous-estimé. Autre constatation : nous pouvons remarquer que le flux calculé à partir des probabilités conditionnelles est surestimé ou sous-estimé lorsque plusieurs villes commencent à s'aligner entre l'origine et la destination. Plus le nombre de villes "fantômes" s'approchera de 0, plus nombreux seront les flux sous-estimés.

Cette différence entre les deux méthodes pourrait s'expliquer par l'utilisation de la probabilité calculée. En étudiant attentivement les différents résultats obtenus, nous pouvons remarquer que la probabilité conditionnelle équivaut à 1 dans le cas homogène. Par contre, elle sera comprise entre 0 et 1 lorsque nous considérons deux groupes de villes. En effet, dans le premier cas, le nombre de paires de villes caractérisées par un même rang et par une même distance équivaut au nombre de villes distantes de cette dernière valeur.

La distance ne se différencie donc pas de la différence d'ordre (qui nous permet ensuite d'obtenir le rang). Dans le second cas, par contre, l'espace entre deux villes peut prendre plus de valeurs que le rang, étant donné que nous imposons un intervalle entre les deux groupes. Ainsi, le nombre de paires de villes caractérisées par un même rang et par une même distance diffère du nombre de villes séparées par cette même distance.

Une étude encore plus approfondie nous aurait sans doute permis de déterminer pour quelles paires de villes le nombre de voyageurs est sous-évalué ou surévalué, et quelles sont leurs caractéristiques. D'autre part, à partir du jeu de données préalablement établi, le choix de la fonction de dissuasion s'est révélé aisé, tout comme il a été assez facile de définir la population totale établie entre une origine et une destination pour le modèle de gravité à partir du rang. Toutefois, les données empiriques sont beaucoup plus complexes, et le choix de la fonction de dissuasion va se révéler bien plus délicat, dans une seconde démarche ...

CHAPITRE 4

Passage en deux dimensions

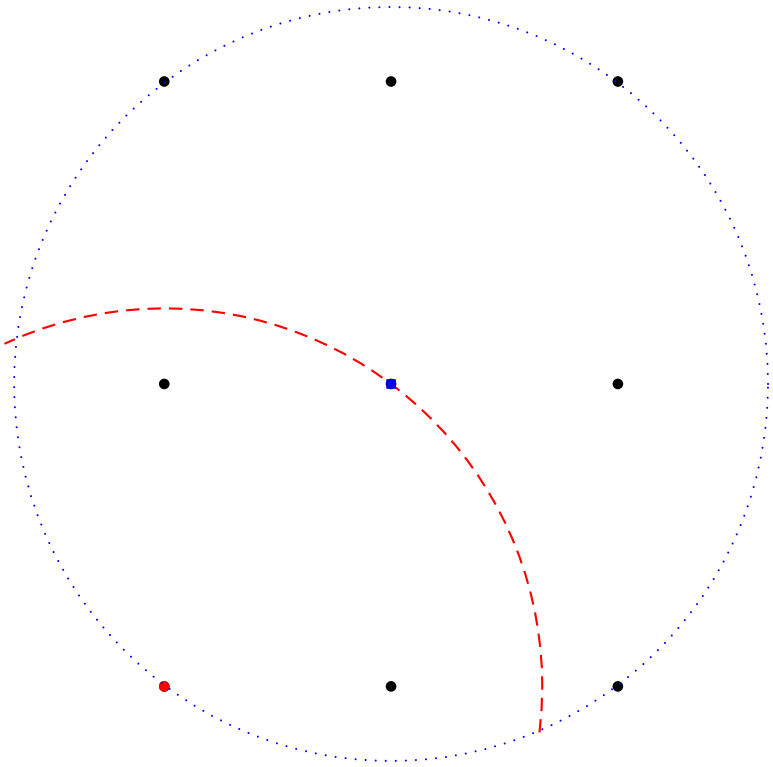
Après avoir étudié ce modèle à une dimension, nous pouvons remarquer qu'il s'agit d'un modèle à emplacement de villes suivant des caractéristiques qui nous permettent de constater des propriétés particulières. Néanmoins, les données empiriques se révèlent beaucoup plus complexes : il y a très peu de chances que des villes voisines possèdent le même nombre d'habitants et soient toutes placées de manière homogène.
Considérant que les hypothèses faites précédemment se révèlent fort simplistes et très peu réalistes, en vue d'une analyse plus pertinente, nous allons suggérer dans cette partie un passage en deux dimensions.

4.1 Rang

La procédure de passage en deux dimensions entraîne une légère modification de la matrice de rang - pour rappel, le rang prend en compte la population totale établie entre les 2 villes - . De fait, à l'inverse de ce que nous avons pu examiner dans le cas de villes placées sur une seule même droite, il n'est pas garanti que cette matrice soit symétrique. Ici, en l'occurrence, le rang sera défini de la manière suivante :

$$s_{ij} = \sum_{\{k:0<d_{ik}\leq d_{ij} \text{ et } k\neq j\}} n_k \quad \forall i,j = 1,\ldots,N.$$

Autrement dit, pour le déterminer, il suffit d'additionner le nombre d'habitants des villes comprises dans le cercle de rayon d_{ij} et de centre i, en retirant toutefois les villes d'origine et de destination. Considérons, par exemple, 9 villes réparties de la manière suivante :



Si nous établissons une comparaison entre les 2 entités représentées en couleur (carré bleu et point rouge), nous pouvons remarquer une différence de rang étant donné que le nombre de villes (illustrées par les points noirs) comprises à l'intérieur des cercles diffère :

- si nous nous déplaçons de ● à ■, 2 villes s'inscrivent dans le cercle rouge tracé en tirets - avec comme centre ● - ;
- par contre, dans le sens inverse, de ■ à ●, nous pouvons en dénombrer 7 dans le cercle bleu en pointillés - avec comme centre ■ - .

4.2 Modèles de gravité et de radiation

En ce qui concerne les flux obtenus, nous considérons les définitions des modèles développées dans la partie théorique de ce mémoire :

$$\begin{aligned} T_{ij}^{R} &= T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \\ T_{ij}^{G} &= \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})}, \end{aligned}$$

où les paramètres α et β sont à déterminer et la fonction de dissuasion $f(r_{ij})$ est choisie pour correspondre au mieux aux données.

Dans le chapitre suivant, cependant, les paramètres α et β seront posés à 1 ou 2 et la fonction de dissuasion sera du type r_{ij}^{γ}, fonction consacrée pour prédire le modèle de la navette entre les comtés américains - développé dans l'article de [Simini *et al.*] - . Aussi, seul cet exposant devra être prédit pour le modèle de gravité grâce à des données réelles. A l'inverse, le modèle de radiation va se révéler libre en paramètres, ce qui va nous permettre d'exploiter plus d'informations et de nous faire directement une première idée du nombre de navetteurs entre 2 comtés. Notre recherche va se poursuivre sur base d'un jeu de données, qui devrait nous permettre de déterminer quel modèle se révèlera le meilleur. De fait, grâce aux données empiriques, il nous sera facile de comparer les flux prédits par les modèles de gravité et de radiation avec les flux observés.

CHAPITRE 5

Etat de New-York

Dans le but d'affiner notre analyse, les deux modèles développés précédemment - à savoir les modèles de gravité et de radiation - vont être expérimentés dans ce chapitre avec des données réelles, obtenues sur le site [Census Commuting] grâce au logiciel MATLAB. Ainsi, nous allons considérer tout particulièrement le nombre de navetteurs répertoriés entre les différents comtés de l'état de New-York.
Afin d'obtenir quelques éléments essentiels à l'élaboration des deux modèles et en tirer des conclusions plus précises, nous utiliserons par ailleurs une application développée par GOOGLE.
Après avoir tenté de dégager la relation existant entre le rang et la distance séparant 2 comtés, nous nous appliquerons à établir les modèles pour le flux réel obtenu.
Enfin, au terme de ce chapitre, nous devrions être en mesure de dégager le meilleur modèle pour ce jeu de données, en prenant en considération ses avantages et inconvénients.

5.1 Rappel théorique sur l'ajustement statistique

Lorsque nous représentons le nuage de points d'une série statistique à 2 variables (X et Y par exemple), force est de constater, généralement, que celui-ci ne correspond pas exactement au graphique d'une fonction. L'ajustement statistique consiste à rechercher la fonction dont le graphique s'approche au mieux de ce nuage. Dans la suite de ce chapitre, nous nous concentrerons à déterminer l'équation de la courbe d'ajustement s'approchant au mieux de ce nuage. La courbe peut se présenter sous la forme d'une droite, d'une parabole, d'une exponentielle, ... Dans le cadre présent, néanmoins, il s'agira tout simplement d'une fonction puissance.

Pour cette section, nous nous sommes basés sur le site [Matlab].

Chaque couple de points (x_i, y_i) - $i = 1, \ldots, N$ où N est l'effectif total - est donc proche d'une courbe du type :

$$\hat{y} = \beta x^{\alpha}$$

où $\hat{y}$ représente l'estimation de la variable Y via l'ajustement statistique.

Afin de déterminer les paramètres α et β, nous nous sommes aidés de l'échelle logarithmique :

$$\begin{aligned} \ln y &= \ln(\beta x^{\alpha}) \\ &= \ln \beta + \alpha \ln x. \end{aligned}$$

En changeant les variables suivantes - $u = \ln x$ et $v = \ln y$ - , nous pouvons facilement déterminer l'équation de la droite de régression de v en fonction de u à partir de la méthode des moindres carrés.

Néanmoins, pour pouvoir appliquer ce changement de variable, il est indispensable que les valeurs prises par les 2 variables soient toujours strictement positives. Dans le cas de nos données, même si elles ne seront jamais affectées d'un signe négatif, elles pourront toutefois s'annuler, dans le cas où, par exemple, aucun comté n'est repris dans le cercle ayant pour rayon la distance entre les 2 entités considérées. Nous avons décidé de ne pas considérer les données nulles lors de l'ajustement statistique, au risque de le fausser - vu qu'il y aura des données manquantes - .

De l'équation $\hat{v} = au + b$ obtenue via la régression linéaire, nous pouvons déduire l'équation de la fonction puissance

$$\hat{y} = \beta x^{\alpha}$$

où $\beta = e^b$ et $\alpha = a$.

Pour émettre un jugement quant à la qualité de l'ajustement linéaire, nous avons utilisé le coefficient de détermination R^2. De fait, celui-ci indique le pourcentage de la variance totale expliqué par l'ajustement linéaire et est défini par :

$$R^2 = 1 - \frac{\sum (v_i - \hat{v}_i)^2}{\sum (v_i - \bar{v})^2}$$

où $\bar{v}$ représente la moyenne de la variable v.

5.2 Quelques indications pratiques sur le modèle de radiation

Rappelons que N représente le nombre de villes considérées.

Via le modèle de radiation, le nombre de personnes se rendant d'un endroit i à un endroit j peut se définir sous la forme de la fonction suivante :

$$T_{ij} = T_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} \qquad (i, j = 1, \ldots, N)$$

où $T_i = \sum_{j \neq i} T_{ij}$ symbolise la quantité d'individus partant du point d'origine i. Seul ce paramètre est inconnu. Si l'on se réfère à l'article de [Simini *et al.*], ce nombre est proportionnel à la population d'origine :

$$T_i = n_i \frac{N_N}{N_T} \quad (i = 1, \ldots, N)$$

où N_N représente le nombre de navetteurs (obtenu en additionnant tous les flux de nos données), et N_T la population totale dans la zone considérée, à savoir, dans le cas présent, l'état de New-York.

Nous avons représenté graphiquement le nombre de navetteurs originaires de chaque comté (T_i), en fonction du nombre d'habitants par comté (n_i) :

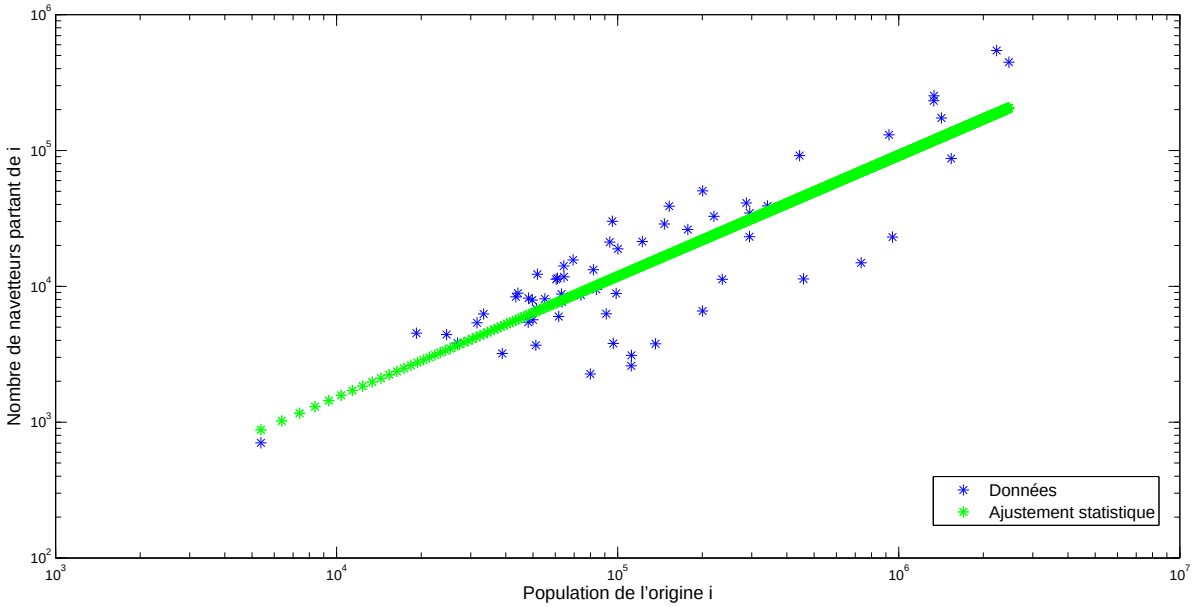


FIGURE 5.1 – Approximation du nombre de navetteurs par le nombre d'habitants

L'ajustement statistique prédit nous donne les informations suivantes :

$$\hat{T}_i = 0.418199\ n_i^{0.890378} \quad (i = 1, \ldots, N)$$

pour lequel un peu plus de 72 % des données sont déterminées par la droite de régression - chiffre obtenu via le coefficient de détermination - . Il s'agit donc d'un très bon ajustement statistique.
Au vu de la proximité de cette équation avec l'équation de la première bissectrice, nous avons choisi de remplacer T_i par $n_i \frac{N_N}{N_T}$.

5.3 Données nécessaires

Vu le nombre important d'états et de comtés constituant les Etats-Unis, nous avons décidé de limiter notre étude aux flux de voyageurs à l'intérieur de l'état de New-York [1] (représenté en rouge sur la carte).

FIGURE 5.2 – Carte des USA avec, en rouge, le comté de New-York

Situé au Nord-Est du pays, l'état de New-York constitue le troisième état le plus peuplé des USA, avec près de 19 millions d'habitants. Sa capitale, la ville d'Albany, se situe à l'est.

L'état de New-York est divisé en 62 comtés - $N = 62$ - (le comté se présentant comme une *forme de gouvernement local, une division territoriale plus petite qu'un état mais plus grande qu'une ville ou une municipalité, dans un état ou un territoire*) :

1. Toutes les informations sur cet état ainsi que les différentes cartes trouvent leur source sur [Wikipédia].

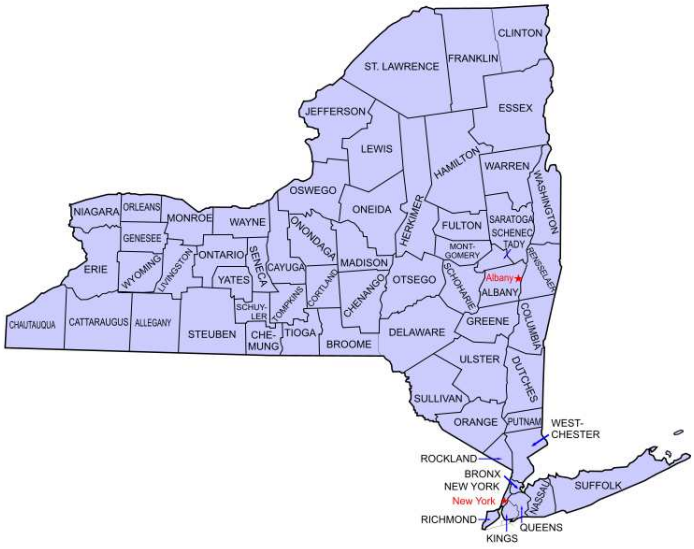


FIGURE 5.3 – 62 comtés de l'état de New-York

Grâce aux modèles de gravité et de radiation et aux données récoltées sur [Census Commuting], nous tenterons de déterminer le nombre de navetteurs résidant dans un comté new-yorkais et se rendant quotidiennement sur un lieu de travail implanté dans n'importe quel autre (ainsi, T_{ij} représente le nombre de personnes vivant en i et travaillant en j).

5.3.1 Matrice de distance

Dans une première étape, nous avons construit la matrice de distance nécessaire à l'élaboration de nos deux modèles. Pour ce faire, nous avons utilisé une application développée par GOOGLE : *"The Google distance Matrix API"* ([Google distance Matrix API]) nous fournissant la distance (en kilomètres) et le temps par trajet (en voiture, à pied, ou à vélo) entre chaque paire d'éléments d'une matrice constituée de villes d'origine et de villes de destination. Nous avons choisi comme moyen de transport la voiture.

Cependant, quelques anomalies sont apparues au niveau de certains comtés. Faute d'avoir trouvé le comté qui nous intéressait, l'application prenait dès lors une autre ville située à New-York même, faussant ainsi toute notre matrice. Afin de remédier à ce problème, nous avons décidé d'identifier plutôt ces comtés par leurs chefs-lieux respectifs[2] :

2. Chefs-lieux obtenus sur le site [Wikipédia 1].

Comté	Chef-lieu
Clinton	Plattsburgh
Essex	Elizabethtown
Madison	Wampsville
Nassau	Mineola

Après avoir attribué à chaque comté une adresse URL grâce à un programme conçu en MATLAB, nous avons pu créer la matrice de distance de dimension $N \times N = 62 \times 62$ via le logiciel EXCEL.

5.3.2 Nombre d'habitants par comté

Dans une seconde étape, nous avons établi un fichier reprenant le nombre d'habitants pour chaque comté new-yorkais. Ceci dans le but de déterminer par la suite le rang séparant chaque paire de comtés. Ces données ont été obtenues via le site [Wikipédia 1]. Ce site répertorie les habitants du 11^{e} état, répartis entre ses différents comtés, en l'an 2000. Les données obtenues sont reprises dans le code via la variable n - qui s'avère être un vecteur - . La population totale établie dans l'état de New-York, notée N_T, s'élève à 18 988 112 habitants.

5.4 Relation entre le rang et la distance

Après avoir représenté le nuage de points du rang (s) en fonction de la distance (d), nous avons cherché l'équation de la courbe d'ajustement s'en approchant au mieux.
Comme nous l'avons pointé plus haut, nous effectuerons une régression linéaire en passant à une échelle logarithmique. Pour rappel, nous avons exclu tous les couples dont une des composantes est nulle. Très naturellement, nos données ne seront jamais négatives, étant donné qu'il s'agit d'une distance et d'une population totale.

Pour établir la relation existant entre le rang et la distance, 4 hypothèses différentes seront traitées :

1. Aucune modification des données ;
2. Placement aléatoire des comtés ;
3. Nombre aléatoire d'habitants dans chacun des comtés ;
4. Nombre identique d'habitants dans chaque comté.

Nous travaillerons avec la définition correcte du rang. D'autre part, comme nous définissons le rang comme la population totale établie dans le cercle de rayon d_{ij} centré à l'origine i, nous pouvons l'approximer par l'équation suivante :

$$s_{ij} = dens * surf_{ij} \; \forall i, j = 1, \ldots, N$$

où *dens* correspond à la densité de population de l'état de New-York - à savoir le nombre moyen d'habitants par km^2, obtenu en divisant le nombre total d'habitants (soit 18 988 112 personnes) par la superficie de cet état (141 205 km^2) - et où $surf_{ij}$ ($\forall i, j = 1, \ldots, N$) détermine l'aire du cercle :

$$\begin{aligned} dens &= \frac{N_T}{141205} = 134.4720; \\ surf_{ij} &= \pi d_{ij}^2. \end{aligned}$$

5.4.1 Définition correcte du rang

Nous allons tenter de définir la relation existant entre la distance et le rang. Ce dernier est défini grâce à la formule suivante, $\forall i, j = 1, \ldots, N$:

$$s_{ij} = \sum_{\{k:0<d_{ik}\leq d_{ij} \text{ et } k\neq j\}} n_k.$$

Pour ce faire, nous avons tout simplement répertorié les groupements de villes comprises entre chaque paire de comtés, et additionné l'ensemble de leurs populations.

Premier cas

Dans un premier temps, nous considérons que les comtés sont placés au bon endroit, avec le nombre correct d'habitants. Nous n'avons donc modifié aucune matrice de distance ni de rang, pas plus que le vecteur de population.

Représentons le nuage de points du rang (s) en fonction de la distance (d).

