Louis Allain

Rapport de stage
2e année

Thème du stage : Analyse exploratoire d'un jeu de données massifs (données d'Orange Cameroun) avec les techniques de network analysis

Structure d'accueil : Orange Innovation

Adresse du stage : 44 avenue de la République, 92320 Châtillon, France

Promotion : 2023

Maître de stage : Erwan Le Quentrec

Référent pédagogique : Basile De Loynes

Remerciements

Je souhaite vivement remercier mes tuteurs Erwan et Stefania, du département SENSE, d'Orange Innovation. Ils m'ont très bien accompagné lors de ce stage, à la fois pour m'intégrer dans l'équipe et pour me guider dans mes travaux pour que mon travail soit à la fois plaisant et utile.
Je tiens également à remercier Virginie, pour m'avoir donné l'opportunité de ce stage, ainsi que pour m'avoir initié au monde de la recherche.

Table des matières

Introduction **4**

1 Environnement de stage **5**
1.1 Historique . . . 5
1.1.1 L'essor d'Orange . . . 5
1.1.2 Orange en Afrique . . . 6
1.2 La recherche . . . 6
1.2.1 Orange Innovation . . . 6
1.2.2 Valorisation des données Orange . . . 7
1.2.3 Une thèse et une collaboration avec Orange Cameroun . . . 8

2 Qualité des données **9**
2.1 Les données à dispositions . . . 9
2.1.1 La table de référence des antennes . . . 9
2.1.2 Les comptes rendus . . . 9
2.1.3 La base de données client . . . 12
2.1.4 Données externes à Orange . . . 12
2.2 Nettoyage et interopérabilité . . . 12
2.2.1 Tirer de l'information des CRA . . . 12
2.2.2 Interopérabilité . . . 13

3 Approche exploratoire **14**
3.1 Analyse du réseau Orange . . . 14
3.2 Analyse de la base client . . . 16
3.3 Analyse des données externes . . . 17
3.3.1 Le recensement de 2005 . . . 17
3.3.2 Les données de l'enquête auprès des ménages . . . 19

4 L'étude d'un graphe social **24**
4.1 Social Network Analysis . . . 24
4.1.1 Définition . . . 24
4.1.2 Statistiques . . . 24
4.1.3 Indicateurs . . . 25
4.1.4 Communautés . . . 26
4.2 Création du graphe . . . 28
4.2.1 Préparer les données . . . 28
4.2.2 Enrichir les noeuds du graphe . . . 31
4.3 Analyse du graphe . . . 31
4.3.1 Statistiques exploratoires . . . 31
4.3.2 Analyse de l'homophilie . . . 32
4.3.3 Étude des communautés . . . 34

Conclusion **38**

Bibliographie **42**

Table des figures

3.1 Distance entre les deux coordonnées pour les antennes ayant deux coordonnées différentes. 14
3.2 Représentation des antennes sur la carte du Cameroun pour les trois code LAC les plus absurde. 15
3.3 Représentation des antennes sur la carte du Cameroun, ainsi que la densité d'antennes par région, en nombre d'antenne par km^2. 15
3.4 Matrice de confusion du premier modèle 16
3.5 Distribution des probabilités d'affectation par genre 17
3.6 Densité de la population Camerounaise, en habitants/km^2, suivant trois niveaux d'administration (2005). 18
3.7 Densité de la population Camerounaise, en habitants/km^2, dans les deux villes les plus peuplées (2005). 18
3.8 Réseau d'antennes d'Orange Cameroun superposé à la densité de population, habitants par km^2, des régions (2005). 19
3.9 Matrice de corrélation des variables de l'enquête auprès des ménages. 20
3.10 Carte des poids spatiaux attribués aux communes 21
3.11 Répartition du taux d'alphabétisation des 15-64 ans, EC-ECAM, 2016. 22
3.12 Comparaison entre la densité des différences entre une commune et ses voisins (rouge) et la densité des différences entre une commune et toutes les autres (gris). 22
3.13 Régression linéaire entre le taux d'alphabétisation et sa variable de décalage spatial associée. 23