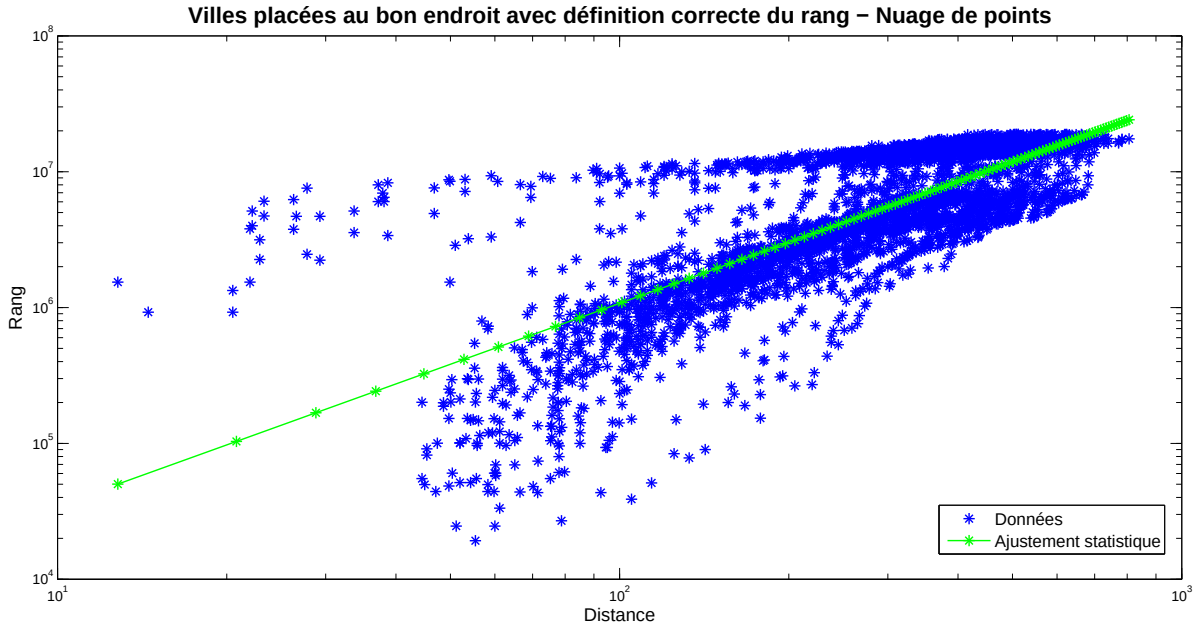


FIGURE 5.4 – Nuage de points des villes placées au bon endroit

Une première constatation saute aux yeux : une tendance croissante des données - plus la distance augmente, plus le rang augmente - . Ceci parait assez logique : plus les comtés sont espacés, plus le nombre de comtés situés dans le cercle risque d'être élevé ; d'où, un rang plus grand. Les données sont relativement éparpillées dans le plan logarithmique et n'y suivent pas vraiment une tendance linéaire - ceci s'explique essentiellement en fonction des comtés peu distants mais avec un rang élevé - .
Les résultats obtenus pour l'ajustement statistique corroborent ces constatations. Ainsi, nous obtenons :

$$\hat{s} = 1119.665632 \cdot d^{1.490921}.$$

Le coefficient de détermination s'élève à 0.58, ce qui se révèle assez moyen - la droite de régression permet de déterminer seulement 58 % de la distribution des points.

Si nous observons plus attentivement ce graphique, nous pouvons remarquer que, pour les grandes distances, le rang commence à stagner - ce phénomène sera encore plus marqué dans les cas qui suivent - . Un phénomène relativement facile à comprendre : si la distance séparant 2 comtés est très grande, le cercle établi avec cette distance comme rayon a toutes les chances de contenir une majorité des comtés (voire la totalité). Le rang s'élèvera donc approximativement - les populations d'origine et de destination n'étant pas comprises - à la population totale, N_T, soit 18 988 112 habitants.

Deuxième cas

Dans une seconde étape, les comtés ont été disposés au hasard à travers l'état de New-York. Pour ce faire, nous avons échangé aléatoirement la position des différentes villes. La matrice des rangs et le vecteur de la population s'en trouvent modifiés, avec une permutation des lignes et/ou des colonnes. La matrice des distances, quant à elle, ne subit aucun changement, étant donné que chaque comté a été remplacé par un autre. La seule différence notable consiste en une modification du nombre d'habitants par ville.

Représentons le rang réel (s) en fonction de la distance (d) :

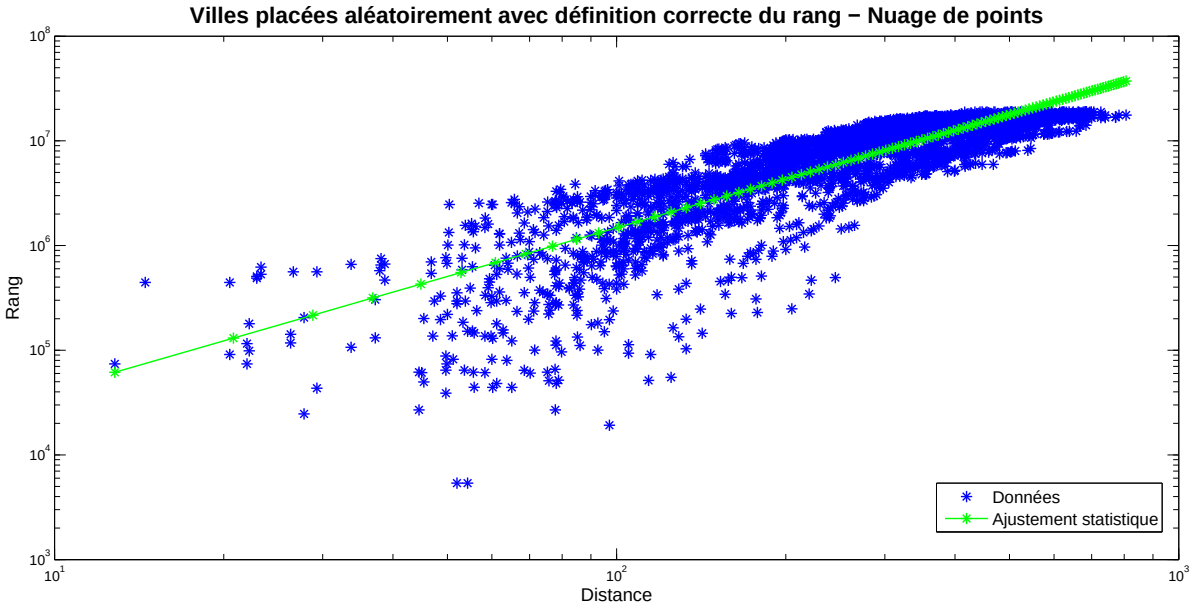


FIGURE 5.5 – Nuage de points des villes placées aléatoirement

De nouveau, nous pouvons remarquer qu'une plus grande distance entre 2 comtés implique un plus grand rang. De fait, comme expliqué précédemment, plus les comtés seront éloignés, plus il y aura de comtés dans l'intervalle, et donc d'habitants.
Le nuage de points apparait plus dispersé tant que l'écart entre les paires de comtés ne dépasse pas 200 km. Au-delà de ce seuil, nous constatons une configuration plus serrée.
Le rang prédit à partir de la distance peut se définir de la manière suivante :

$$\hat{s} = 2380.253405 \cdot d^{1.439255}.$$

Dans le cas présent, l'ajustement statistique permet d'augmenter l'exactitude des points prédits : 63.7 % des données sont expliquées par la droite de régression.

Troisième cas

Dans une troisième étape, nous avons décidé de replacer les comtés au bon endroit, avec, toutefois, cette fois une répartition aléatoire du nombre total d'habitants N_T entre les différents comtés. Ceci implique une modification du vecteur n reprenant la population par entité, ainsi que le rang entre chaque paire.

Le nuage de points du rang réel (s) en fonction de la distance (d) peut être représenté de la manière suivante :

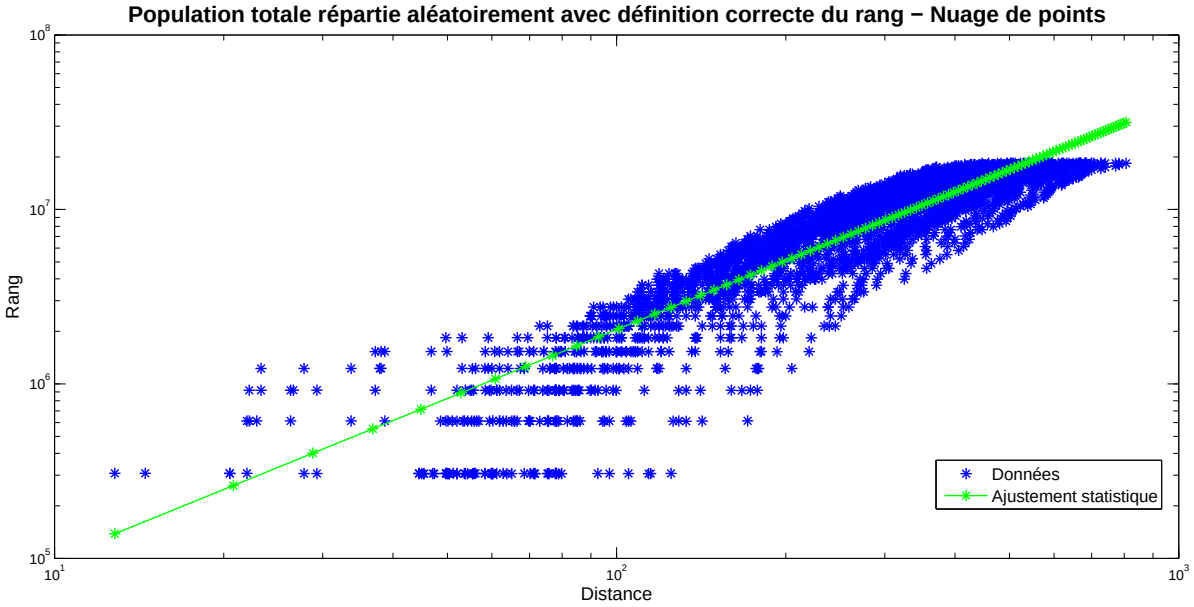


FIGURE 5.6 – Nuage de points des villes avec population totale répartie aléatoirement

Une tendance croissante se dégage nettement de ce graphique : plus la distance augmente, plus le rang augmente, avec, toutefois, un palier pour les distances maximales. De fait, nous pouvons remarquer que le rang commence à stagner, phénomène déjà expliqué précédemment. Contrairement aux deux cas précédents, les données semblent moins éparpillées, si ce n'est dans le cas de paires de villes peu distantes.

L'ajustement statistique nous permet de définir le rang par rapport à la distance via la fonction suivante :

$$\hat{s} = 4881.156519 \cdot d^{1.311413}.$$

Même si nous constatons une faible variation du paramètre α, l'exposant de la fonction, le coefficient de détermination se révèle nettement meilleur : 0.83. Désormais, la droite de régression permet de déterminer 83 % de la distribution des points. L'ajustement statistique se révèle dès lors nettement plus efficace.

Quatrième cas

Enfin, dans une dernière hypothèse de travail, nous avons supposé les comtés tout à fait homogènes, constitués d'un même nombre d'habitants. Nous avons ainsi placé $\frac{N_T}{62} = 306259.871$ personnes par comté. De nouveau, seule notre matrice de rang subit une modification.

Représentons le nuage de points du rang (s) en fonction de la distance (d) :

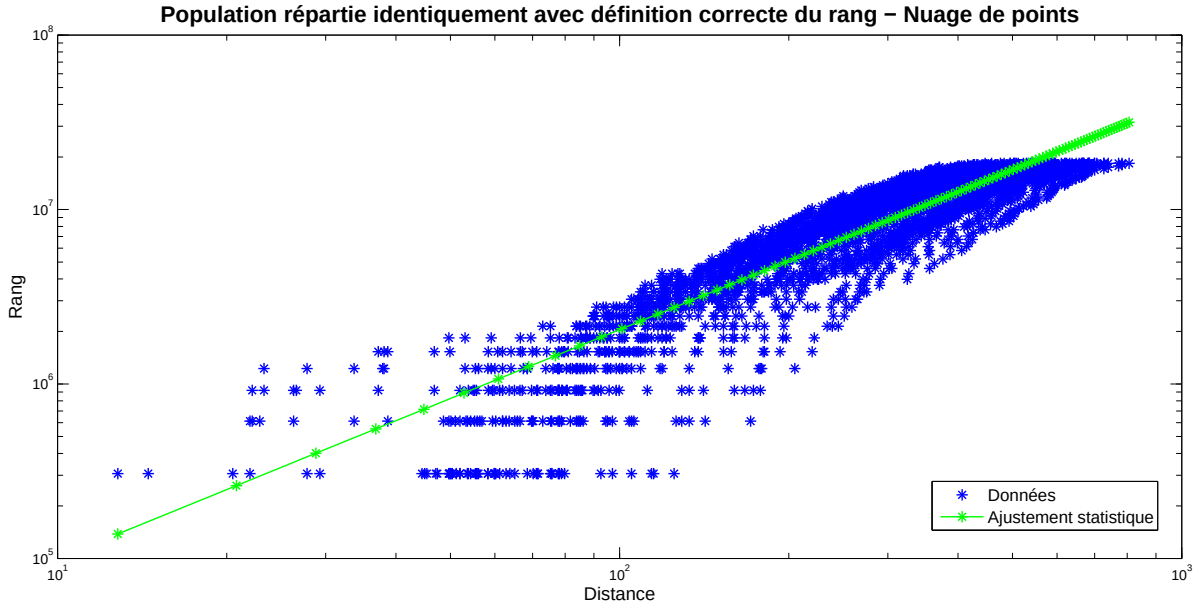


FIGURE 5.7 – Nuage de points des villes avec population totale répartie identiquement

Nous remarquons un nuage de points similaire au cas précédent. La configuration générale présente toujours la même tendance.
L'équation de l'ajustement statistique se révèle légèrement différente :

$$\hat{s} = 4881.217242 \cdot d^{1.311405}.$$

Le coefficient de détermination restant identique, la droite de régression détermine toujours 83 % de la distribution des points.

5.4.2 Aire du cercle

Considérons à présent que nous pouvons approximer le rang par l'aire du cercle. De manière assez évidente, cette définition sera identique quel que soit le cas de figure choisi :

$$s_{ij} = dens \cdot \pi \cdot d_{ij}^2.$$

De fait, les différents éléments intervenant dans sa définition n'ont jamais été modifiés, comme nous l'avons expliqué précédemment : nous n'avons jamais transformé la matrice des distances quel que soit le cas considéré.
Aussi, représentons l'unique nuage de points du rang (s) en fonction de la distance (d) :

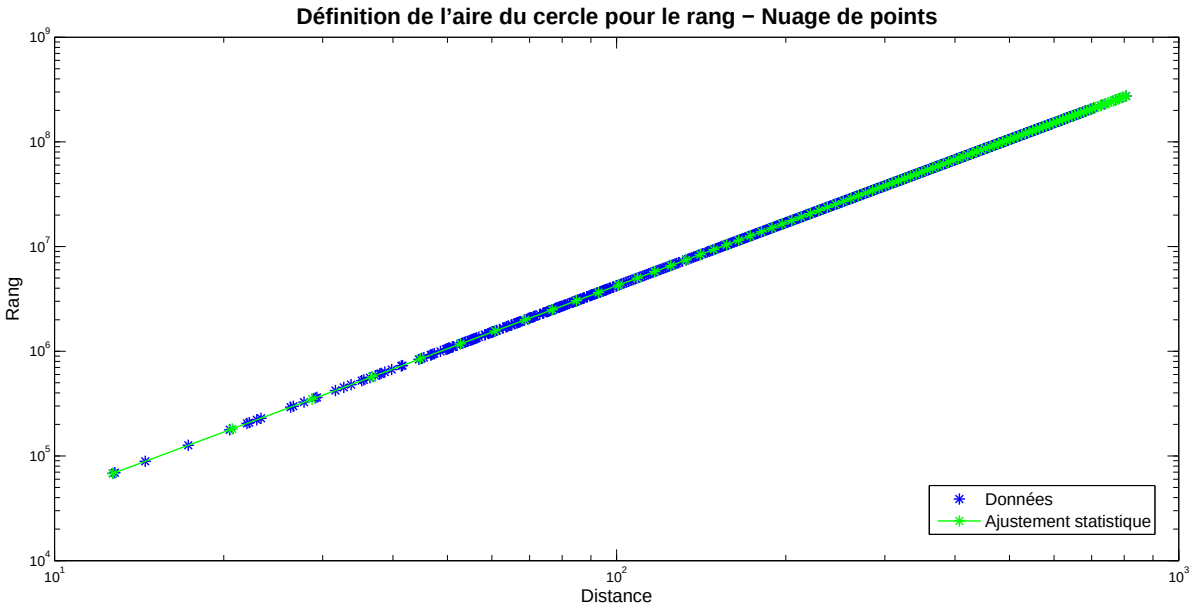


FIGURE 5.8 – Nuage de points des villes

L'ajustement statistique, décrit précédemment, trouve la fonction puissance suivante :

$$\hat{s} = 422.456097 \cdot d^2.$$

Ce résultat semble logique, dans la mesure où nous avons défini le rang comme étant approximativement le carré de la distance.
De plus, le coefficient de détermination s'élève à 1 : les valeurs observées sont identiques à celles prédites par l'ajustement statistique. De fait, nous pouvons constater sur la figure 5.8 que les données réelles sont situées sur une même droite.

Cependant, dans la réalité, il s'avère très peu probable - voire impossible - que la population soit répartie de manière totalement homogène à travers l'état de New-York. De plus, le rang dépasse largement la population établie dans l'état de New-York ($\approx 10^7$).

5.5 Données réelles

Les données récoltées sur le site [Census Commuting] correspondent au flux de navetteurs circulant régulièrement entre chaque paire de comtés des différents états américains (et même avec des villes étrangères) durant l'année 2000. Les fichiers reprennent les flux de 166 248 paires d'origines-destinations entre 3141 entités. Il est clair que traiter l'entièreté des états s'avèrerait une tâche beaucoup trop vaste. Nous nous concentrerons dès lors essentiellement sur l'état de New-York, dont les 62 comtés joueront à la fois le rôle de points d'origine et de destination.

Dans cette partie, nous nous attacherons à représenter les différents flux réels en fonction de la distance et du rang. En ce qui concerne le rang, nous traiterons à la fois la définition correcte, mais également la définition via l'aire d'un cercle et celle obtenue grâce à l'ajustement statistique.
Nous obtiendrons ainsi 4 nuages de points, pour lesquels un ajustement statistique sera de nouveau effectué.

5.5.1 Flux réel en fonction de la distance

Représentons le flux réel (T_{obs}) en fonction de la distance (d) séparant chaque paire de comtés :

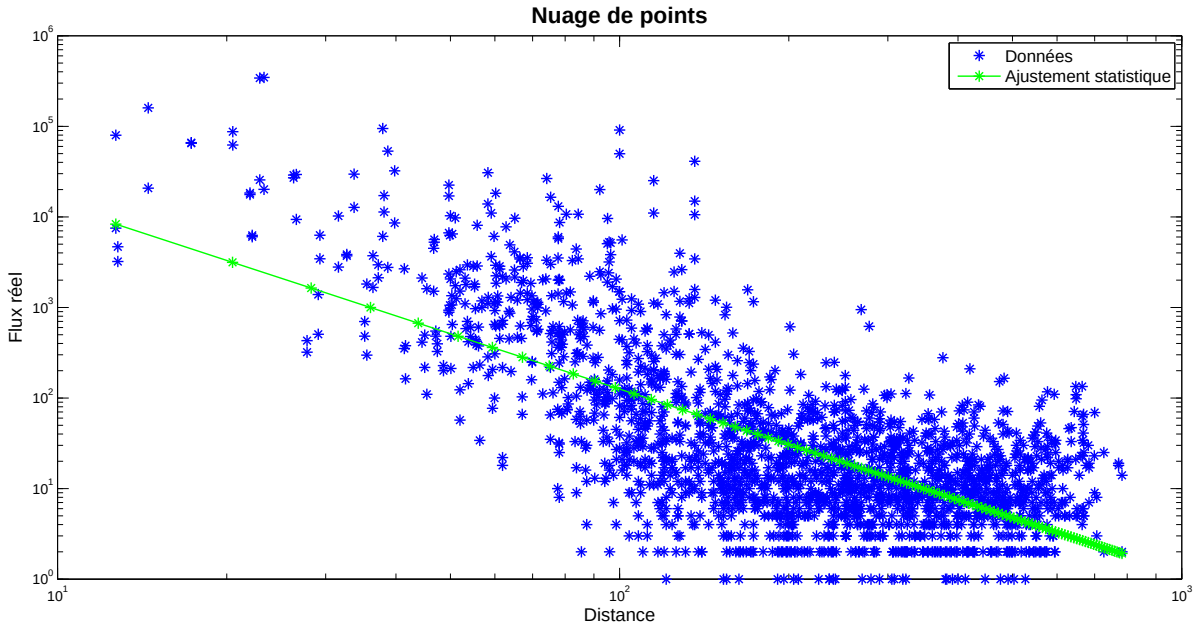


FIGURE 5.9 – Flux réel en fonction de la distance

De ce nuage de points se dégage une tendance négative, dans le sens où plus la distance entre une paire de comtés augmente, moins de navetteurs effectueront ce déplacement. Nous avons déjà fait cette constatation dans la partie théorique de ce travail : en effet, les individus vont plus facilement travailler à proximité de chez eux.

Grâce à l'ajustement statistique, le flux estimé peut être défini par l'équation suivante :

$$\hat{T}_{\text{obs}_D} = 1452172.949531 \cdot d^{-2.031753}.$$

Le coefficient de détermination s'élève à 0.48. L'ajustement statistique explique seulement 48 % de la dispersion des points. En fait, un examen de la figure nous permet de constater que la droite est influencée par la forte proportion de comtés fort éloignés. Cela entraîne que la majorité des flux entre des comtés peu distants se révèlent sous-estimés par cet ajustement statistique.

5.5.2 Flux réel en fonction du rang

Dans cette partie, nous nous attacherons à étudier le flux réel en fonction des différentes notions du rang (s). Premièrement, nous travaillerons avec la définition correcte du rang, à savoir la population totale établie dans le cercle ayant pour rayon la distance entre chaque paire de comtés. Deuxièmement, nous considérerons que nous pouvons approximer le rang par l'aire du cercle. Enfin, nous utiliserons la définition du rang obtenue par l'ajustement statistique dans la section précédente.

Définition correcte du rang

Pour représenter le flux réel (T_{obs}) en fonction du rang (s), nous procéderons, dans une première étape, avec la définition correcte, à savoir la population totale établie dans le cercle :

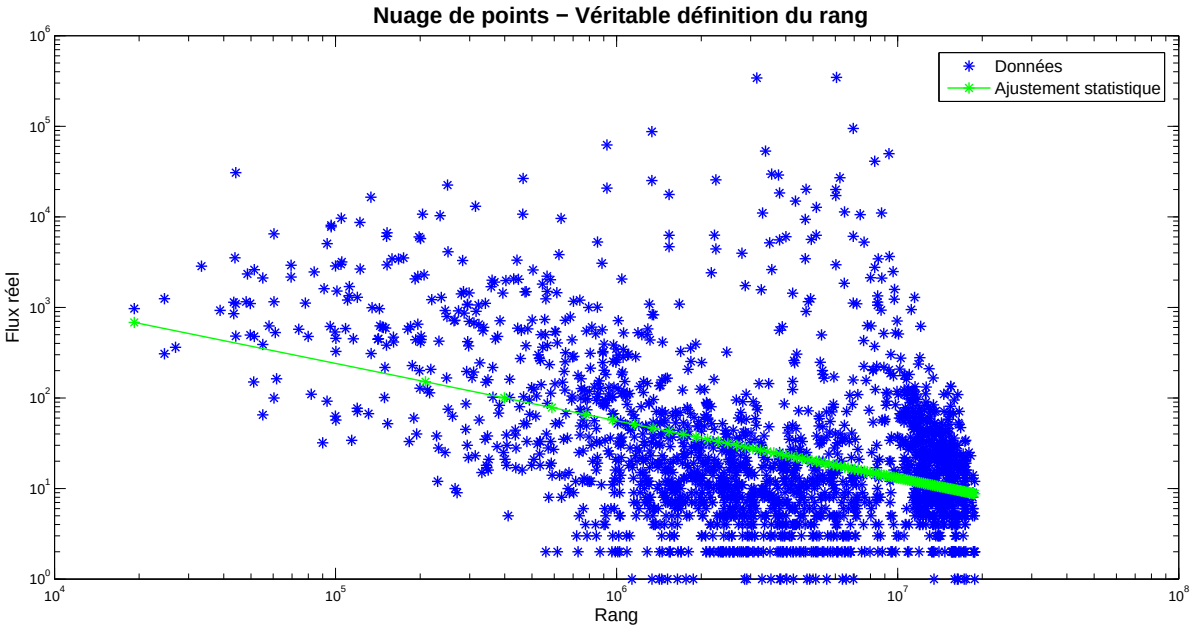


FIGURE 5.10 – Flux réel en fonction de la distance - Vraie définition du rang

La figure 5.10 met en évidence une répartition moins linéaire des données : de fait, nous pouvons remarquer quelques valeurs aberrantes - certaines paires de comtés avec un rang élevé (entre 10^6 et 10^7) présentent un nombre considérable de navetteurs (entre 10^4 et 10^6) - . Ceci pourrait s'expliquer par l'attrait de certaines villes en matière d'emploi.
De plus, de nombreuses villes possèdent un rang élevé, ce qui influence ainsi la droite d'ajustement. Si nous observons attentivement ce graphique, nous pouvons retrouver un phénomène relevé précédemment : le rang commence à stagner dans la partie droite et ne dépassera pas la population totale, N_T, soit 18 988 112 habitants.
Alors qu'un rang inférieur à 10^5 habitants implique un grand nombre de voyageurs, nous ne pouvons pas vraiment tirer de conclusions quant aux paires de villes avec un rang supérieur à ce seuil. De fait, si nous considérons par exemple une population totale établie dans le cercle comprise entre 10^6 et 10^7, le nombre de voyageurs peut aller de 10^0 à 10^6, avec, néanmoins, une plus forte tendance pour les faibles flux.

Le flux prédit par l'ajustement statistique est déterminé par l'équation suivante :

$$\hat{T}_{\text{obs}_S} = 350770.500758 \cdot s^{-0.632795}.$$

Le coefficient de détermination confirme notre observation : 18.7 % seulement des données sont expliquées par l'ajustement linéaire. L'ajustement n'est donc pas fiable.

Définition du rang via l'aire du cercle

Représentons à présent le flux réel (T_{obs}) en fonction du rang (s) via la définition de l'aire du cercle :

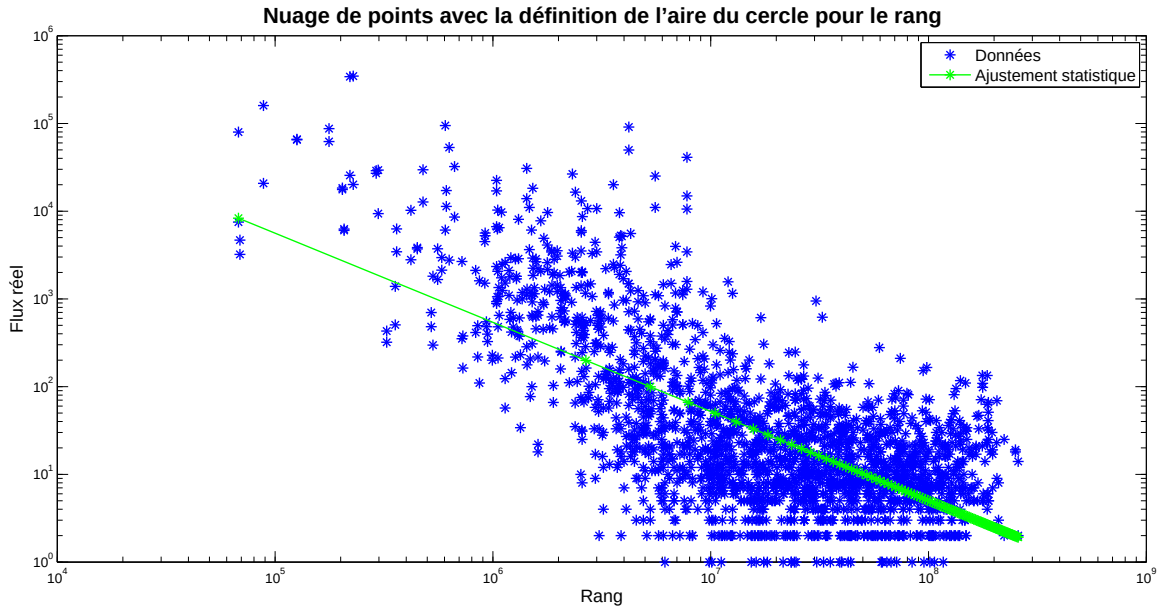


FIGURE 5.11 – Flux réel en fonction du rang - Aire du cercle

Contrairement au cas précédent, la figure ci-dessus met nettement plus en évidence une répartition des données autour d'une ligne droite de pente négative : plus la population établie entre 2 comtés augmente, moins de navetteurs effectuent le trajet. Mais comment expliquer cette répartition plus linéaire ? Serait-ce dû au fait que le rang est plus structuré, et présente une tendance quadratique ? Nous savons, par définition, que plus la distance augmente, plus le rang croît ($s \approx d^2$). De plus, le graphique 5.9 nous a amenés au constat que les individus se déplacent de préférence à proximité de chez eux. Il semble dès lors évident que le flux réel, ici, suit une tendance similaire à ce graphique 5.9.
Cependant, il nous faut de nouveau relever que de nombreuses paires de comtés ont un rang élevé (entre 10^7 et 10^8). L'ajustement statistique va donc s'en trouver influencé. De même, en considérant l'aire du cercle, nous réalisons que le rang dépasse largement la population totale de l'état de New-York, ce qui parait assez absurde.

Le flux prédit s'écrit :

$$\hat{T}_{\text{obs}_S} = 672878966.5537 \cdot s^{-1.015876}.$$

Cet ajustement se révèle plus fiable, dans la mesure où il explique 47.9 % de la dispersion des données.

Définition du rang via l'ajustement statistique

Représentons enfin le flux réel (T_{obs}) en fonction du rang via la définition obtenue par l'ajustement statistique (soit, pour rappel : $s = 1119.665632 \cdot d^{1.490921}$) :

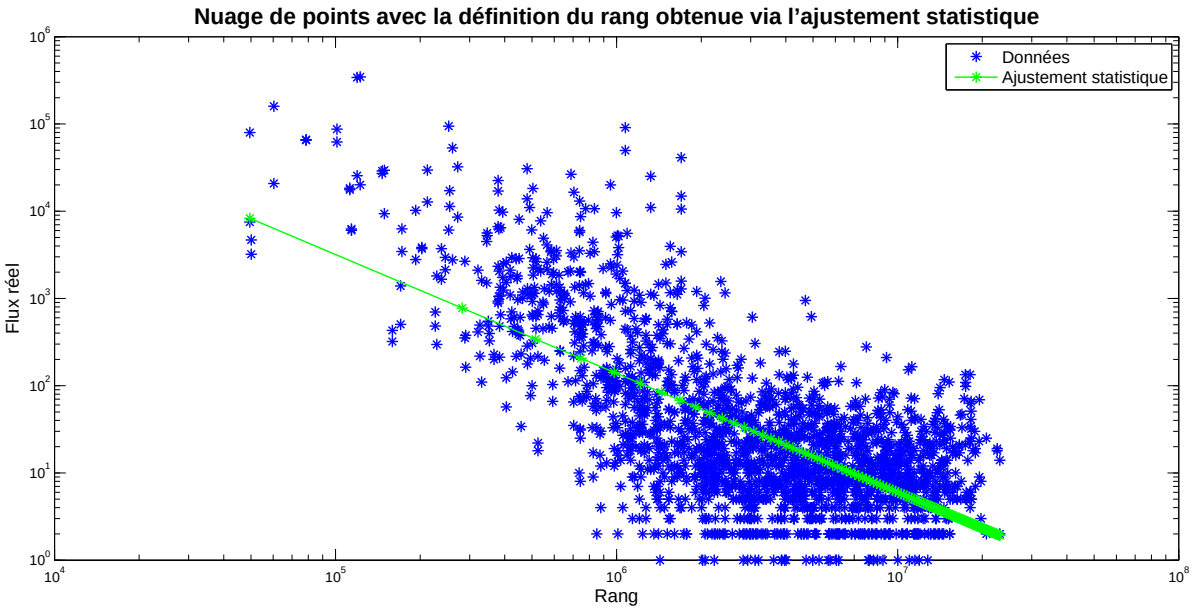


FIGURE 5.12 – Flux réel en fonction du rang - Ajustement statistique

Une tendance linéaire négative se dégage clairement de la figure 5.12 : plus le rang augmente, plus le flux diminue. Comme dans le cas précédent, ceci pourrait s'expliquer par le fait que le rang est défini à partir de la distance ($s \approx d^{1.5}$), et que plus les villes sont distantes, plus le flux décroît (graphique 5.9). Néanmoins, nous pouvons à nouveau relever une forte concentration de paires de comtés avec un rang élevé, dans le coin inférieur droit. La droite de régression s'en trouve inévitablement influencée.
Alors que, dans le cas précédent, la population totale établie entre 2 comtés dépassait largement la population totale de l'état de New-York, nous pouvons noter ici un seuil maximal d'environ 10^7 grâce à l'ajustement statistique (*cf.* le graphique 5.4 qui prend une valeur maximale).

Le flux prédit peut être défini de la manière suivante :

$$\hat{T}_{\text{obs}_S} = 20756984407.114010 \cdot s^{-1.362750}.$$

La performance de l'ajustement est similaire au cas précédent, étant donné que le coefficient de détermination est pareil.

5.6 Comparaison des deux modèles

Etant donné la différence d'échelle entre le rang et la distance, nous ne pouvons comparer les modèles de radiation et de gravité en nous basant sur les méthodes précédentes. Aussi, nous allons procéder autrement en représentant graphiquement le flux réel (T_{obs}) en fonction du flux prédit (T_{calc}). Plus les couples de points s'aligneront sur la droite $T_{\text{calc}} = T_{\text{obs}}$, plus notre modèle sera fiable. Le meilleur des cas se traduisant mathématiquement par une équivalence entre chaque flux réel et chaque flux calculé.

Afin de comparer les deux modèles, nous allons calculer la variabilité des données autour de la première bissectrice $y = x$ de la manière suivante :

$$V = 1 - \frac{\sum (T_{\text{calc}_{ij}} - \hat{T}_{ij})^2}{\sum (T_{\text{calc}_{ij}} - \bar{T}_{\text{calc}})^2}$$

où $\hat{T}_{ij} = T_{\text{obs}_{ij}}$ - comme nous comparons par rapport à la première bissectrice - et $\bar{T}_{\text{calc}}$ représente la moyenne du flux calculé.
Si le modèle prédit exactement le flux observé, alors $T_{\text{calc}_{ij}} = T_{\text{obs}_{ij}}$, $\forall i, j = 1, \ldots, N$, impliquant une variabilité maximale, avec l'obtention d'une droite correspondant à la première bissectrice du repère orthonormé.
Cette variabilité, comme le coefficient de détermination, peut prendre des valeurs comprises entre 0 et 1.

5.6.1 Modèle de gravité

Nous avons d'abord calculé le flux obtenu avec le modèle de gravité. Pour ce faire, nous avons utilisé la définition du nombre de navetteurs établie précédemment via l'ajustement statistique :

$$\hat{T}_{\text{obs}_D} = 1452172.949531 \cdot d^{-2.031753}.$$

Représentons le flux obtenu en fonction du flux observé :

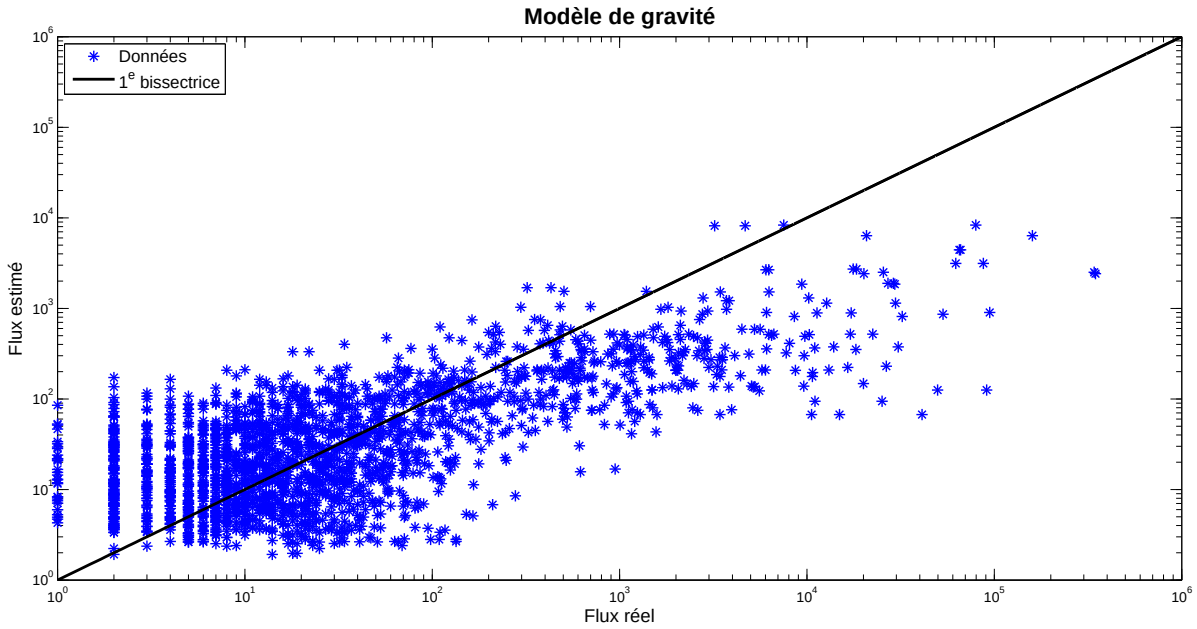


FIGURE 5.13 – Modèle de gravité

Nous distinguons ici une tendance linéaire croissante du nuage de points. Néanmoins, le modèle de gravité sous-estime les grands flux (à partir de 10^3) et surestime les plus petits. Entre les deux, les points se répartissent de part et d'autre de la droite d'équation $y = x$.

Autre observation : le groupe des données s'avère assez compact. De fait, si nous effectuons un ajustement statistique sur ce nuage de points, près de la moitié des données (48 %) peuvent être expliquées par la régression linéaire. Cette droite, cependant, se tient éloignée de la première bissectrice. De plus, le flux maximal pour le modèle de gravité équivaut à 10^4, contrairement au réel 10^6. Si nous totalisons les différents flux obtenus, nous constatons que le modèle de gravité via cet ajustement statistique prévoit seulement 289287 navetteurs, soit quasi à peine un trentième du nombre réel.

Cette mauvaise estimation pourrait s'expliquer par le fait que nous n'avons jamais pris en compte le nombre d'habitants propre à chaque comté et que seule la distance importe. Comment améliorer dès lors le modèle ? Nous tenterons d'y remédier par une procédure spéciale décrite dans la prochaine section (*Amélioration des modèles*).

Si nous calculons la variabilité V par rapport à la première bissectrice, nous obtenons un nombre négatif (-0.089348) alors qu'elle devrait être comprise entre 0 et 1. A titre d'explication, il peut être raisonnable d'avancer que la première bissectrice ne prend en compte qu'une partie des données et n'ajuste dès lors pas du tout le nuage de points.

5.6.2 Modèle de radiation

Pour le modèle de radiation, nous pouvons procéder de 2 manières différentes pour étudier le flux obtenu :

1. via la définition de [Simini *et al.*] ;
2. via le mauvais ajustement statistique trouvé dans la section précédente.

Définition de Simini

Rappelons la définition de Simini *et al.* :

$$T_{ij}^R = \frac{N_N}{N_T} \frac{n_i^2 n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})}$$

où N_N représente le nombre total de navetteurs et N_T le nombre d'habitants dans l'état de New-York.

En suivant la définition de Simini, nous obtenons le nuage de points suivant du flux obtenu en fonction du flux réel :

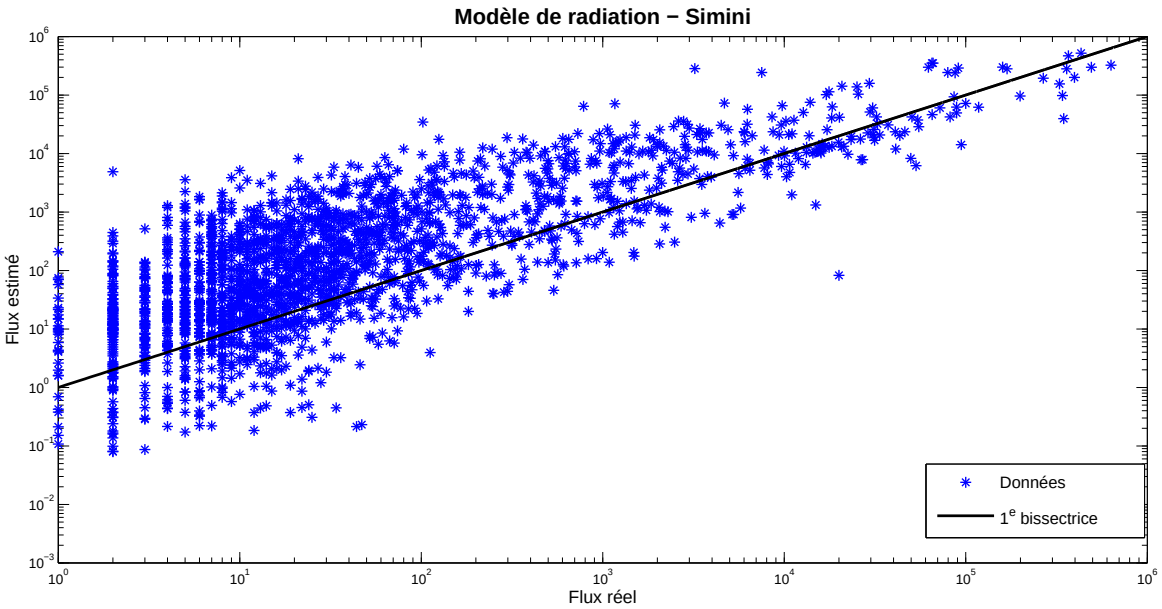


FIGURE 5.14 – Modèle de radiation - Simini

De prime abord, ce modèle semble meilleur que le précédent, étant donné que les données "suivent" plus la droite. Néanmoins, il aurait tendance à sous-estimer le nombre de navetteurs entre les comtés pour des flux supérieurs à 10^4, et à le surestimer pour des flux inférieurs à 10^4. De surcroît, contrairement au cas précédent, la configuration du nuage

semble plus dispersée.

Après avoir effectué un ajustement statistique sur ce nuage de points, la droite de régression se révèle proche de la première bissectrice, avec un coefficient de détermination de 0.58. En nous penchant sur les flux de navetteurs entre chaque comté, le modèle de radiation prévoit 11 447 868 personnes se déplaçant entre les différentes paires de comtés. Même si ce chiffre s'avère plus élevé que le nombre total réel de navetteurs, il est inférieur au nombre d'habitants de l'état de New-York, ce qui se révèle être un point positif.

La variabilité, quant à elle, s'avère meilleure qu'au modèle précédent : de fait, elle s'élève à 0.32. La première bissectrice semble mieux traverser le nuage de points, avec, toutefois, une plus forte concentration autour des "petits" flux.