4.1 Un exemple de graphe à 6 noeuds et 8 arêtes. 24
4.2 Un exemple de graphe dirigé à 6 noeuds et 8 arêtes. 24
4.3 Un exemple de graphe pondéré et dirigé avec des boucles. 25
4.4 Illustration de l'algorithme de Leiden, Figure 3 du papier *From Louvain to Leiden : guaranteeing well-connected communities* de Traag, Waltman et van Eck [57]. 28
4.5 Visualisation du graphe social aggrégé au niveau des communes. 30
4.6 Détection de communautés avec deux algorithmes différents. 34
4.7 Répartition, par communautés de Leiden, de quatres variables qui informent sur le niveau de richesse. 35
4.8 Répartition, par communautés, du taux d'emploi et du taux de non salarié. 36
4.9 Étude de la pauvreté par communautés de Leiden. 36
4.10 Analyse de la cenralité de degré par communautés de Leiden. 37

Liste des tableaux

2.1 Enregistrement dans la base des CRA d'un appel reçu par A à 12h00. 10
2.2 Enregistrement dans la base des CRA d'un appel passé entre B à A (tout deux clients Orange), à 12h00. 10
2.3 Échantillon choisi de la table des CRA que j'ai traité. 10
2.4 Extrait des comptes rendus des transactions Orange Money [1] 11
2.5 Extrait de la table des comptes rendu d'utilisation des données [2]. 11
2.6 Extrait de la base de données clients [3] 12
2.7 Exemple de la table construite à partir des CRA qui permet de construire le graphe social (les données sont fictives). 13

4.1 Exemple de la matrice des arêtes qui constitue le graphe social. 29
4.2 Listes des communes ayant le plus d'appels sortant... 31
4.3 Listes des communes classées par degré et par poids. 31
4.4 Listes des communes classées par degré et poids entrant. 32
4.5 Listes des communes classées par degré et poids sortant. 32
4.6 Indicateurs de centralité pour les 3 communes les plus élevé et les 3 les plus faible. 33
4.7 Valeur de l'homophiliepour quelques attributs choisis. 33

Introduction

Orange est un fleuron de la technologie française qui participe activement à la recherche et développement en France. Orange est également présent en Afrique, un continent clé du monde de demain. De nombreux projets d'accompagnement du développement sont menés chez Orange en Afrique. Mon stage s'inscrit parfaitement dans cette optique. Un partenariat a été signé avec le Cameroun pour utiliser leurs données dans le cadre d'une thèse. Cette thèse vise à utiliser la théorie des graphes et des hypergraphes pour analyser les données de réseau de groupe, afin d'identifier des acteurs importants pour l'inclusion des femmes dans le domaine de l'entreprenariat. Cette thèse est donc à la croisée entre la recherche en mathématiques et en sociologie quantitive, dans le cadre des hypergraphes, et entre l'application de théories complexes pour des usages concrets. Mon stage vise quant à lui, à utiliser les outils de *social network analysis* pour explorer les premiers échantillons de données fournis par les équipes d'Orange Cameroun.

Les premières traces d'une théorie rigoureuse sur ce type de graphes date de 2003 avec des papiers fondateurs publiés en majorité par Mark Newman [21, 46, 45, 47, 12]. C'est un des auteurs les plus prolifiques dans ce domaine. Il publie un livre référence, *Networks*, en 2011, réédité en 2018 [48]. La détection de communautés dans des graphes est un domaine de recherche encore actif. L'algorithme central dans ce domaine est celui dit de Louvain [7], publié en 2009. De nombreux algorithmes ont été publiés depuis avec de nouvelles possibilités, comme détecter des niveaux hiérarchiques de communautés et la possibilité d'avoir des individus qui font partis de plusieurs communautés à la fois [33]. D'autres algorithmes ont permis d'améliorer à la fois les performances et la qualité de la partition détectée par l'algorithme de Louvain [57]. Puis la communauté scientifique s'est intéressée à de modèles plus complexes que des graphes à interactions par paires (*pairwise interactions*). D'abord, dans les années 2010, avec des graphes multi-couches [31, 8] et des graphes avec une dimension temporelle [22, 19]. Puis, plus récemment, avec des graphes dit d'ordres supérieurs (*higher-order networks*) tels que les complexes simpliciaux et les hypergraphes [32, 26, 2, 4, 40, 3, 6, 25]. Certains résultats montrent qu'il est possible de passer d'un graphe à interactions par paires à un graphe d'ordres supérieurs [60]. L'intérêt de ces outils est que l'analyse de graphes sociaux permets de mettre en lumière des tendances de comportements au niveau individuel, à la fois sur un grand nombre de personnes et à la fois sur une grande période.