Définition via l'ajustement statistique

Rappelez-vous, grâce à l'ajustement statistique, nous avons défini le nombre de navetteurs suivant le rang par :

$$\hat{T}_{\text{obs}_S} = 350770.500758 \cdot s^{-0.632795}.$$

Nous obtenons dès lors le nuage de points suivant :

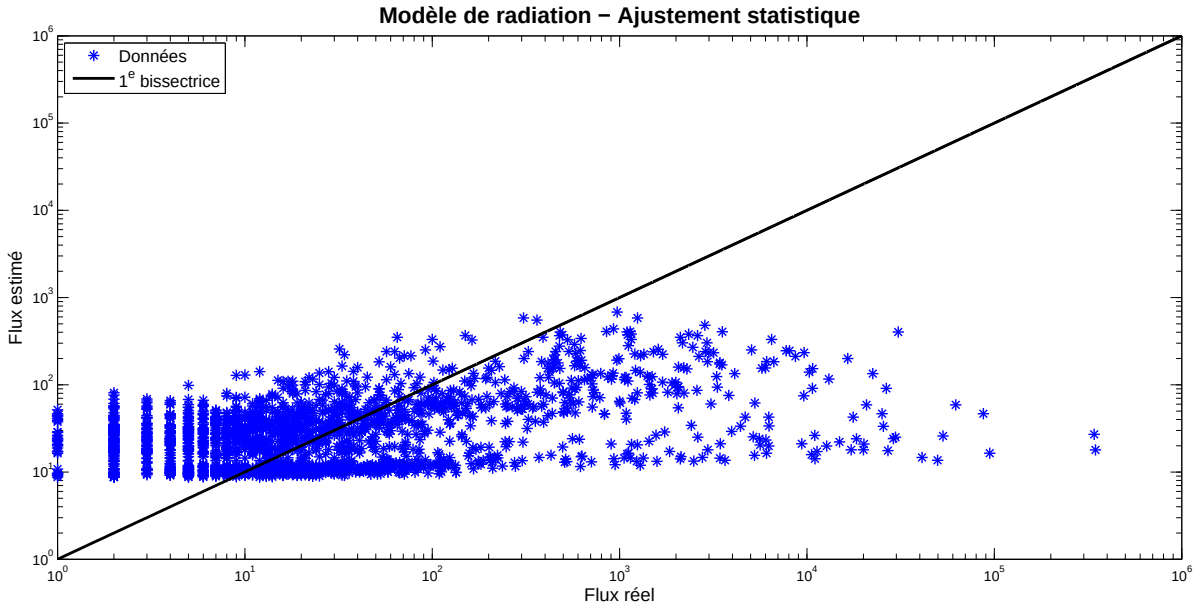


FIGURE 5.15 – Modèle de radiation - Ajustement statistique

Nous pouvons remarquer que les estimations, dans ce cas-ci, sont extrêmement mauvaises.

Alors que les faibles flux sont surestimés par le modèle, les plus élevés, à l'inverse, se trouvent sous-estimés. De plus, si nous établissons un ajustement statistique sur ce nuage de points, seuls 18.8 % des points sont prédits par ce modèle.

La variabilité s'avère également médiocre : elle prend une valeur de -3.34, une valeur peu surprenante au vu de l'allure du graphique. La première bissectrice n'ajuste aucunement les points du nuage.

Si nous totalisons les différents flux obtenus, seules 116 817 personnes sont considérées comme voyageant d'un comté à un autre, soit même pas 1% de la population totale de New-York et même pas un septantième du nombre réel de navetteurs.
Comme nous l'avons pointé pour le modèle de gravité, ce modèle de radiation ne prend pas en compte le nombre d'habitants propre à chaque comté. Nous renvoyons dès lors à nouveau à la partie suivante *Amélioration des modèles*.

5.6.3 Amélioration des modèles

Nous l'avons relevé précédemment, le modèle de gravité et un des modèles de radiation, prédits à partir d'un ajustement statistique, ne prennent pas en compte le nombre d'habitants par comté.
De plus, les deux matrices de flux ont des diagonales nulles - signifiant ainsi l'absence totale de navetteurs au sein d'une même ville - . En effet, dans le cas d'une entité jouant à la fois le rôle d'origine et de destination, le rang et la distance sont nuls ; dès lors, il est impossible de prédire le flux ; par conséquent, nous le mettons à 0. Dans ce contexte, nous avons décidé de développer un programme susceptible de donner un meilleur ajustement statistique, dans le sens où le coefficient angulaire de la droite de régression se rapprocherait de 1 - en effet, l'objectif est d'obtenir une droite, ce qui impliquerait une analogie entre les données réelles et les données prédites - . Pour ce faire, nous avons élaboré trois définitions différentes des modèles dépendant d'un paramètre α - que nous allons faire varier - .

Modèle de gravité

Pour prédire le flux moyen du modèle de gravité, nous avons décidé d'utiliser la définition suivante :

$$T_{ij}^{G} = \frac{n_i n_j}{d_{ij}^{\alpha}}.$$

Dans un premier temps, intéressons-nous au coefficient angulaire obtenu via l'ajustement statistique en fonction du paramètre α choisi :

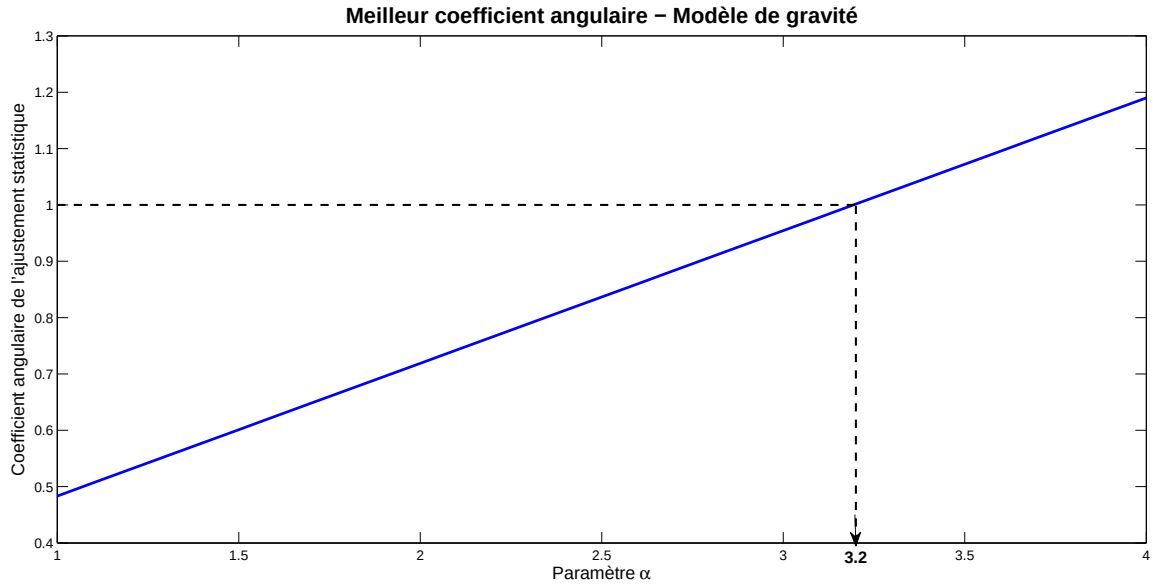


FIGURE 5.16 – Coefficient angulaire dans le modèle de gravité

De prime abord, nous remarquons que la pente a tendance à augmenter lorsque le paramètre α croît. Comme nous pouvons le constater à la lecture de la figure ci-dessus, le paramètre α offrant le meilleur ajustement statistique s'élève à 3.2.

Représentons à présent le flux obtenu avec ce paramètre α en fonction du flux réel :

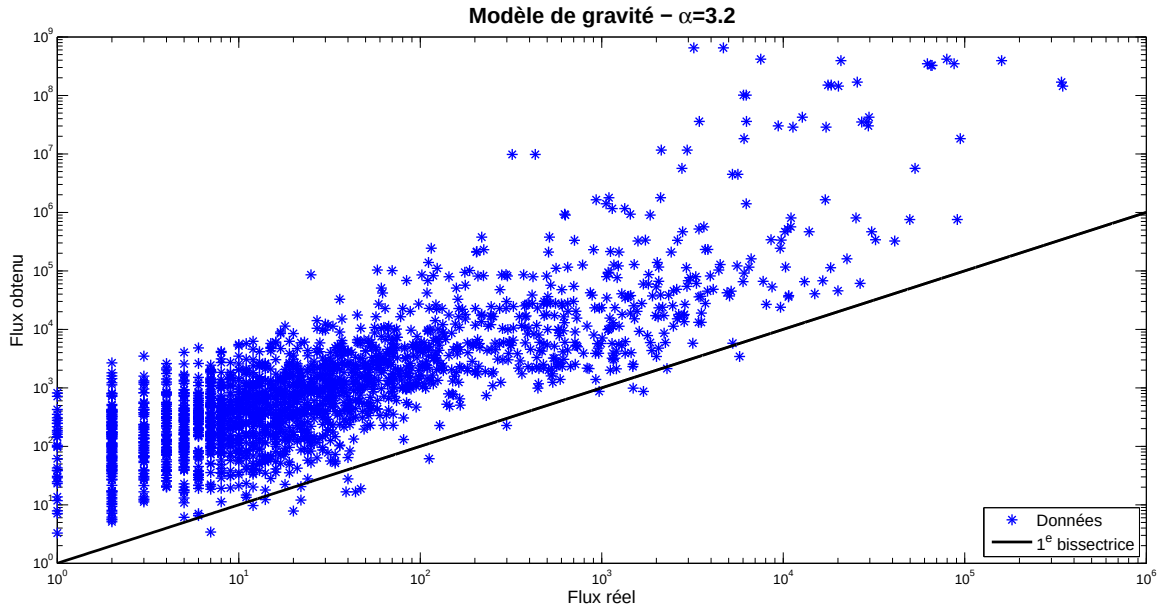


FIGURE 5.17 – Modèle de gravité - $\alpha = 3.2$

Ceci nous conduit d'emblée à noter une amélioration du modèle de gravité. Les données suivent plutôt une tendance linéaire. Néanmoins, ce modèle aurait tendance à surestimer de manière générale le flux réel : de fait, seuls quelques points se situent sous la première bissectrice. De plus, il ressort du graphe que le nombre prédit de navetteurs ralliant 2 comtés peut grimper jusqu'à 10^9, soit un écart de $5.8788 \cdot 10^9$ de personnes par rapport aux données réelles.

En d'autres termes, 5 886 790 931 navetteurs seraient prévus dans l'état de New-York, autrement dit, la quasi totalité des terriens - à un milliard près - se déplaceraient dans cet état. Nous rejoignons ici la remarque établie précédemment dans le chapitre plus théorique de cette recherche : plus la population d'origine s'accroît, plus le nombre de navetteurs augmente. Aussi, il serait sans doute utile de modifier le numérateur du modèle de gravité en intégrant 2 paramètres supplémentaires (comme mettre des puissances aux différentes populations).

Si nous nous intéressons à la variabilité V par rapport à la première bissectrice, celle-ci s'avère négative (-1.121085). Un résultat tout à fait logique étant donné que quasi l'entièreté des points se situent au-dessus de cette droite. Néanmoins, le coefficient de détermination par rapport à l'ajustement statistique établi permet de prédire la variation de 66.6 % des données. Le nuage de points, quant à lui, semble également moins dispersé.

Le modèle de gravité prévoit une surestimation du flux réel, qui lui est toutefois proportionnelle. Si nous normalisons par une constante e^c (où c symbolise l'ordonnée à l'origine de la régression linéaire précédente), nous obtenons le nuage de points suivant :

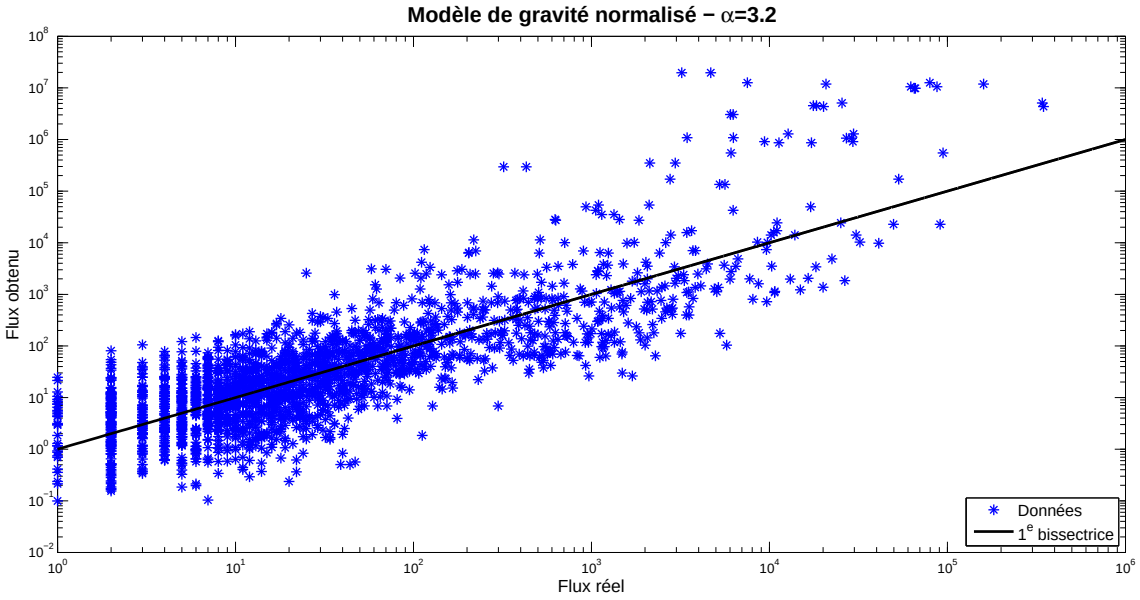


FIGURE 5.18 – Modèle de gravité normalisé

Grâce à cette normalisation, le modèle de gravité s'avère être plus correct : les données sont peu éparpillées autour de la première bissectrice, et la variabilité quant à elle s'élève à 0.665570. Ainsi, 66.6 % des données sont expliquées par la première bissectrice, ce qui constitue la meilleure estimation jusqu'à présent.
D'autre part, le nombre prédit de navetteurs peut aller jusqu'à 10^8, contrairement aux 10^6 personnes maximales. Ce qui implique aussi, malheureusement, un plus grand nombre prédit de navetteurs, dépassant largement la quantité réelle de résidents dans l'état de New-York.
L'ajout de paramètres pourrait peut-être nous aider à obtenir un meilleur résultat.

Deuxième modèle de gravité

Pour prédire le flux du modèle de gravité, nous avons également utilisé la définition suivante :

$$T_{ij} = \frac{n_i^2 n_j}{d_{ij}^{\alpha}},$$

ceci dans un but d'obtenir une définition similaire à celle utilisée par [Simini *et al.*].
Dans un premier temps, intéressons-nous au coefficient angulaire obtenu via l'ajustement statistique en fonction du paramètre α :

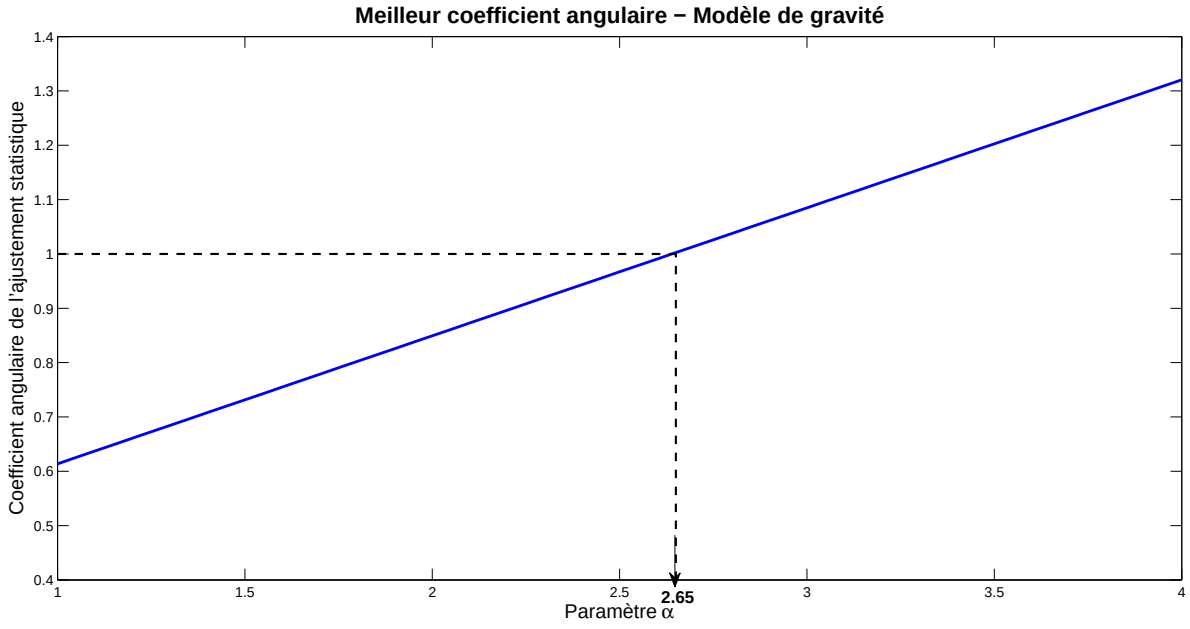


FIGURE 5.19 – Coefficient angulaire dans le deuxième modèle de gravité

Nous pouvons à nouveau constater une augmentation de la pente lorsque le paramètre α croît. De ce graphique, nous pouvons déduire que le paramètre α offrant le meilleur ajustement statistique s'élève à 2.65.

Représentons à présent le flux obtenu avec ce paramètre α en fonction du flux réel :

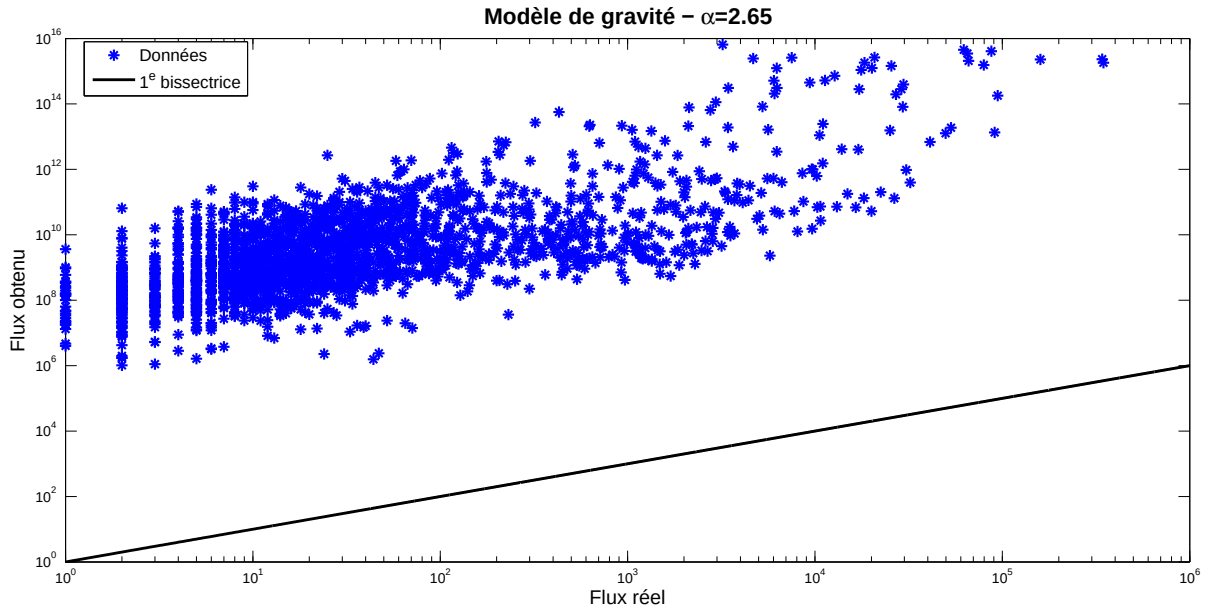


FIGURE 5.20 – Modèle de gravité - $\alpha = 2.65$

Malgré une tendance linéaire, ce modèle de gravité surestime de manière significative le flux réel : une minorité de flux observés (tous supérieurs à 10^6) prennent des valeurs réelles. Le nombre prédit de navetteurs entre deux comtés peut s'élever jusqu'à 10^{16}, ce qui paraît tout simplement impossible.

La médiocrité de ce modèle pourrait s'expliquer par le fait que le numérateur $(n_i^2 n_j)$ tend à prendre des valeurs démesurées par rapport au dénominateur (d_{ij}^{α}). De fait, alors que la distance maximale entre 2 comtés s'élève à 806 kilomètres, le nombre d'habitants dans un même comté peut atteindre près de 2.5 millions. Ainsi, ce type de modèle prévoit plus de $4 \cdot 10^{16}$ navetteurs dans l'état de New-York, nombre dépassant largement le nombre total de terriens.

Si nous nous intéressons à la variabilité V par rapport à la première bissectrice, celle-ci s'avère fortement négative (-33.68), résultat peu surprenant vu la "distance" séparant la première bissectrice du nuage de points. Par contre, le coefficient de détermination par rapport à l'ajustement statistique permet de prédire la variation de 46.6 % des données - une valeur inférieure à celle obtenue via le modèle précédent étant donné que le nuage de points semble plus dispersé - .

Effectuons une démarche similaire au modèle précédent. Comme ce modèle prévoit une surestimation toutefois proportionnelle au flux réel, nous normalisons par une constante e^c (où c symbolise l'ordonnée à l'origine de la régression linéaire précédente) et obtenons le nuage de points suivant :

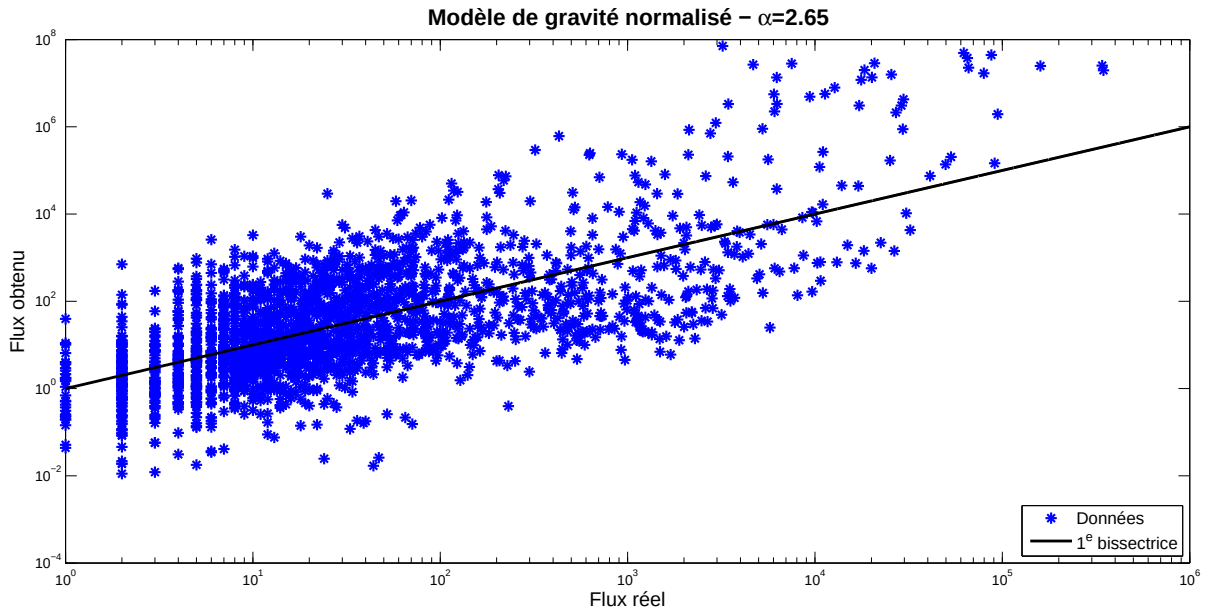


FIGURE 5.21 – Modèle de gravité normalisé

La normalisation améliore nettement ce modèle de gravité : le nuage de points autour de la première bissectrice semble légèrement plus dense que dans le cas précédent ; cette constatation peut être confirmée par la valeur de la variabilité : elle redescend à 0.4656. Près de la moitié des flux est expliquée par la première bissectrice.
Toutefois, le modèle de gravité normalisé prévoit un nombre maximal de navetteurs de 10^8 contre les 10^6 de la réalité. Par conséquent, comme dans le cas précédent, le nombre total de voyageurs ($\approx 10^8$) se révèle plus grand et dépasse largement la quantité de résidents dans l'état de New-York.

Modèle de radiation

Nous avons procédé de manière similaire pour le modèle de radiation :

$$T_{ij}^R = \frac{N_N m_i^2 n_j}{N_T \cdot s_{ij}^\alpha}$$

où N_N symbolise le nombre de navetteurs et N_T le nombre d'habitants dans l'état de New-York.

Le coefficient angulaire de l'ajustement statistique en fonction du paramètre α peut être représenté par la fonction linéaire suivante :

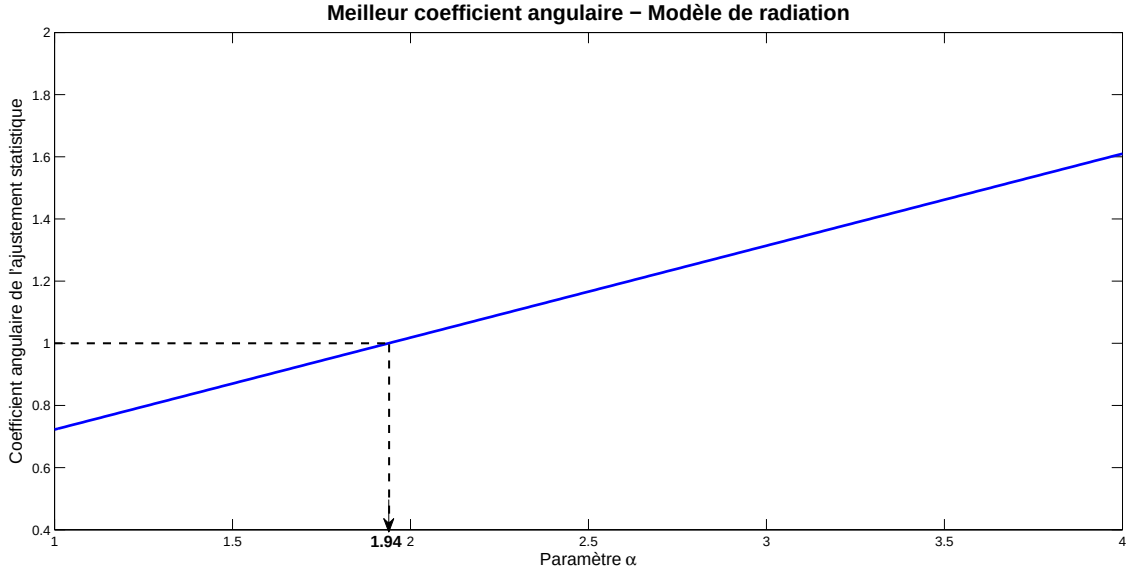


FIGURE 5.22 – Coefficient angulaire dans le modèle de radiation

Le paramètre α nous donnant le coefficient angulaire le plus proche de 1 s'élève à 1.94.

Représentons donc le flux observé en utilisant ce paramètre, en fonction du flux moyen :

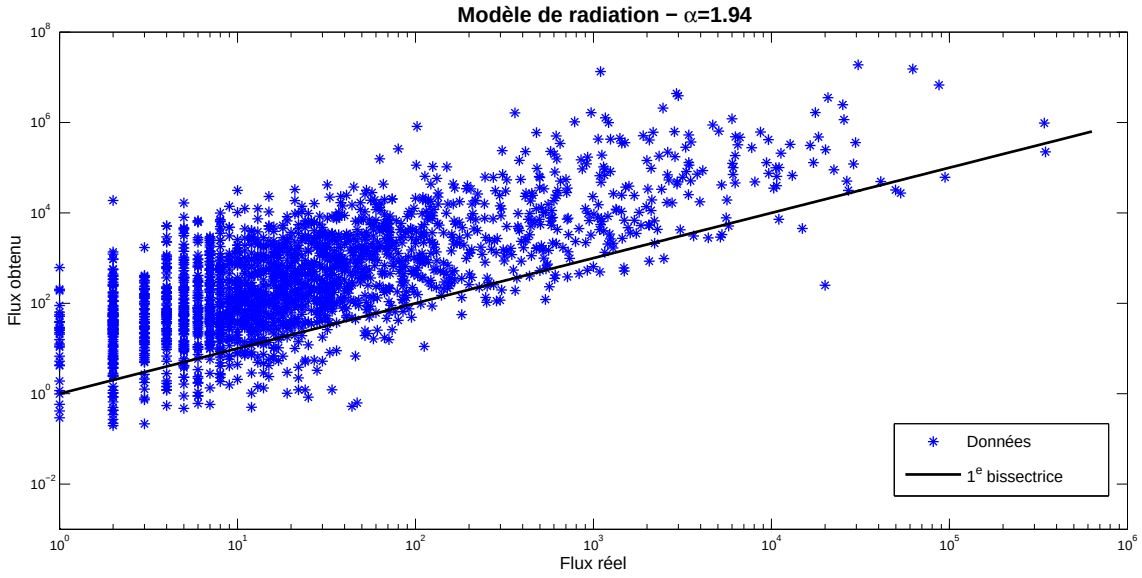


FIGURE 5.23 – Modèle de radiation - $\alpha = 1.94$

Contrairement au graphique établi pour le modèle de radiation via l'ajustement statistique, le modèle se trouve ici nettement amélioré. De fait, les données calculées diffèrent moins des données réelles - contrairement aux cas précédents - . Cependant, ces données se révèlent toujours supérieures à la première bissectrice. Nous pouvons remarquer une similarité avec le modèle de radiation de Simini, ce qui paraît assez logique, étant donné que la définition utilisée pour ce modèle est similaire à celle de l'article.
Ce modèle de radiation prévoit 108 068 287 navetteurs new-yorkais, soit un nombre plus éloigné de la réalité ainsi que de celui du modèle de radiation obtenu via la définition de Simini, et encore supérieur au nombre total d'habitants dans cet état. Néanmoins, il se trouve nettement en deçà du chiffre trouvé avec les modèles de gravité décrits précédemment.

Si nous nous intéressons à présent à la variabilité, elle se révèle supérieure à celle du modèle de gravité. Une observation assez logique étant donné que les données se trouvent plus réparties autour de la première bissectrice. Néanmoins, ce nombre, qui s'élève à -0.35, ne nous satisfait guère.
Par contre, l'ajustement statistique permet d'expliquer la variabilité de 51 % des données, avec dès lors une dispersion plus répandue autour de la droite de régression.

Le modèle de radiation prévoit également une surestimation du flux réel, tout en lui étant toutefois proportionnelle. Aussi, comme dans le cas précédent, nous allons normaliser les données avec la constante e^c (où c représente toujours l'ordonnée à l'origine de la régression linéaire). Nous obtenons le nuage de points suivant :

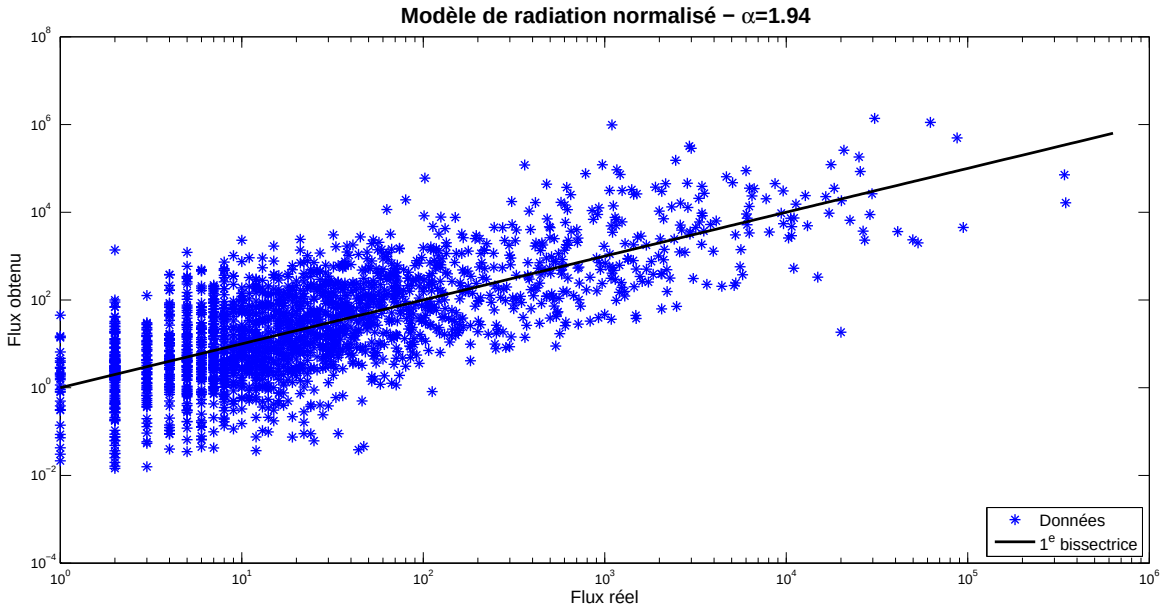


FIGURE 5.24 – Modèle de radiation normalisé - $\alpha = 1.94$

Ce modèle de radiation normalisé semble assez proche de celui de Simini, si ce n'est qu'il possède une meilleure variabilité : un peu plus de la moitié des flux observés (51 %) sont expliqués par la première bissectrice, soit un taux inférieur à celui obtenu avec le modèle de gravité normalisé.
Cependant, ce modèle-ci offre l'avantage de ne pas dépasser le flux réel maximal. En outre, le flux total de navetteurs se révèle seulement inférieur de 79075 unités par rapport au flux total réel. Ce qui le situe dès lors également en deçà de la population totale établie dans l'état de New-York.
Il serait donc intéressant d'étudier plus en détail ce modèle afin d'essayer d'ajuster encore le modèle de Simini. Néanmoins, sans données préalables, il aurait tout simplement été impossible de le normaliser. Mis à part cette remarque, il s'avère donc que le modèle développé par Simini permet de fournir une première estimation des flux de navetteurs dans le comté de New-York.

5.7 Avantages et inconvénients des modèles

Dans cette dernière partie, nous tenterons de dégager les limites et avantages des modèles de gravité et de radiation développés tout au long de cette implémentation. Avant toute chose, il convient d'établir quelques remarques générales.

5.7.1 Remarques générales

En premier lieu, au vu des valeurs gigantesques travaillées aussi bien au niveau du flux que de la distance et du rang, nous avons dû représenter les axes sur une échelle logarithmique. Un avantage de cette utilisation est de donner une représentation et une interprétation immédiates et très utiles des phénomènes où nos ordonnées ont une fonction puissance de nos abscisses. Par contre, certaines valeurs vont être "exclues" - de fait, les rangs et flux nuls ne pourront pas être représentés sur ce type de graphique - .

En ce qui concerne notre ajustement statistique, étant donné que nous avons effectué un changement de variables - $y = \log(x)$ - , nous avons délibérément décidé d'exclure les valeurs nulles liées à cette modification. Ainsi, les coefficients de notre régression linéaire existent.

Dans les deux cas, cependant, il convient de garder à l'esprit que certaines valeurs ont été retirées, ce qui biaise les résultats. Ainsi, par exemple, 123 valeurs égales à 0 se retrouvent dans la matrice du rang, 62 dans celle des distances, 1613 dans la matrice des flux réels, ...

Dans la même optique, certains flux calculés ont été posés à 0. De fait, suite aux différents ajustements statistiques effectués, nous nous trouvions de nouveau confrontés à des flux inexistants (nombre/0). Afin de remédier à ce problème, nous avons choisi de poser ces flux à 0. Cependant, cette hypothèse se révèle fort simpliste et naïve. Ainsi par exemple, aucun navetteur n'irait travailler dans son comté de résidence.

Enfin, la somme totale n'excède jamais le nombre d'habitants de l'état de New-York, excepté pour les modèles améliorés et les 2 de gravité normalisés - de fait, la quasi totalité de notre nuage de points se situe au-dessus de la première bissectrice, et/ou les flux observés se révèlent excessivement grands - . Dans la pratique, certaines personnes peuvent faire la navette jusqu'à un autre état (dans le cas, par exemple, de comtés limitrophes), d'autres n'ont pas d'activité professionnelle (enfants, pensionnés, personnes au chômage, ...) ; ceci induit que le flux total de navetteurs à l'intérieur même de l'état de New-York soit inférieur à sa population totale. Néanmoins, en ajoutant des paramètres ou des contraintes, ce problème pourrait - on l'espère - être facilement résolu.

Notons également que les distances obtenues avec l'application de GOOGLE ne sont pas données en ligne droite, à vol d'oiseau, mais plutôt comme si les navetteurs se déplaçaient en voiture à travers routes et autoroutes ... en utilisant toutefois le chemin le plus rapide.

5.7.2 Modèle de gravité

Le premier inconvénient qui vient directement à l'esprit au niveau du modèle de gravité réside dans la nécessité de disposer de données antérieures - ce qui n'est pas toujours le cas - .
Dans notre recherche, nous avons d'abord essayé de trouver une relation entre le flux réel et la distance, afin de déterminer le flux observé. Sans guère de résultat satisfaisant.
Afin d'obtenir les flux prédits à partir du modèle de gravité amélioré pour les deux définitions, nous avons ensuite dû estimer le paramètre utilisé pour la fonction de dissuasion (noté α ci-dessus, et qui offre la meilleure proximité avec les données réelles). Comme nous l'avons vu dans la partie théorique, il était nécessaire, pour ce faire, de disposer de données antérieures. Même si les modèles normalisés s'avèrent meilleurs que celui défini par l'ajustement statistique - dans le sens où les données sont mieux réparties autour de la première bissectrice, avec une très bonne variabilité - , certains flux se révèlent supérieurs à la valeur maximale du nombre observé de navetteurs, entraînant dès lors un taux de navetteurs supérieur à la population de l'état de New-York.
Cependant, sans ces données, il nous aurait été tout simplement impossible de réaliser le modèle de gravité. A moins d'avancer par essai-erreur, et de comparer avec le flux obtenu pour le modèle de radiation.

En nous intéressant un peu plus longuement à la matrice des flux des modèles de gravité, outre les flux nuls sur la diagonale, nous pouvons tirer les constations suivantes :

– En ce qui concerne l'obtention d'une matrice symétrique :

  • Pour le modèle de gravité prédit à partir de l'ajustement statistique, la distance entre 2 comtés étant identique, le nombre de voyageurs de l'un à l'autre ou inversément, s'avère identique. Le comté d'origine et de destination importe peu : seule la distance compte.

  • De manière similaire, au vu de la définition du nombre de voyageurs entre deux comtés pour le premier modèle de gravité amélioré, il est évident que le calcul de i vers j, ou inversément, est identique - le rôle de chacun étant tout simplement inversé - .

  Cette propriété particulière semble néanmoins très peu réaliste : il est en effet très peu probable que les comtés s'échangent en total équilibre leurs navetteurs. Ainsi, par exemple, il y a plus de chances que le comté de New-York accueille plus de travailleurs en provenance des comtés avoisinants, plutôt que l'inverse - à titre d'exemple 346 268 habitants du Queens exercent leur activité professionnelle dans le comté de New-York, alors que seulement le 20^{e} (20121) font le chemin inverse.

– Toutefois, dans le dernier modèle étudié (la population d'origine ayant un poids plus important), les flux prédits semblent plus réalistes. De fait, aucune symétrie n'existe, les comtés n'échangent pas leurs navetteurs. Malheureusement, ce modèle ne se distinguera pas : le nombre de navetteurs prédit dépasse largement le nombre réel.

Pour pallier à ces deux inconvénients, il aurait été intéressant d'étudier un cas plus général au niveau du dénominateur : mettre chaque poids de ville à une puissance différente (sans toutefois mettre au carré la population d'origine). Néanmoins, 3 paramètres auraient alors dû être découverts dans ce cas, et un tel procédé aurait pris beaucoup trop de temps.

Enfin, dans les 3 cas d'application, la somme de tous les flux au sein de l'état de New-York pour le modèle de gravité, diffère totalement de celle des flux réels : alors que l'ajustement statistique prévoit seulement un trentième du nombre réel de voyageurs, les 2 modèles améliorés et les 2 normalisés prédisent un nombre supérieur à la population de l'état de New-York.

5.7.3 Modèle de radiation

Le principal avantage du modèle de radiation réside dans le fait qu'il est totalement libre de paramètres, une fois approximé le nombre de navetteurs partant de chaque comté. Néanmoins, ce nombre n'aurait pu être découvert sans l'apport préalable d'un jeu de données.

De plus, la somme de tous les flux au sein de l'état de New-York pour le modèle de radiation diffère également du nombre réel total de navetteurs. La surestimation du flux observé par rapport au flux réel en est une conséquence dans les deux premiers cas - Simini et amélioré - . A l'inverse, le modèle de radiation normalisé permettrait d'obtenir un taux de navetteurs inférieur au taux réel.

Malgré cet inconvénient, le modèle développé en 2011 par Simini *et al.* fournirait une bonne première estimation des flux réels observés dans l'état de New-York même si la variabilité en soi n'est pas extraordinaire - avec, toutefois, 3 471 500 de navetteurs prédits en plus - . Le modèle de radiation normalisé, par contre, prédit moins de 100 000 personnes en-dessous du nombre réel. De plus, la première bissectrice permet d'expliquer un peu plus de la moitié des données. Toutefois, cette amélioration n'a été possible que grâce à l'utilisation des données réelles.

En observant plus longuement la matrice des flux prédits, il apparait que le modèle de Simini présage un nombre de navetteurs à l'intérieur d'un même comté et que la somme de ce flux avec le nombre de travailleurs originaires de cet endroit est inférieur à sa population établie. A l'inverse du modèle de gravité, les flux prédits avec les 4 prototypes - Simini, ajustement statistique, amélioration et normalisé - semblent plus réalistes : de fait, aucune symétrie n'existe dans les matrices. Ainsi, par exemple, le taux de travailleurs quittant quotidiennement leur comté pour se rendre dans celui de New-York diffère du nombre de navetteurs effectuant le chemin inverse.

Conclusion

A travers la partie théorique, nous avons pu expérimenter toute la difficulté d'élaborer et de trouver des modèles statistiques capable de prédire les flux de navetteurs, au vu de la complexité de la mobilité humaine. Certes, grâce au développement et à l'utilisation de plus en plus courante de la géolocalisation depuis les années 1930, de nombreux jeux de données ont été créés : au fur et à mesure des décennies, ceux-ci se sont révélés de plus en plus précis, pour en arriver aujourd'hui à une précision spatiotemporelle jamais égalée auparavant. A partir de ces observations empiriques, de nombreux modèles ont été conçus afin de prédire les déplacements des personnes. Les modèles de gravité ont permis de mettre en avant l'effet de la distance physique et ont longtemps figuré comme précurseurs voir leaders en la matière ... jusqu'au jour où les premiers travaux se sont concentrés sur la densité de points entre une origine et une destination. Les modèles de radiation ont été remis au goût du jour grâce à l'existence de jeux de données qui permettent de montrer leur qualité. Parmi ceux-ci, le modèle de Simini *et al.* créé en 2011 se révèle très facile : les personnes choisissent de préférence la meilleure opportunité de travail la plus proche de chez eux et aucun paramètre ne doit être évalué. Ainsi, il est possible de prédire statistiquement les déplacements des travailleurs, quelle que soit la région, étant donné que la densité de population est connue à travers le monde. Autres résultats théoriques obtenus : les deux modèles évoqués sont similaires dans le cas où la population est uniforme, et le nombre de navetteurs sature lorsque la population d'origine ou de destination s'accroît.

Dans la partie pratique, nous avons d'abord analysé le nombre de voyageurs prédits par les modèles de gravité et de radiation, dans le cas d'une population homogène et de villes placées sur une même ligne droite. Faisant écho à la phase théorique, les modèles s'équivalent, quel que soit le nombre d'habitants considéré. Nous avons alors testé les 2 modèles en prenant pour référence le nombre de navetteurs circulant dans l'état de New-York. Le choix du meilleur modèle se révèle ici beaucoup plus délicat ... Rappelons d'emblée la nécessité de disposer de données empiriques antérieures, pour le modèle de gravité. De plus, les modèles de gravité améliorés surestiment le flux réel. Malgré leur normalisation par une constante et une bonne variabilité, ces modèles n'arrivent pas à prédire

un nombre de navetteurs inférieur à la population totale établie dans l'état de New-York. A l'inverse, alors que le modèle de radiation amélioré et normalisé permet de prédire un nombre de navetteurs en deçà du nombre réel, il nous faut toutefois disposer de données antérieures afin de trouver les deux paramètres (choix de l'exposant α associé au rang et de la constante de normalisation). Par contre, le modèle défini par Simini *et al.* offre une première bonne estimation des flux réels observés et semble plus ajusté et réaliste, vu sa non-symétrie. Cependant, gardons bien à l'esprit que le nombre de navetteurs partant du point d'origine doit être également obtenu via des données empiriques.