Orange, en tant qu'opérateur mobile, dispose de données extrêmement riches, qui s'appliquent parfaitement avec la théorie des graphes sociaux. L'utilisation de ces données peut s'appliquer dans bon nombre de domaines, comme la sociologie ou l'épidémiologie [53, 38]. Mais ces modèles en réseaux ne se limitent pas à la fine granularité des individus. En effet, l'individu est déterminé et/ou influencé par exemple sa classe sociale. L'étude d'un niveaux agrégés comme la classes sociales est donc primordiales piur démeler ce qui tient des caractéristiques individuelle ou des effets de groupes. Les études sociologiques montrent que l'individu influe sur ses paires, les communautés auxquelles il appartient et sa classe sociale en l'occurence. L'étude des interactions au niveau des individus est alors essentielle pour comprendre le comportement des individus. Les graphes sociaux permettent justement de modéliser ces différents niveaux hiérarchiques par un même modèle, et à partir des mêmes données. De plus, l'étude de ces interactions d'ordres supérieurs (entre l'individu et son environnement) et l'étude de graphes d'ordres supérieurs semblent se rejoindre. Ces questions sont évidemment hors de portée de mon stage. Les données peuvent également être agrégées à des niveaux hiérarchiques des appels non pls au niveau des personnes mais au niveau des communes, comme je l'appliquerai.

Les travaux que j'ai menés se concentre sur l'analyse de la qualité des données et l'analyse exploratoire, entre autre en mobilisant les graphes sociaux. L'objectif est de comprendre les données, la façon dont elles sont stockées et la façon dont elles peuvent interopérer. Puis, de m'approprier les méthodes de *social network analysis* pour extraire de premiers résultats sur l'organisation sociale du pays. Quelles sont les premiers enseignements que nous pouvons tirer, à l'aide de la théorie des graphe sociaux, des données réseau d'Orange Cameroun ? Après une rapide description de mon environnement de stage, je m'attèle à étudier la qualité des données à disposition. Puis j'analyse les données notamment à travers une analyse spatiale des antennes du réseau au Cameroun et des variables socio-économiques récupérées à l'Institut National de Statistiques [28] du Cameroun. Enfin, je détaille et j'applique les outils de *social network analysis*, en dégageant les premiers enseignements de cette étude.

Chapitre 1: Environnement de stage

1.1 Historique

Le groupe Orange est aujourd'hui un leader dans l'industrie des télécommunications. C'est le onzième groupe mondial par chiffre d'affaires en 2021, avec 42,5 milliard d'euros. Orange est aussi le troisième opérateur européen derrière Deutsche Telekom (Allemagne) et Vodafone (Royaume-Uni), et africain derrière MTN (Afrique du Sud) et Vodacom (Afrique du Sud). Fin 2021, Orange comptait 270 millions de clients dans le monde et embauchait plus de 140 000 personnes à travers le monde [17].