Ce modèle de radiation offre sans conteste une nouvelle voie d'exploration, plus fiable et performante, dans la compréhension des phénomènes de mobilité en terme de mouvements de population, de flux de navetteurs (de la province vers capitale, et inversément, par exemple), de géographie urbaine, d'épidémiologie, ... Cet outil pourrait permettre de prévoir la propagation d'épidémies, d'ajuster certaines applications en terme de moyens de transport, ... En outre, aucun paramètre ne devant être évalué, nous avons ici les propriétés d'une théorie universelle.
Sans doute aurait-il été intéressant d'étudier l'ajout d'un terme prenant en compte l'avantage supplémentaire de se trouver en l'origine, et d'observer s'il était susceptible d'améliorer l'exactitude des prédictions, comme développé dans le second chapitre. Depuis quelque temps, le modèle de radiation est déjà légèrement remis en question par différents chercheurs : ainsi par exemple, [Masucci *et al.*] suggèrent de normaliser le flux de voyageurs car la définition proposée par Simini *et al.* ne peut être utilisée que pour une région à population totale très importante ; il manque dès lors un "chaînon" pour comprendre la mobilité à une échelle plus réduite (par exemple, dans le cas d'une ville telle que Londres). Dans la même idée, selon [Liang *et al.*], le modèle de radiation ne conviendrait pas pour prévoir les mouvements de passagers en taxi à Pékin. Néanmoins, faute de temps, nous n'avons pu développer ces éléments, qui pourraient éventuellement prolonger cette étude. Il aurait sans doute été également pertinent d'analyser les données pour une région moins peuplée.

Néanmoins, à grande échelle, le modèle de radiation permet d'approximer la mobilité grâce à la densité de points entre origine et destination. Le paramètre le plus important ne réside pas en la distance séparant les deux villes, mais bien, par exemple, dans le nombre d'opportunités d'emploi présentes sur ce chemin. Dans les années à venir, il conviendrait de déterminer pour quelles paires de villes le flux est sous-estimé ou surestimé et quelles sont leurs caractéristiques (aussi bien dans les villes que dans les pays) afin de prévoir plus précisément la mobilité humaine et d'établir un modèle universel.

Bibliographie - Sitographie

[Bamis] I. Bamis (2012), Thesis : Constrained Gravity Models for Networks Flows, *Imperial College London* ;

[Batty] M. Batty (2010), Symmetry, Networks, Flows and Spatial Interaction : Notes and Reflections, *unpublished* ;

[Brockmann *et al.*] D. Brockmann, L. Hufnagel and T. Geisel (2006), The scaling laws of human travel, *Nature* **439**, pp 462-465 ;

[Brockmann 1] D. Brockmann (2010), Statistical Mechanics : The physics of where to go, *Nature Physics* **6**, pp 720-721 ;

[Brockmann 2] D. Brockmann (2010), Following the money, *Physics World*, February 2010, pp 31-34 ;

[Brockmann 3] D. Brockmann (2012), Spotlight on mobility, *Nature News & Views* **48**, pp 40-41 ;

[Clauset] A. Clauset (2011), *Inference, Models and Simulation for Complex Systems - Lectures 2* - , CSCI 7000/4830, University of Colorado Boulder ;

[Lambiotte] R. Lambiotte (2012), *Questions de probabilités et statistiques*, Notes de cours à l'usage des étudiants mathématiciens de Master 1, FUNDP ;

[Lenormand *et al.*] M. Lenormand, S. Huet, F. Gargiulo and G. Deffuant (2012), A Universal Model of Commuting Networks, *PLoS ONE 7(10) : e45985*, doi :10.1371/journal.pone.0045985 ;

[Liang *et al.*] X. Liang, J. Zhao, L. Dong and K. Xu (2012), Modeling collective human mobility : Understanding exponential law of intra-urban movement, *arXiv :1212.6331v1 [physics.soc-ph]* ;

[Masucci *et al.*] A.P. Masucci, J. Serras, A. Johansson and M. Batty (2012), Gravity vs radiation model : on the importance of scale and heterogeneity in commuting flows, *arXiv :1206.5735v1 [physics.soc-ph]* ;

[Noulas *et al.*] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo (2011), A tale of many cities : universal patterns in human urban mobility, *CoRR*, abs/1108.5355 ;

[Simini *et al.*] F. Simini, M. C. González, A. Maritan and A.-L. Barabási (2012), A universal model for mobility and migration patterns, *Nature* **484**, pp 96-100 ;

[Song *et al.*] C. Song, T. Koren, P. Wang and A.-L. Barabási (2010), Modelling the scaling properties of human mobility, *Nature Physics* **6**, pp 818-823 ;

[Stouffer] S. Stouffer (1940), Intervening opportunities : a theory relating mobility and distance, *American Sociological Review* **5**, pp 845-867.

[Census Commuting] Données trouvées sur `http://www.census.gov/population/www/cen2000/commuting`, consulté le 01/02/2013 ;

[Google distance Matrix API] Matrice obtenue sur `https://developers.google.com/maps/documentation/distancematrix/`, consulté le 02/02/2013 ;

[Matlab] Recherche sur le site `http://dmpeli.math.mcmaster.ca/matlab/math1j03/lecturenotes/lecture3_2.htm`, consulté le 01/03/2013 ;

[Massilia] Carte trouvée sur le site `http://www.geographie-muniga.fr/BOITE_OUTILS_SDLV/Alphabet.aspx`, consulté le 03/04/2012 ;

[Wikipédia] Recherche sur le site `http://fr.wikipedia.org/wiki/\%C3\%89tat_de_New_York`, consulté le 01/02/2013 ;

[Wikipédia 1] Recherche sur le site `http://fr.wikipedia.org/wiki/Comt\%C3\%A9s_de_l\%27\%C3\%89tat_de_New_York`, consulté le 04/02/2013.

Annexes

Le contenu des annexes se compose de 4 entités :

A. Lois de puissance et leur invariance d'échelle ;
B. Définition de l'erreur entre le modèle de gravité et le modèle de radiation ;
C. Programmes créés pour le chapitre *Comparaison des modèles à une dimension* ;
D. Programme créés pour le chapitre *Etat de New-York*.

ANNEXE A

Lois de puissance et leur invariance d'échelle

Cette annexe a été développée grâce aux cours de [Clauset] et de [Lambiotte].

Une distribution de loi de puissance est une distribution spéciale de probabilité. Nous pouvons la définir ainsi :

$$p(x) = Cx^{-\alpha} \qquad \forall x \geq x_{min}, \tag{A.1}$$

où α est un paramètre, dit exposant ou puissance ou degré de la loi, et C est la constante de normalisation, $C = (\alpha - 1)x_{min}^{\alpha-1}$, dite constante de proportionnalité. La loi puissance est donc déterminée par ces deux paramètres.

De nombreuses quantités empiriques tournent autour d'une valeur typique, la moyenne. Cependant, ces quantités peuvent varier grâce à l'écart-type, sans toutefois présenter de grands écarts par rapport à cette valeur typique ; de sorte que la moyenne représente la plupart des observations. Par exemple, aux USA, un homme mesure en moyenne 170 cm. Néanmoins, les déviations plus grandes (comme par exemple la taille du plus grand) sont toujours assez proches de cette moyenne, avec une différence de deux fois cet écart-type. La distribution de la taille masculine aux USA peut donc être parfaitement caractérisée par cette moyenne et cet écart-type.

Malheureusement, ce n'est pas toujours le cas. Ainsi, les quantités distribuées selon la loi de puissance sont souvent rencontrées et peuvent caractériser la distribution de quantités familières. Cette loi est observée dans des domaines scientifiques (physique, biologie,...). Certaines valeurs moyennes peuvent perdre tout leur sens face à certaines données, comme, par exemple, les populations des 600 plus grandes villes aux USA - il y a énormément de petites villes et peu de grandes villes - . L'écart-type associé à ce type de données dépasse largement la moyenne.

Une propriété intéressante de la loi puissance est l'invariance d'échelle. En effet, en comparant $p(x)$ et $p(cx)$, avec une constante c, nous pouvons constater que ces deux densités sont toujours proportionnelles. En effet :

$$\begin{aligned} p(cx) &= (\alpha - 1)\, x_{min}^{\alpha-1}\, (cx)^{-\alpha} \\ &= c^{-\alpha} \left[(\alpha - 1)\, x_{min}^{\alpha-1}\, x^{-\alpha}\right] \\ &\propto p(x). \end{aligned}$$

Cette invariance d'échelle n'est rencontrée que pour cette fonction. Il s'ensuit donc que la probabilité relative entre de petits et de grands évènements est identique.

Nous pouvons représenter cette fonction dans un système de coordonnées en log-log via une droite. De fait, en prenant le logarithme des deux côtés de l'équation ((A.1)), nous obtenons :

$$\begin{aligned} \ln p(x) &= \ln \left[(\alpha - 1)\, x_{min}^{\alpha-1}\, x^{-\alpha}\right] \\ &= \ln C \; - \; \alpha \; \ln x \end{aligned}$$

où α représente la pente de la droite.

Reprenons l'exemple des villes américaines. Comme dit précédemment, il y a énormément de petites villes et peu de grandes villes. En fait, dans les distributions de loi de puissance, nous observons une population à grande fréquence suivie par des populations avec des fréquences de plus en plus faibles, diminuant graduellement en une traine. Souvent les événements peu fréquents, constituant cette longue traine, peuvent gagner en importance, entrainant ainsi une plus grande probabilité d'observer les événements extrêmes.

ANNEXE B

Définition de l'erreur entre le modèle de gravité et le modèle de radiation

Réécrivons l'erreur entre le modèle de radiation et le modèle de gravité (2.5) via :

$$\begin{aligned} \bar{m} &= \tfrac{n_i+n_j+s_{ij}}{N_{ij}}; \\ n_i &= \bar{m}(1+\delta_i); \\ n_j &= \bar{m}(1+\delta_j). \end{aligned}$$

Nous obtenons ainsi :

$$\begin{aligned} E &= \frac{1}{N}\sum_{\{i,j:i\neq j\}}\left[\ln\frac{C\ n_i^{\alpha}\ n_j^{\beta}}{d_{ij}^{\gamma}} - \ln\frac{n_i^2\ n_j}{(n_i+s_{ij})(n_i+n_j+s_{ij})}\right]^2 \\ &= \frac{1}{N}\sum_{\{i,j:i\neq j\}}\left[\ln\frac{C\ \bar{m}^{\alpha+\beta}\ (1+\delta_i)^{\alpha}\ (1+\delta_j)^{\beta}}{d_{ij}^{\gamma}}\right. \\ &\qquad \left. -\ln\frac{\bar{m}^3\ (1+\delta_i)^2\ (1+\delta_j)}{(\bar{m}(1+\delta_i)+s_{ij})\ (\bar{m}(1+\delta_i)+\bar{m}(1+\delta_j)+s_{ij})}\right]^2. \end{aligned}$$

Développons le dénominateur, D, du second terme :

$$\begin{aligned} D &= (\bar{m}(1+\delta_i)+s_{ij})\ (\bar{m}(1+\delta_i)+\bar{m}(1+\delta_j)+s_{ij}) \\ &= \bar{m}^2(1+\delta_i)^2+\bar{m}^2(1+\delta_i)(1+\delta_j)+2\bar{m}(1+\delta_i)s_{ij}+\bar{m}s_{ij}(1+\delta_j)+s_{ij}^2. \end{aligned}$$

Comme $\bar{m} = \frac{n_i+n_j+s_{ij}}{N_{ij}}$, la variable s_{ij} peut se réécrire comme :

$$\begin{aligned} s_{ij} &= \bar{m}N_{ij} - n_i - n_j \\ &= \bar{m}N_{ij} - \bar{m}(1+\delta_i) - \bar{m}(1+\delta_j). \end{aligned}$$

En injectant cette variable dans l'équation précédente, nous obtenons :

$$\begin{aligned} D &= \bar{m}^2 N_{ij}^2 - \bar{m}^2 N_{ij}(1+\delta_j) \\ &= \bar{m}^2 N_{ij}(N_{ij} - 1 - \delta_j). \end{aligned}$$

L'erreur peut donc être définie comme :

$$E = \frac{1}{N} \sum_{\{i,j:i\neq j\}} \left[\ln \frac{\bar{m}^{\alpha+\beta}\ (1+\delta_i)^{\alpha}\ (1+\delta_j)^{\beta}}{\rho^2\ d_{ij}^{\gamma}} \ - \ \ln \frac{\bar{m}\ (1+\delta_i)^2\ (1+\delta_j)}{N_{ij}(N_{ij}-1-\delta_j)} \right]^2$$

où $\rho^2 = \frac{1}{C}$.

ANNEXE C

Programmes créés pour le chapitre *Comparaison des modèles à une dimension*

Les deux programmes suivants ont été créés avec le logiciel MATLAB.

Comparaison des modèles à une dimension

```
%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                        %
%                                                                        %
%              COMPARAISON DES MODELES A UNE DIMENSION                   %
%              ---------------------------------------                   %
%                                                                        %
%Représentation du flux moyen attendu en fonction du rang ou de la       %
%distance dans le cas où les villes se situent sur une seule droite.     %
%Nous nous intéresserons tout d'abord à une population homogène (l=0)    %
%uniformément située à égale distance, ensuite à une population plus     %
%hétérogène, composée de 2 groupes distincts séparés par une distance    %
%quelconque (l>0).                                                       %
%Le nombre d'habitants par ville sera identique.                         %
%Pour obtenir le flux moyen du modèle de radiation, deux méthodes seront %
%développées.                                                            %
%                                                                        %
%=========================================================================%


clc
clear all
close all
```

```
%Quelques élements nécessaires
%----------------------------

N=100; %Nombre de villes
d=zeros(1,N);  %Vecteur de la position des différentes villes
dist=zeros(N,N); %Matrice des distances entre chaque paire de villes
rang=zeros(N,N); %Matrice des différences d'ordre entre i et j
Tgrav=ones(N,N); %Flux prédit par le modèle de gravité
Trad=ones(N,N); %Flux prédit par le modèle de radiation

l=10; %Nombre de villes fantômes entre les 2 groupes de villes.
     %Si l=0, nous avons le cas homogène.


n=zeros(1,N);  %Vecteur du nombre d'habitants par ville
for j=1:N
    n(j)=10;  %On suppose population homogène de 1 ou 10 habitant(s)
end


%*************************************************************************
%                           PLACEMENT DES VILLES
%*************************************************************************


%On suppose que les villes sont divisées en 2 groupes communs via le
%paramètre l.

D=N+2*l; %DISTANCE ENTRE DEUX EXTREMITES

if mod(N,2)==0 %Nombre pair de villes : Même nombre dans chaque groupe
    for i=1:(N/2)
        d(i)=i;
    end
    for i=((N/2)+1+l):(D-l)
        d(i-l)=i;
    end
else %Nombre impair de villes : 1 ville supplémentaire dans un des groupes
    for i=1:(N+1)/2
        d(i)=i;
    end
    for i=(((N+1)/2)+1+l):(D-l)
        d(i-l)=i;
    end
end

r=1:N; %Vecteur comprenant l'ordre des différentes villes


%*************************************************************************
%          CALCUL DES RANGS ET DES DISTANCES ENTRE DEUX VILLES
%*************************************************************************


for i=1:N
    for j=i:N
        %Comme la matrice est symétrique, on ne travaille qu'avec la
        %partie supérieure de la matrice.

        %-> DISTANCE ENTRE 2 VILLES
        %   ----------------------
```

```
        d1=abs(d(j)-d(i));
        d2=D-d(j)+d(i);
        if d1<d2
            dist(i,j)=d1;
        else
           dist(i,j)=d2;
        end

        %-> DIFFERENCE D'ORDRE ENTRE 2 VILLES
        %   --------------------------------

        r1=abs(r(j)-r(i));
        r2=N-r(j)+r(i);
        if r1<r2
            rang(i,j)=r1;
        else
            rang(i,j)=r2;
        end
    end

    for j=1:i
        dist(i,j)=dist(j,i);
        rang(i,j)=rang(j,i);
    end
end


%*************************************************************************
%             CALCUL DU FLUX MOYEN ATTENDU POUR LES 2 MODELES
%*************************************************************************


%-> CALCUL DU RANG POUR LE MODELE DE RADIATION
%   ----------------------------------------

s=ones(N,N); %Rang pour radiation

for i=1:N
    for j=1:N
        if i==j
            s(i,i)=0;
        else
            s(i,j)=(abs(rang(i,j))-1)*n(1); %Nombre de villes considérées
                                            %entre les 2 villes FOIS 1 ou 10
                                            %(population de 1 ou 10)
        end
    end
end

%-> CALCUL DU RANG POUR LE MODELE DE GRAVITE
%   ---------------------------------------

sg=ones(N,N); %Rang pour gravité

for i=1:N
    for j=1:N
        if i==j
            sg(i,i)=0;
        else
            if dist(i,j)==rang(i,j)
                sg(i,j)=s(i,j);
```

```
            else
                sg(i,j)=s(i,j)+l*n(1);
            end
        end
    end
end

%-> CALCUL DU FLUX POUR LE MODELE DE GRAVITE
%   ---------------------------------------
%Remarque: on prend le même modèle que celui de radiation

f=zeros(N,N);

for i=1:N
    for j=1:N
       f(i,j)=(n(i)+sg(i,j))*(n(i)+n(j)+sg(i,j));
       Tgrav(i,j)=(n(i)^2*n(j)/f(i,j));
    end
end

%-> CALCUL DU FLUX POUR LE MODELE DE RADIATION
%   -----------------------------------------

for i=1:N
    for j=1:N
        Trad(i,j)=(n(i)^2*n(j))/((n(i)+s(i,j))*(n(i)+n(j)+s(i,j)));
    end
end

%-> CALCUL DU FLUX MOYEN POUR LES MODELES DE RADIATION ET DE GRAVITE
%   ------------------------------------------------------------------

temp=unique(s);
%Somme des flux des paires de villes séparées par une distance de dst :
Tg=zeros(1,floor(D/2));

%Nombre de paires de villes séparées par une distance de dst :
Nbrg=zeros(1,floor(D/2));

%Somme des flux des paires de villes séparées par un rang de rg :
Tr=zeros(1,length(temp));

%Nombre de paires de villes séparées par un rang de rg :
Nbrr=zeros(1,length(temp));


for rg=1:length(temp)
    for i=1:N
        for j=i:N
            if s(i,j)==temp(rg)

                %Si un élément de la matrice de rang équivaut à
                %une valeur de la variable rg, on rajoute +1 au nombre de
                %paires de villes séparées par un rang de cette valeur,
                %ET on rajoute le flux du modèle de radiation pour cette
                %paire de villes (d'origine i et de destination j) à la
                %somme des flux des paires de villes séparées par ce rang.

                Tr(rg)=Tr(rg)+Trad(i,j);
                Nbrr(rg)=Nbrr(rg)+1;
            end
        end
```

```
    end
end

%Même principe que la boucle précédente

for dst=1:floor(D/2)
    for i=1:N
        for j=i:N
            if dist(i,j)==dst
                Tg(dst)=Tg(dst)+Tgrav(i,j);
                Nbrg(dst)=Nbrg(dst)+1;
            end
        end
    end
end

%Flux moyen de voyageurs pour le modèle de radiation :
Tmoyenr=zeros(1,length(temp));

%Flux moyen de voyageurs pour le modèle de gravité :
Tmoyeng=zeros(1,floor(D/2));


for rg=1:length(temp)
    Tmoyenr(rg)=Tr(rg)/Nbrr(rg);
    if Nbrr(rg)==0
        Tmoyenr(rg)=0;
    end
end

for dst=1:floor(D/2)
    Tmoyeng(dst)=Tg(dst)/Nbrg(dst);
    if Nbrg(dst)==0
        Tmoyeng(dst)=0;
    end
end

%************************************************************************
%                        REPRESENTATION GRAPHIQUE
%************************************************************************


%-> COMPARAISON DES 2 MODELES
%   ------------------------

semilogy(1:floor(D/2),Tmoyeng,'-*r')
hold on;
semilogy(1:length(temp),Tmoyenr,'-b')
legend('Modèle de gravité','Modèle de radiation');
xlabel('Rang ou Distance');
ylabel('T_{{moyen}_G} ou T_{{moyen}_R}')
title('Flux moyen attendu')

figure
loglog(Tmoyeng(1:length(Tmoyenr)),Tmoyenr,'*b')
hold on
loglog(10^(-3):0.1:10,10^(-3):0.1:10,'k')
xlabel('Modèle de gravité')
ylabel('Modèle de radiation')
title('Flux prédit par les deux modèles')
legend('Flux prédit','1^e bissectrice')
```

```
%************************************************************************
%      2e METHODE POUR OBTENIR LE FLUX POUR LE MODELE DE RADIATION
%************************************************************************

methode2(N,D,s,dist,Nbrg,Tgrav,Tmoyenr)
```

Deuxième méthode pour prédire le flux en fonction du rang

```
function methode2(N,D,rang,dist,Nbrg,Tgrav,Tmoyenr)

%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%                                                                         %
%     *Deuxième méthode pour prédire le flux en fonction du rang*         %
%   ***************************************************************       %
%                                                                         %
%BUT  -> Représenter le flux en fonction du rang obtenu via la deuxième   %
%        méthode (utilisation des probabilités conditionnelles)           %
%ENTREES : N       = Nombre de villes                                     %
%          D       = Distance entre les deux extrémités                   %
%          rang    = Matrice du rang entre chaque paire de villes         %
%          dist    = Matrice de la distance entre chaque paire de villes  %
%          Nbrg    = Vecteur reprenant le nombre de villes ayant 1 rang rg%
%          Tgrav   = Matrice des flux prédits par le modèle de gravité    %
%          Tmoyenr = Vecteur des flux moyens prédits par le modèle de     %
%                    radiation                                            %
%SORTIE : /                                                               %
%                                                                         %
%=========================================================================%

%Définition des vecteurs
%-----------------------
temp=unique(rang);
Tnew1=zeros(1,length(temp));
proba=zeros(length(temp),length(temp));
proba1=zeros(length(temp),length(temp));


%************************************************************************
%                CALCUL DU FLUX SUIVANT LA DEUXIEME METHODE
%************************************************************************

for rg=1:length(temp)

    %Initialisation des variables
    %----------------------------

    %Initialisation du vecteur comprenant les villes d'origine des paires
    %séparées par un rang de rg :
    xnum=[];

    %Initialisation du vecteur comprenant les villes de destination
    %des paires séparées par un rang de rg :
    ynum=[];


    %Recherche des paires de villes avec un rang de rg
    %------------------------------------------------
```

```
for i=1:N
    for j=i:N
        if rang(i,j)==temp(rg)

            %On ajoute dans les vecteurs la ville d'origine et la ville
            %de destination séparées par un rang de rg

            xnum=[xnum i];
            ynum=[ynum j];

        end
    end
end

%Calcul des différentes probabilités P(dst), P(dst et rg) et P(rg|dst)
%-----------------------------------------------------------------------

nbrp=length(xnum); % Nombre de paires de villes avec un rang de rg

%-> Pour chaque paire de villes séparées par un rang de rg, on
%   enregistre dans le vecteur distp la distance séparant les 2 villes

%Initialisation et définition du vecteur comprenant la distance
%séparant les villes distantes d'un rang de rg :
distp=zeros(1,nbrp);
for  i=1:nbrp
    distp(i)=dist(xnum(i),ynum(i));
end

%-> Définition du nombre de paires de villes (combinaison de 2 parmi N)

%nbrpaire=factorial(N)/(factorial(N-2)*factorial(2));
nbrpaire=N*(N+1)/2;

%-> Calcul de la probabilité d'avoir une distance dist entre 2 villes :
%   Nombre de villes séparées par une distance dist divisée par le
%   nombre de paires de villes au total (P(dist)).

pr=Nbrg/nbrpaire;

%-> Pour chaque distance, on compte le nombre de paires de villes
%   séparées par une distance de i et on divise par le nombre total de
%   paires de villes considérées pour obtenir au final la probabilité
%   qu'une paire de villes soit séparée par une certaine distance et
%   avec un certain rang. On obtient proba(rg et i).
%   Ensuite on calcule la probabilité conditionnelle P(rg|dist)

for i=1:floor(D/2)
    proba(rg,i)=length(find(distp==i))/nbrpaire;
    if pr(i)~=0
        proba1(rg,i)=proba(rg,i)/pr(i); %Probabilité conditionnelle
    else
        proba1(rg,i)=0;
    end
end

%Calcul du flux moyen du modèle de radiation via la deuxième méthode
%-----------------------------------------------------------------------

%-> Pour chaque paire de villes séparées par une distance de k,
%   on enregistre dans la matrice tij la valeur obtenue par le modèle
%   de gravité.
```

```
    tij=zeros(1,floor(D/2));

    for i=1:nbrp
        for k=1:floor(D/2)
            if distp(i)==k
                tij(distp(i))=Tgrav(xnum(i),ynum(i));
                continue
            end
        end
    end

    %-> Calcul du flux moyen pour les villes séparées par un rang de rg

    Tnew1(rg)=0;
    for i=1:floor(D/2)
        Tnew1(rg)=Tnew1(rg)+proba1(rg,i)*tij(i);
    end

end


%*************************************************************************
%                       REPRESENTATIONS GRAPHIQUES
%*************************************************************************


%Comparaison des 2 méthodes
%--------------------------
figure
semilogy(temp,Tmoyenr,'*-b',temp,Tnew1,'-r')
legend('1e méthode','2e méthode')
xlabel('rang')
ylabel('Flux moyen')
title('Comparaison des flux moyens attendus pour le modèle de radiation')

figure
loglog(Tmoyenr,Tnew1)
end
```

ANNEXE D

Programmes créés pour le chapitre *Etat de New-York*

Les programmes suivants ont été créés avec le logiciel MATLAB.

Obtention de la matrice de distance

```
%========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                               %
%                                                                        %
%               OBTENTION DE LA MATRICE DE DISTANCE                      %
%               -----------------------------------                      %
%                                                                        %
%Nous voulons obtenir une matrice de distance pour des données réelles,  %
%par exemple entre les 62 comtés de l'Etat de New-York aux USA.          %
%Nous avons donc utilisé le fichier NY1bis.csv - obtenu grâce au site    %
%http://www.census.gov/population/www/cen2000/commuting - , mais également%
%l'application de Google nous donnant la matrice de distance - Google    %
%distance Matrix API - . Ce code reprend les différentes étapes pour la  %
%création de l'adresse url utilisée pour chaque comté et ouvre la page web%
%nous intéressant. L'utilisation du logiciel Excel nous a ensuite permis %
%de créer la matrice de distance de taille 62*62 (DistanceNY.csv).       %
%                                                                        %
%REMARQUE IMPORTANTE : Certaines villes utilisées par l'application Google%
%ne correspondent pas au comté qui nous intéresse. Ainsi, pour 4 comtés, %
%nous avons décidé de les remplacer par leur chef-lieu respectif à savoir:%
%           -> comté de Clinton (10) : Plattsburgh                       %
%           -> comté de Essex   (16) : Elizabethtown                     %
%           -> comté de Madison (27) : Wampsville                        %
%           -> comté de Nassau  (30) : Mineola                           %
%                                                                        %
%========================================================================%
```

```
clc
clear all

%Lecture du fichier NY1bis.csv, reprenant les différents comtés
%************************************************************

[Origine,Dest,Nbr]=textread('NY1bis.csv','%s %s %f','delimiter',';');
tailletableau=length(Origine);

%Définition d'un vecteur reprenant les différents comtés considérés (Ville)
%et le nombre de comtés (nbrville - à savoir 62)
%*************************************************************************

Ville=unique(Origine);
nbrville=length(Ville);

%Nombre de destinations possibles pour chaque comté d'origine

taille=zeros(nbrville,1);

for i=1:nbrville
    taille(i)=length(find(ismember(Origine,Ville(i))));
end

%Création de l'adresse internet avec modification (cf remarque précédente)
%************************************************************************

%-> COMTE D'ORIGINE
%   ---------------

k=43; %Numéro du comté d'origine

if k==10
    Cheflieu='Plattsburgh NY';
elseif k==16
    Cheflieu='Elizabethtown NY';
elseif k==27
    Cheflieu='Wampsville NY';
elseif k==30
    Cheflieu='Mineola NY';
else
    Cheflieu=char(Ville(k));
end

%-> COMTE(S) DE DESTINATION
%   ----------------------
%Comme la matrice est symétrique, on ne considère que la partie supérieure,
%à savoir les comtés "strictement plus grands" dans l'ordre alphabétique.

if k+1==10
    Town='Plattsburgh NY';
elseif k+1==16
    Town='Elizabethtown NY';
elseif k+1==27
    Town='Wampsville NY';
elseif k+1==30
    Town='Mineola NY';
else
    Town=char(Ville(k+1));
end

const=strcat(Town,'|');
```

```
%Utilisation de strcat, qui concatène les différentes comtés de destination

for i=k+2:nbrville
    if i<nbrville
        if i==10
            Town='Plattsburgh NY';
        elseif i==16
            Town='Elizabethtown NY';
        elseif i==27
            Town='Wampsville NY';
        elseif i==30
            Town='Mineola NY';
        else
            Town=char(Ville(i));
        end
        const=strcat(const,Town,'|');
    else
        const=strcat(const,Ville(i));
    end
end

%-> CREATION DE L'ADRESSE INTERNET + OUVERTURE DE LA PAGE
%   ------------------------------------------------

url=sprintf(strcat...
    ('http://maps.googleapis.com/maps/api/distancematrix/xml?origins=',...
    char(Cheflieu),'&destinations=',char(const),...
    '&mode=driving&language=fr-FR&sensor=false'));
web(url)
```

Relation entre le rang et la distance

```
%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%              RELATION ENTRE LE RANG ET LA DISTANCE                      %
%              ------------------------------------                       %
%                                                                         %
%Représentation du nuage de points du rang en fonction de la distance.   %
%Recherche de la relation existant entre ces 2 quantités grâce à une      %
%régression linéaire - en utilisant une échelle logarithmique - .        %
%                                                                         %
%Pour ce faire, 4 hypothèses seront traitées :                           %
%       - Aucune modification des données                    (Cas 1)     %
%       - Placement aléatoire des comtés                     (Cas 2)     %
%       - Nombre d'habitants aléatoire dans chacun des comtés (Cas 3)     %
%       - Nombre d'habitants identique dans chaque comté      (Cas 4)     %
%                                                                         %
%=========================================================================%

clc
clear all
close all

%Quelques élements nécessaires
%-----------------------------

Sup=141205; %Superficie de l'état de New York : 141 205 km^2

%Coordonnées géographiques des différents comtés
[Ville,Latitude,Longitude]=textread('CoordNY.csv','%s %f %f','delimiter'...
```

```
    ,';');

%Nombre d'habitants par comté
[Ville,n]=textread('PopulationComtéNY.csv','%s %f','delimiter',';');

%Matrice des distances
dist=importdata('DistanceNY.csv');

N=length(dist); %Nombre de comtés
Nt=sum(n); %Population totale établie dans l'état

surf=pi*dist.^2; %Aire du cercle
dens=(Nt/Sup); %Densité de population de l'état


%************************************************************************
%                   CAS 1 : VILLES PLACEES AU BON ENDROIT
%************************************************************************


disp('CAS 1 -> VILLES PLACEES AU BON ENDROIT')

%-> AIRE DU CERCLE
%   --------------

rg=dens.*surf; %Rang défini comme l'aire d'un cercle
temp=unique(dist);

% 1) Nuage de points

disp('Nuage de points : ')
nuagepts(temp,dist,rg,N,1)
xlabel('Distance');
ylabel('Rang');
title('Villes placées au bon endroit avec définition de l"aire du cercle pour le rang - Nuage de points')
legend('Données','Fit')

% 2) Moyenne des rangs pour chaque distance

disp('Moyenne : ')
figure
rga=moyenne(N,temp,dist,rg);
subplot(1,2,1)
loglog(temp,rga,'*-k')
title('Villes placées au bon endroit avec définition correcte du rang - Moyenne par distance')
fit(temp,rga')
legend('Moyenne','Fit')


%-> VRAIE DEFINITION DU RANG
%   ------------------------

s=calculrang(N,dist,n); %Rang calculé "manuellement"

% 1) Moyenne des rangs pour chaque distance

rgb=moyenne(N,temp,dist,s);
subplot(1,2,2)
loglog(temp,rgb,'-*k')
title('Villes placées au bon endroit avec définition correcte du rang - Moyenne par distance')
fit(temp,rgb')
legend('Moyenne','Fit')
```

```
% 2) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,s,N,1)
xlabel('Distance');
ylabel('Rang');
title('Villes placées au bon endroit avec définition correcte du rang - Nuage de points')
legend('Données','Fit')


%************************************************************************
%               CAS 2 : VILLES PLACEES ALEATOIREMENT
%************************************************************************


fprintf('\n')
disp('CAS 2 -> VILLES PLACEES ALEATOIREMENT')

%-> AIRE DU CERCLE
%   --------------

choix2=randperm(N); %Permutation aléatoire des villes
n2=zeros(1,N); %Nouveau vecteur des populations

for i=1:N
    n2(i)=n(choix2(i));
end

% 1) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,rg,N,1)
title('Villes placées aléatoirement avec définition de l''aire du cercle pour le rang - Nuage de points')
xlabel('Distance');
ylabel('Rang');

% 2) Moyenne des rangs pour chaque distance

disp('Moyenne :')
rg2a=moyenne(N,temp,dist,rg);
figure
subplot(1,2,1)
loglog(temp,rg2a,'*-k')
title({'Villes placées aléatoirement avec définition de l''aire'...
    'du cercle pour le rang - Moyenne par distance'})
fit(temp,rg2a)
xlabel('Distance');
ylabel('Rang');
legend('Moyenne','Fit')

%-> VRAIE DEFINITION DU RANG
%   ------------------------

s2=calculrang(N,dist,n2); %Rang calculé "manuellement"

% 1) Moyenne des rangs pour chaque distance

rg2b=moyenne(N,temp,dist,s2);
subplot(1,2,2)
loglog(temp,rg2b,'*-k')
title('Villes placées aléatoirement avec définition correcte du rang - Moyenne par distance')
```

```
fit(temp,rg2b)
xlabel('Distance');
ylabel('Rang');
legend('Moyenne','Fit')

% 2) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,s2,N,1)
title('Villes placées aléatoirement avec définition correcte du rang - Nuage de points')
xlabel('Distance');
ylabel('Rang');
legend('Données','Fit')


%************************************************************************
%           CAS 3 : POPULATION PLACEES ALEATOIREMENT
%************************************************************************


fprintf('\n')
disp('CAS 3 -> POPULATIONS PLACEES ALEATOIREMENT')

%-> AIRE DU CERCLE
%   --------------

%Répartition aléatoire de la population entre les comtés

n3=zeros(1,N);
for i=1:Nt
    choix3=ceil(rand*N); %Nombre aléatoire compris entre 1 et 62
    n3(choix3)=n3(choix3)+1;
end

% 1) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,rg,N,1)
title('Population placée aléatoirement avec définition de l"aire du cercle pour le rang - Nuage de points')
xlabel('Distance');
ylabel('Rang');
legend('Données','Fit')

% 2) Moyenne des rangs pour chaque distance

disp('Moyenne :')
rg3a=moyenne(N,temp,dist,rg);
figure
subplot(1,2,1)
loglog(temp,rg3a,'*-k')
title('Population placée aléatoirement avec définition de l"aire du cercle pour le rang - Moyenne par distance')
fit(temp,rg3a)
legend('Moyenne','Fit')
xlabel('Distance');
ylabel('Rang');

%-> VRAIE DEFINITION DU RANG
%   ------------------------

s3=calculrang(N,dist,n3); %Rang calculé "manuellement"

% 1) Moyenne des rangs pour chaque distance
```

```
rg3b=moyenne(N,temp,dist,s3);
subplot(1,2,2)
loglog(temp,rg3b,'-*k')
title('Population placée aléatoirement avec définition correcte du rang - Moyenne par distance')
fit(temp,rg3b)
legend('Moyenne','Fit')
xlabel('Distance');
ylabel('Rang');
legend('Données','Fit')

% 2) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,s3,N,1)
title('Population placée aléatoirement avec définition correcte du rang - Nuage de points')
xlabel('Distance');
ylabel('Rang');
legend('Données','Fit')


%************************************************************************
%                     CAS 4 : MEME POPULATION PAR VILLE
%************************************************************************


fprintf('\n')
disp('CAS 4 -> POPULATION IDENTIQUE PAR VILLE')
format long
n4=ones(N,1)*Nt/N; %Population totale répartie identiquement

%-> AIRE DU CERCLE
%   --------------

rg4=dens*surf;

% 1) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,rg4,N,1)
title('Population placée identiquement avec définition de l"aire du cercle pour le rang - Nuage de points')
xlabel('Distance');
ylabel('Rang');
legend('Données','Fit')

% 2) Moyenne des rangs pour chaque distance

rg4a=moyenne(N,temp,dist,rg4);
disp('Moyenne :')
figure
subplot(1,2,1)
loglog(temp,rg4a,'*-k')
title('Population placée identiquement avec définition de l"aire du cercle pour le rang - Moyenne par distance')
fit(temp,rg4a);
legend('Moyenne','Fit')
xlabel('Distance');
ylabel('Rang');

%-> VRAIE DEFINITION DU RANG
%   ------------------------

s4=calculrang(N,dist,n4); %Rang calculé "manuellement"
```

```
% 1) Moyenne des rangs pour chaque distance

rg4b=moyenne(N,temp,dist,s4);
subplot(1,2,2)
loglog(temp,rg4b,'*-k')
title('Population placée identiquement avec définition correcte du rang - Moyenne par distance')
fit(temp,rg4b)
legend('Moyenne','Fit')
xlabel('Distance');
ylabel('Rang');

% 2) Nuage de points

disp('Nuage de points :')
nuagepts(temp,dist,s4,N,1)
title('Population placée identiquement avec définition correcte du rang - Nuage de points')
xlabel('Distance');
ylabel('Rang');
legend('Données','Fit')
```

Données réelles

```
%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%                        DONNEES REELLES                                  %
%                        ---------------                                  %
%                                                                         %
%Représentation du nuage de points :                                      %
%       1° Du nombre de navetteurs partant de chaque origine en fonction  %
%          du nombre d'habitants de celle-ci;                             %
%       2° Des données réelles en fonction de la distance ou du rang;     %
%       3° Des données observées en fonction des données réelles.         %
%Recherche de la relation existant entre ces 2 quantités pour chaque cas, %
%grâce à une régression linéaire - via une échelle logarithmique - .      %
%                                                                         %
%=========================================================================%

clc
clear all
close all

%Quelques éléments nécessaires
%-----------------------------

%Flux réel -Nbr- pour chaque paire de comtés (Origine->Dest)
%(Attention : quand la paire n'est pas considérée, le flux sera posé à 0)
[Origine,Dest,Nbr]=textread('NY1bis.csv','%s %s %f','delimiter',';');

%Nombre d'habitants (n) par comté (Ville)
[Ville,n]=textread('PopulationComtéNY.csv','%s %f','delimiter',';');

%Matrice des distances (dist)
dist=importdata('DistanceNY.csv');

N=length(dist); %Nombre de comtés
Nt=sum(n); %Population totale établie dans l'état de New-York

Ville=unique(Origine); %Différents comtés considérés
```

```
%Nombre de comtés de destination pour chaque comté
taille=zeros(N,1);
for i=1:N
    taille(i)=length(find(ismember(Origine,Ville(i))));
end

%Création de la matrice des flux réels
%-------------------------------------

T=zeros(N,N); %Matrice des flux réels
nbr=0;

%On parcourt chaque ligne du vecteur Nbr, en compartimentant par comté
%d'origine.
%(par exemple, le premier comté reprend les lignes 1 à taille(1)=44;
%le deuxième les lignes taille(1)+1=45 à taille(1)+taille(2)=80; etc)

for i=1:N
    for j=nbr+1:nbr+taille(i)
        for k=1:N
            if length(char(Ville(k)))==length(char(Dest(j)))
                if char(Ville(k))==char(Dest(j))
                    T(i,k)=Nbr(j);
                end
            end
        end
    end
    nbr=nbr+taille(i);
end

%Estimation du nombre de navetteurs partant de chaque origine par rapport
%au nombre de résidents du comté
%-------------------------------------------------------------------------

sommeT=zeros(1,N);

for i=1:N
    sommeT(i)=sum(T(i,:))-T(i,i);
end

loglog(n,sommeT,'*')
fit(n,sommeT)
xlabel('Population de l"origine i')
ylabel('Nombre de navetteurs partant de i')


%*************************************************************************
%                       DONNEES REELLES ET DISTANCE
%*************************************************************************


disp('Modèle de gravité ')

%-> REPRESENTATION DU FLUX REEL EN FONCTION DE LA DISTANCE
%   ---------------------------------------------------

dst=unique(dist);
nuagepts(dst,dist,T,N,1)
xlabel('Distance')
ylabel('Flux réel')
title('Nuage de points')
legend('Données', 'Ajustement statistique')
```

```
%-> REPRESENTATION DU FLUX REEL MOYEN EN FONCTION DE LA DISTANCE
%   ------------------------------------------------------------
%
%Treel=moyenne(N,dst,dist,T); %Flux réel moyen
%
% figure
% loglog(dst,Treel,'*b')
% xlabel('Distance')
% ylabel('Flux réel')
% fit(dst,Treel)
% title('moyenne')

%-> REPRESENTATION DU FLUX OBSERVE EN FONCTION DU FLUX REEL
%   ------------------------------------------------------

Tgravf=fluxgravF(N,dist); %Flux observé par le modèle de gravité
nuagepts(unique(T),T,Tgravf,N,1)
xlabel('Flux réel')
ylabel('Flux estimé')
title('Modèle de gravité')

%Représentation de la première bissectrice
d=10^0:0.1:10^6;
hold on
loglog(d,d,'-k')


%*************************************************************************
%                       DONNEES REELLES ET RANG
%*************************************************************************


%-> SUIVANT SIMINI
%   --------------

disp('Modèle de radiation - Simini')
s=calculrang(N,dist,n); %Rang calculé "manuellement"
rg=unique(s);

% 1) Représentation du flux réel en fonction du rang

nuagepts(rg,s,T,N,2)
xlabel('Rang')
ylabel('Flux réel')
title('Vraie définition du rang - Nuage de points')

% 2) Représentation du flux réel moyen en fonction du rang
%
%Treelr=moyenne(N,rg,s,T);
%figure
%loglog(rg,Treelr,'*b')
%xlabel('Rang')
%ylabel('Flux réel')
%title('Vraie définition du rang - moyenne')

% 3) Représentation du flux observé en fonction du flux réel

Trad=fluxradS(N,n,s); %Flux observé par le modèle de radiation
nuagepts(unique(T),T,Trad,N,1)
xlabel('Flux réel')
ylabel('Flux estimé')
```

```
title('Modèle de radiation - Simini')

%Représentation de la première bissectrice
hold on
loglog(d,d,'-k')


%-> SUIVANT L'AJUSTEMENT STATISTIQUE
%   -------------------------------

disp('Modèle de radiation - Ajustement statistique')
rg2=1119.665632*dist.^(1.490921);
yo=unique(rg2);

% 1) Réprésentation du flux réel en fonction du rang

nuagepts(yo,rg2,T,N,2)
xlabel('Rang')
ylabel('Flux réel')
title('Définition du rang via fit - Nuage de points')

% 2) Représentation du flux réel moyen en fonction du rang
%
%Treelr2=moyenne(N,yo,rg2,T); %Flux réel moyen
%loglog(yo,Treelr2,'*r')
%xlabel('Rang')
%ylabel('Flux réel')
%title('Définition rang via fit - moyenne')
%fitrg(yo,Treelr2)

% 3) Représentation du flux observé en fonciton du flux réel

Tradfit=fluxradF(N,s); %Flux calculé via l'ajustement statistique
nuagepts(unique(T),T,Tradfit,N,1)
xlabel('Flux réel')
ylabel('Flux estimé')
title('Modèle de radiation - Ajustement statistique')

%Représentation de la première bissectrice
hold on
loglog(d,d,'-k')

%-> SUIVANT L'AIRE DU CERCLE
%   -----------------------

disp('Modèle de radiation - Aire du cercle')
Sup=141205;
surf=pi*dist.^2;
dens=round(Nt/Sup);
s1=dens*surf; %Rang calculé via l'aire du cercle
rg1=unique(s1);

% 1) Réprésentation du flux réel en fonction du rang

nuagepts(rg1,s1,T,N,2)
xlabel('Rang')
ylabel('Flux réel')
title('Aire du cercle')

% 2) Représentation du flux réel moyen en fonction du rang
%figure
%Treelr1=moyenne(N,rg1,s1,T); %Flux réel moyen
```

```
%loglog(rg1,Treelr1,'*r')
%xlabel('Rang')
%ylabel('Flux réel')
%title('Définition avec aire du cercle pour le rang - Moyenne')
```