1.1.1 L'essor d'Orange

C'est en France qu'est développé le premier réseau de communication, par Claude Chappe en 1794. En 1876, Alexander Graham Bell améliore le réseau de télégraphie optique de Chappe en utilisant la télégraphie électrique, il invente le téléphone. L'État français crée alors le ministère des Postes et Télégraphes en 1878, qui devient le ministère des Postes, Télégraphes et Téléphones (PTT) en 1929. À la fin de la seconde guerre mondiale, le Centre d'études des télécommunications (CNET) a pour mission de mener la recherche et le développement de l'industrie des télécommunications. Au même moment est créée la Direction Générale des Télécommunications au sein du ministère PTT, avec comme objectif l'indépendance de la France dans les télécommunications. Dans la deuxième moitié du XX^{ème} siècle, les inventions se succèdent, en 1947 le transistor, en 1961 le laser, en 1962 réception des premières images télévisées provenant des ÈtatsUnis en France, à Pleumeur-Bodou, en 1970 les premières fibres optiques et en 1974 la carte à puce, pour ne citer qu'elles. En 1975, la France mets en place le programme "Delta LP" ou "Plan de rattrapage téléphonique" qui vise à installer 14 millions de lignes en 7 ans.

La Direction Générale des Télécommunications se renomme France Télécom en 1990 pour se conformer aux directives européennes, et devient un établissement de droit public. France Télécom étend sa gamme de produits et services en 1995 en créant Wanadoo et lance l'ADSL (internet haut débit) en 1999, disponible pour le grand public. En 1998, France Télécom est le premier partenaire technique de la coupe du monde et en 1999 du Tour de France, ce qui accroît sa popularité. France Télécom s'internationalise en signant des partenariats avec des entreprises comme Deutsche Telekom (Allemagne) ou encore Sprint (ÈÉtats-Unis) et en achetant de nombreuses parts dans des sociétés de télécommunications. Entre autres, France Télécom rachète Orange à Vodafone (Royaume-Uni) en 2000 pour 40 milliards d'euros, et regroupe toutes ses activités mobiles sous cette marque. France Télécom est alors le quatrième opérateur mondial et le nom Orange s'impose peu à peu chez France Télécom. Le groupe subit de plein fouet la crise de la bulle internet et perd 96% de sa capitalisation boursière entre 2000 et 2002 et devient la deuxième entreprise la plus endettée du monde avec 70 milliards d'euros de dettes. Le plan "Ambition FT 2005" est mis en place par Thierry Breton avec trois leviers principaux : une augmentation du capital à travers l'état à hauteur de 15 milliards, un réaménagement de la dette prise en charge à hauteur de 15 milliards par les banques et investisseurs et enfin 15 milliards de trésorerie à dégager grâce à des efforts internes. Le plan NeXT, mis en place par Didier Lombard, vise quant à lui à réduire les coûts, la masse salariales et à unifier les différentes filiales. La marque France Télécom prend fin en 2013, au profit d'Orange.

Aujourd'hui Orange n'a plus le même rayonnement international que France Télécom du début du XXI^{ème} siècle, mais reste le premier opérateur français devant SFR, Bouygues Telecom et Free Mobile. Orange gère 28% des connexions 5G en France, 31% des connexions 4G, 38% des connexions 3G et plus de 55% des connexions 2G [29]. Orange n'est pas uniquement leader dans les nouvelles technologies mais est aussi l'opérateur qui assure le plus le suivi des anciennes technologies, principalement en raison d'obligations légales. C'est également le premier fournisseur de fibre en Europe avec 65 millions de foyers connectés au premier janvier 2023. Pour autant, le groupe fait face à plusieurs défis dans son organisation. Orange est un acteur majeur dans le domaine de la télécommunication, avec en contrepartie des paris technologiques pas toujours aisés à relever, qui s'illustrent par le relatif échec dans le lancement commercial d'internet, notamment par rapport à Free. Orange fait également face à une crise de son modèle social,

étant issue de l'administration, Orange s'est inscrit dans un modèle social décrit par John Kenneth Galbraith [18] puis par Daniel Cohen : "À l'image de la société féodale, la société industrielle du XX^{ème} siècle lie un mode de production et un mode de protection. Elle scelle l'unité de la question économique et de la question sociale..." [13]. Un modèle qu'il semble aujourd'hui difficile de maintenir, ce qui crée des tensions, notamment sociales. En atteste la crise qu'a traversé le groupe et la condamnation de Dider Lombard, suite à la vague de suicide en 2009. Enfin, Orange doit faire face aux enjeux de sa masse salariale française et à une pyramide d'âge déformée, due aux embauches des années 1975. La solution la plus socialement acceptable est le temps partiel senior, même si elle couvre bon nombre d'insuffisances. En 2022, c'est 10% des effectifs d'Orange France qui sont partis en temps partiel senior [37]. Il faut également recruter des jeunes qui apportent des qualifications plus en adéquation avec les besoins de l'entreprise.