Amélioration des modèles

```
%========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                               %
%                                                                        %
%                  AMELIORATION DES MODELES                              %
%                  ------------------------                              %
%                                                                        %
%Représentation du nuage de points du flux prédit par les modèles de     %
%gravité et de radiation en fonction du flux réel, suivant le MEILLEUR   %
%ajustement statistique.                                                 %
%Pour ce faire :                                                         %
%   1° Elaboration de deux définitions différentes des modèles dépendant %
%      d'un paramètre alpha;                                             %
%   2° Recherche de la relation existant entre ces 2 quantités grâce à une%
%      régression linéaire - en utilisant une échelle logarithmique - ;  %
%   3° Choix de l'ajustement statistique où la pente se rapproche le plus %
%      de 1.                                                             %
%                                                                        %
%========================================================================%

clc
clear all
close all

%Quelques éléments nécessaires
%-----------------------------

%Flux réel -Nbr- pour chaque paire de comtés (Origine->Dest)
%(Attention : quand la paire n'est pas considérée, le flux sera posé à 0)
[Origine,Dest,Nbr]=textread('NY1bis.csv','%s %s %f','delimiter',';');

%Nombre d'habitants (n) par comté (Ville)
[Ville,n]=textread('PopulationComtéNY.csv','%s %f','delimiter',';');

%Matrice des distances (dist)
dist=importdata('DistanceNY.csv');

N=length(dist); %Nombre de comtés
Nt=sum(n); %Population totale établie dans l'état de New-York

Ville=unique(Origine);  %Différents comtés considérés

%Nombre de comtés de destination pour chaque comté
taille=zeros(N,1);
for i=1:N
    taille(i)=length(find(ismember(Origine,Ville(i))));
end

%Création de la matrice des flux réels
%-------------------------------------

T=zeros(N,N); %Matrice des flux réels
nbr=0;

%On parcourt chaque ligne du vecteur Nbr, en compartimentant par comté
```

```
%d'origine.
%(par exemple, le premier comté reprend les lignes 1 à taille(1)=44;
%le deuxième les lignes taille(1)+1=45 à taille(1)+taille(2)=80; etc)

for i=1:N
    for j=nbr+1:nbr+taille(i)
        for k=1:N
            if length(char(Ville(k)))==length(char(Dest(j)))
                if char(Ville(k))==char(Dest(j))
                    T(i,k)=Nbr(j);
                end
            end
        end
    end
    nbr=nbr+taille(i);
end


%**************************************************************************
%                         MODELE DE GRAVITE
%**************************************************************************


disp('Modèle de gravité')

%-> INITIALISATION DES VARIABLES
%   ----------------------------

Fpred=zeros(N,N); %Flux prédit par le modèle de gravité
coeff=[]; %Vecteur reprenant les pentes de l'ajustement suivant alpha

%-> AJUSTEMENT LINEAIRE EN FAISANT VARIER ALPHA
%   ------------------------------------------

for alpha=1:0.05:4

    %Calcul du flux prédit pour le modèle de gravité

    for i=1:N
        for j=1:N
            Fpred(i,j)=n(i)*n(j)/(dist(i,j)^alpha);
        end
    end
    for i=1:N
        for j=1:N
            if dist(i,j)==0
                Fpred(i,j)=0; %Pour chaque rang nul, on pose un flux nul.
            end
        end
    end

    %Ajustement statistique

    c=fitbest(unique(T),T,Fpred,N);
    coeff=[coeff c(1)]; %Rajout de la nouvelle pente

end

%-> REPRESENTATION GRAPHIQUE ET DETERMINATION DU MEILLEUR ALPHA
%   ----------------------------------------------------------

figure
```

```
alpha=1:0.05:4;
plot(alpha,coeff)
xlabel('Paramètre \alpha');
ylabel('Coefficient angulaire du fit');
title('Meilleur coefficient angulaire');

%Choix du meilleur alpha
m=(min(find(coeff>=1)));
hold on
plot([1 alpha(m)],[1 1],'--k')
plot([alpha(m) alpha(m)],[0.4 coeff(m)],'--k')

%-> REPRESENTATION GRAPHIQUE EN FONCTION DU ALPHA CHOISI
%   ---------------------------------------------------

for i=1:N
    for j=1:N
        if dist(i,j)==0
            Fpred(i,j)=0;
        else
            Fpred(i,j)=n(i)*n(j)/(dist(i,j)^alpha(m));
        end
    end
end

nuagepts(unique(T),T,Fpred,N,1) %Flux prédit en fonction du flux réel
hold on
%Première bissectrice
d=10^0:0.1:10^6;
loglog(d,d,'-k');
xlabel('Flux réel')
ylabel('Flux obtenu')
title('Modèle de gravité - \alpha=3.2')


%*************************************************************************
%                          MODELE DE RADIATION
%*************************************************************************


disp('Modèle de radiation')

%-> INITIALISATION DES VARIABLES
%   ---------------------------

s=calculrang(N,dist,n); %Rang
Fpredr=zeros(N,N); %Flux prédit par le modèle de radiation
coeff1=[]; %Vecteur reprenant les pentes de l'ajustement suivant alpha

%-> AJUSTEMENT LINEAIRE EN FAISANT VARIER ALPHA
%   ------------------------------------------

for alpha=1:0.01:4

    %Calcul du flux prédit par le modèle de radiation

    for i=1:N
        for j=1:N
            Fpredr(i,j)=n(i)^2*n(j)*sum(sum(T))/(s(i,j)^alpha*sum(n));
        end
    end
    for i=1:N
```

```
        for j=1:N
            if s(i,j)==0
                Fpredr(i,j)=0; %Pour chaque rang nul, on pose un flux nul
            end
        end
    end

    %Ajustement statistique

    c=fitbest(unique(T),T,Fpredr,N);
    coeff1=[coeff1 c(1)]; %Rajout de la nouvelle pente

end

%-> REPRESENTATION GRAPHIQUE ET DETERMINATION DU MEILLEUR ALPHA
%   ------------------------------------------------------------

figure
alpha=1:0.01:4;
plot(alpha,coeff1)
xlabel('Paramètre \alpha');
ylabel('Coefficient angulaire du fit');
title('Meilleur coefficient angulaire');

%Choix du meilleur alpha
m=(min(find(coeff1>=1)));
hold on
plot([1 alpha(m)],[1 1],'--k')
plot([alpha(m) alpha(m)],[0.4 coeff1(m)],'--k')

%-> REPRESENTATION GRAPHIQUE EN FONCTION DU ALPHA CHOISI
%   -------------------------------------------------

for i=1:N
    for j=1:N
        if s(i,j)==0
            Fpredr(i,j)=0;
        else
            Fpredr(i,j)=n(i)^2*n(j)*sum(sum(T))/(s(i,j)^alpha(m)*sum(n));
        end
    end
end

nuagepts(unique(T),T,Fpredr,N,1) %Flux prédit en fonction du flux réel
hold on
%Première bissectrice
loglog(unique(T),unique(T),'-k');
xlabel('Flux réel')
ylabel('Flux obtenu')
title('Modèle de radiation - \alpha=1.94')
```

Quelques fonctions utiles

```
function fit(x,y)

%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%                                                                         %
%    *Ajustement statistique des données en fonction de la distance*      %
%  ***************************************************************      %
%                                                                         %
%BUTS -> Détermination de l'équation de la courbe d'ajustement (type      %
%        puissance) s'approchant au mieux d'un nuage de points :          %
%        y=beta x^alpha                                                   %
%     -> Détermination du coefficient de détermination                    %
%ENTREES : 2 vecteurs (x et y)                                            %
%SORTIE  : /                                                              %
%                                                                         %
%PARTICULARITE : Cette fonction affiche à l'écran principal les deux      %
%paramètres de la fonction ainsi que la valeur du coefficient de          %
%détermination.                                                           %
%=========================================================================%


%Définition des vecteurs ne reprenant pas les valeurs nulles
%-----------------------------------------------------------

ess=[];
d=[];

for i=1:length(x)
    if x(i)>=min(x) && x(i)<=max(x) && y(i)>0 && x(i)>0
        ess=[ess y(i)];
        d=[d x(i)];
    end
end

%Ajustement statistique
%----------------------

%-> Calcul des coefficients de l'ajustement statistique

c=polyfit(log(d),log(ess),1);
fprintf('\n');
fprintf('  - Les coefficients du fit sont : alpha=%f et beta=%f \n',...
    c(1),exp(c(2)));

%-> Représentation graphique de la droite d'ajustement en vert (g)

%xint=linspace(d(1),d(length(d)));
xint=[min(x):1:10 10:10:10^3 10^3:1000:max(x)];
yint=exp(c(2))*xint.^(c(1));

hold on
loglog(xint,yint,'*g');

%-> Détermination du coefficient de détermination r^2

yfit= c(1) * log(d) + c(2);

ymean=mean(log(ess));
SStotal=sum((log(ess)-ymean).^2);
```

```
SSresid=sum((log(ess)-yfit).^2);
rsq = 1 - SSresid/SStotal;
fprintf('\n');
fprintf('    Avec un coefficient de détermination de : %f  \n',rsq);
fprintf('\n');

%-> Détermination du coefficient de détermination par rapport à la première
%   bissetrice

%rsqnoir=erreur(d,ess);
end


function s=calculrang(N,dist,n)

%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%                                                                         %
%               *Calcul du rang suivant la définition *                   %
%             ****************************************                    %
%                                                                         %
%BUT -> Calculer le rang d'un ensemble de paires de villes                %
%ENTREES : N    = nombre de villes                                        %
%          dist = matrice des distances                                   %
%          n    = vecteur reprenant le nombre d'habitants par ville       %
%SORTIE : matrice du rang                                                 %
%                                                                         %
%=========================================================================%

%Initialisation de la matrice du rang
%------------------------------------

s=zeros(N,N);

%Définition de la matrice du rang
%--------------------------------

%Remarque : Le rang entre deux villes se définit comme la population totale
%établie dans le cercle ayant pour rayon la distance entre ces 2 villes,
%excepté les populations d'origine et de destination.

for i=1:N
    for j=1:N
        for k=1:N
            if dist(i,k)<=dist(i,j)  && dist(i,k)>0 && k~=j && k~=i
                s(i,j)=s(i,j)+n(k);
            end
        end
    end
end


function y=moyenne(N,dst,dist,x)

%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%                                                                         %
%  *Calcul de la moyenne pour des valeurs ayant la même caractéristique*  %
%*************************************************************************%
%                                                                         %
%(exemple: Moyenne de flux ayant la même distance)                        %
```

```
%                                                                          %
%BUT -> Calculer la moyenne pour des valeurs (x) ayant la même            %
%       caractéristique (dst)                                              %
%ENTREES : N    = nombre de villes                                         %
%          dst  = vecteur des différentes valeurs distinctes             %
%                 (caractéristiques)                                       %
%          dist = matrice des différentes valeurs caractéristiques       %
%                 pour chaque paire de villes                             %
%          x    = matrice symétrique dont on veut déterminer la moyenne   %
%                 pour des valeurs ayant les mêmes caractéristiques       %
%SORTIE : Un vecteur y reprenant la moyenne pour chacune de ces           %
%         caractéristiques                                                 %
%                                                                          %
%==========================================================================%


%Initialisation des vecteurs
%---------------------------

y=zeros(1,length(dst)); %Moyenne

%Nombre de villes ayant les mêmes caractéristiques
nombre=zeros(1,length(dst));

%Calcul de la moyenne
%--------------------

for k=1:length(dst)
    for i=1:N
        for j=1:N
            if dist(i,j)==dst(k)
                y(k)=x(i,j)+y(k);
                %Somme des valeurs x ayant la même caractéristique
            end
        end
    end
    nombre(k)=length(find(dist == dst(k)))/2;
    %Calcul du nombre de paires de villes ayant les mêmes caractéristiques
    %(Attention, comme la matrice est symétrique, on divise par 2).
end

for i=1:length(dst)
    y(i)=y(i)/nombre(i); %Moyenne pour chaque caractéristique
end


function fitrg(x,y)

%==========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                 %
%                                                                          %
%                                                                          %
%       *Ajustement statistique de données en fonction du rang*            %
%     ********************************************************             %
%                                                                          %
%BUTS -> Détermination de l'équation de la courbe d'ajustement (type       %
%        puissance) s'approchant au mieux d'un nuage de points :           %
%        y=beta x^alpha                                                    %
%     -> Détermination du coefficient de détermination                     %
%ENTREE : 2 vecteurs (x=rang et y)                                         %
%SORTIE : /                                                                %
%                                                                          %
%PARTICULARITE : Cette fonction affiche à l'écran principal les deux       %
```

```
%paramètres de la fonction ainsi que la valeur du coefficient de          %
%détermination.                                                           %
%=========================================================================%


%Définition des vecteurs ne reprenant pas les valeurs nulles
%-----------------------------------------------------------

ess=[];
d=[];
for i=1:length(x)
    if x(i)>=min(x) && x(i)<=max(x) && y(i)>0 && x(i)>0
        ess=[ess y(i)];
        d=[d x(i)];
    end
end

%Ajustement statistique
%----------------------

%Calcul des coefficients du fit

c=polyfit(log(d),log(ess),1);
fprintf('\n');
fprintf('   - Les coefficients du fit sont : %f et %f \n',c(1),exp(c(2)));

%Représentation graphique de la droite d'ajustement

xint=linspace(d(1),d(length(d)),100);
yint=exp(c(2))*xint.^(c(1));
hold on
loglog(xint,yint,'*g');

%Détermination du coefficient de détermination r^2

yfit= c(1) * log(d) + c(2);

ymean=mean(log(ess));
SStotal=sum((log(ess)-ymean).^2);
SSresid=sum((log(ess)-yfit).^2);
rsq = 1 - SSresid/SStotal;
fprintf('\n');
fprintf('    Avec un coefficient de détermination de : %f  \n',rsq);
fprintf('\n');

end


function nuagepts(dst,dist,T,N,num)

%=========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                %
%                                                                         %
%                                                                         %
%                    *Représentation du nuage de points*                  %
%                  ***************************************                %
%                                                                         %
%BUTS -> Représenter le nuage de points d'une série statistique à 2       %
%        variables                                                        %
%     -> Trouver ensuite l'ajustement statistique correspondant           %
%ENTREES : dst  = vecteur reprenant les valeurs distinctes de la matrice  %
%                 suivante                                                %
%          dist = matrice pour les abscisses                              %
```

```
%          T    = matrice pour les ordonnées                                %
%          N    = nombre de villes                                          %
%          num  = 1 (si on a en abscisse la distance) ou                    %
%                 2 (si on a en absisse le rang)                            %
%SORTIE : /                                                                 %
%                                                                           %
%===========================================================================%


%Définition des vecteurs correspondant aux abscisses et aux ordonnées
%--------------------------------------------------------------------

x=[];
y=[];
for k=1:length(dst)
    for i=1:N
        for j=1:N
            if dist(i,j)==dst(k)
                x=[x dist(i,j)]; % Abscisses triées par ordre croissant
                y=[y T(i,j)]; % Ordonnées correspondant aux abscisses
            end
        end
    end
end

%Nuage de points
%---------------

figure
loglog(x,y,'*')

%Ajustement statistique
%----------------------

if num==1
    fit(x,y);
elseif num==2
    fitrg(x,y);
end

%Normalisation
%-------------

%figure
%loglog(x,y/33.151860,'*') %Premier modèle de gravité
%fit(x,y/33.151860);
%figure
%loglog(x,y/13.659483 ,'r*') %Modèle de radiation
%fit(x,y/13.659483)
%figure
%loglog(x,y/91956466.164303 ,'*') %Deuxième modèle de gravité
%fit(x,y/91956466.164303);


function Trad=fluxradS(N,n,s)

%===========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                  %
%                                                                           %
%                                                                           %
%  *Calcul du flux suivant le modèle de radiation - Définition de Simini* %
%***************************************************************************%
```

```
%                                                                          %
%BUT  -> Calculer le flux entre chaque paire de villes suivant le modèle  %
%        de radiation (en particulier via la définition de Simini)        %
%ENTREES : N = nombre de villes                                           %
%          n = vecteur reprenant le nombre d'habitants par ville          %
%          s = matrice du rang                                            %
%SORTIE : matrice des flux                                                %
%                                                                          %
%==========================================================================%

t=zeros(1,N); %Nombre de navetteurs partant de chaque ville
Trad=zeros(N,N); %Flux entre chaque paire de villes

for i=1:N
    t(i)=(n(i)*7990669)/sum(n);
    for j=1:N
        Trad(i,j)=t(i)*(n(i)*n(j))/((n(i)+s(i,j))*(n(i)+n(j)+s(i,j)));
    end
end


function Tradfit=fluxradF(N,rg)

%==========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                 %
%                                                                          %
%                                                                          %
% *Calcul du flux suivant le modèle de radiation - Définition via le fit* %
%**************************************************************************%
%                                                                          %
%BUT  -> Calculer le flux entre chaque paire de villes suivant le modèle  %
%        de radiation (en particulier via la définition du fit)           %
%ENTREES : N  = nombre de villes                                           %
%          rg = matrice du rang                                           %
%SORTIE : matrice des flux                                                %
%                                                                          %
%==========================================================================%


Tradfit=350770.500758*rg.^(-0.632795); %Obtenu via l'ajustement statistique

for i=1:N
    for j=1:N
        if rg(i,j)==0
            Tradfit(i,j)=0; %Pour chaque rang nul, on pose un flux nul
        end
    end
end

function Tgravfit=fluxgravF(N,d)


%==========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                 %
%                                                                          %
%                                                                          %
%          *Calcul du flux suivant le modèle de gravité*                   %
%        ************************************************                  %
%                                                                          %
%BUT  -> Calculer le flux entre chaque paire de villes suivant le modèle  %
%        de gravité                                                        %
%ENTREES : N = nombre de villes                                            %
```

```
%          d = matrice des distances                                        %
%SORTIE : matrice des flux                                                  %
%                                                                           %
%===========================================================================%

Tgravfit= d.^(-2.031753)*1452172.949531;
%Obtenu via l'ajustement statistique

for i=1:N
    for j=1:N
        if d(i,j)==0
            Tgravfit(i,j)=0; % Flux nul lorsque la distance est nulle
        end
    end
end

function rsq=erreur(x,y)

%===========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                  %
%                                                                           %
%     *Calcul de la variabilité autour de la première bissectrice*          %
%   ***************************************************************         %
%                                                                           %
%BUT  -> Calculer le coefficient de détermination de données par rapport    %
%        à la première bissectrice                                          %
%ENTREES : x  = vecteur des abscisses                                       %
%          y  = vecteur des ordonnées                                       %
%SORTIE : Variabilité autour de la première bissectrice                     %
%                                                                           %
%===========================================================================%


%Définition des valeurs observées et des valeurs prédites
%--------------------------------------------------------

yfit= log(x); % Valeurs prédites - y=x est la droite de régression -
yobs= log(y); % Valeurs observées

ymean=mean(yobs); %Moyenne des valeurs observées

%Calcul de la variabilité
%------------------------

SStotal=sum((yobs-ymean).^2);
SSresid=sum((yobs-yfit).^2);
rsq = 1 - SSresid/SStotal;
fprintf('\n');
fprintf('  - Par rapport à la 1ère bissectrice, le coefficient de détermination est : %f  \n',rsq);
fprintf('\n');

function c=fitbest(dst,dist,Fpred,N)

%===========================================================================%
%Pauline Lucas - Master 2 Mathématiques FD                                  %
%                                                                           %
%                                                                           %
%                 *Meilleur ajustement statistique *                        %
%               ************************************                        %
%                                                                           %
%BUT  -> Détermination des paramètres de l'équation de la courbe            %
%        d'ajustement (type puissance) s'approchant au mieux d'un nuage de%
```

```
%        points : y=beta x^alpha                                              %
%ENTREES : dst   = vecteur reprenant les différentes valeurs de dist          %
%          dist  = matrice des distances (en abscisse)                        %
%          Fpred = matrice des ordonnées                                      %
%          N     = nombre de villes                                           %
%SORTIES : Vecteur reprenant les coefficients (beta et alpha)                 %
%                                                                             %
%=============================================================================%

%Définition des vecteurs ne reprenant pas les valeurs nulles
%-----------------------------------------------------------

x=[];
y=[];

for k=1:length(dst)
    for i=1:N
        for j=1:N
            if dist(i,j)==dst(k) && Fpred(i,j)>0 && dist(i,j)>0
                x=[x dist(i,j)]; % Abscisses triées par ordre croissant
                y=[y Fpred(i,j)]; % Ordonnées correspondant aux abscisses
            end
        end
    end
end

%Ajustement statistique
%----------------------

%Calcul des coefficients de l'ajustement statistique

c=polyfit(log(x),log(y),1);
```