1.1.2 Orange en Afrique

L'Afrique est aujourd'hui le continent le moins développé du monde. Malgré des progrès notables bon nombre de pays sont politiquement instables, certains gouvernements tournent les élections à leur avantage dans une population qui compte encore de nombreux analphabètes (un tiers de la population africaine est analphabète [27]). De plus, les récents coup d'État dans plusieurs pays africains (Soudan, Mali, Guinée, Burkina-Faso, Niger et Gabon) amplifient cette instabilité [56]. Mais le continent africain est un continent riche, par ses ressources naturelles abondantes et surtout par sa population : une personne sur deux à moins de vingt ans [42]. Aujourd'hui 18% de la population mondiale vit en Afrique (1,4 milliard). Et, selon l'ONU, *plus de la moitié de la croissance démographique dans le monde d'ici à 2050 aura lieu en Afrique* [42]. L'Afrique représente donc un marché majeur par son dynamisme. Le continent devrait compter 1 milliard d'utilisateurs de smartphones d'ici 2030, contre 500 millions aujourd'hui [49]. Orange se doit donc de saisir les opportunités de marché qui se développent en Afrique, aussi pour accompagner le développement Africain, notamment à travers le numérique et les télécommunications.

Orange est aujourd'hui présent dans 18 pays de la zone Afrique Moyen-Orient et comprend 146 millions clients mobile et 29 millions clients pour son service Orange Money [49]. Orange est le troisième opérateur de la région en chiffre d'affaires sur l'année 2022 et a pour ambition dans sa stratégie d'être un opérateur multi-services. Au-delà de rendre le numérique accessible à tous, l'opérateur porte une attention particulière à développer son réseau mobile dans les zones les plus reculées. Dans le cadre de sa stratégie multi-services, Orange travaille pour optimiser l'accès a l'énergie, à la finance digitale, la santé et l'éducation. Un des facteur qui ralentit l'accès aux téléphones et au réseau mobile dans les pays africain est l'accès à l'énergie. Plusieurs solutions sont mises à l'essai et déployées, comme des kits solaires comprenant des panneaux solaires qui permettent de recharger un téléphone ou un smartphone par USB. Les services financiers développés à travers Orange Money ouvrent de nouvelles perspectives, sécurité et fluidité des paiements, développement de services d'assurance (scolaire, vie, et santé) et des compteurs électriques intelligents. En tant qu'opérateur, Orange peut également jouer un rôle dans l'accès à la santé et à l'éducation. En fournissant un accès réseau aux populations, Orange peut favoriser le développement de la télémédecine, particulièrement utile dans des régions rurales et difficiles d'accès. À titre d'illustration, Orange a lancé une application en Côte d'Ivoire, m-Vaccin, en partenariat avec l'Alliance du vaccin (GAVI) et les autorités locales. Plus spécifiquement, concernant l'éducation, la connectivité est utile aux universités et aux écoles, en particulier en zones rurales. Le digital peut également répondre aux besoin de formation des enseignants [15]. Enfin, Orange s'investit dans le développement et la professionnalisation de l'entrepreneuriat informel, particulièrement en soutenant les initiatives des femmes. Des sondages réalisés par Orange au Kenya montrent que la première barrière à l'entrée au numérique, pour les femmes, est le prix du téléphone. Proposer un accès à internet à un prix accessible est une priorité pour Orange. Une attention particulière est donc déployée pour faciliter l'accès au crédit, à la formation et à l'entrepreneuriat féminin.

1.2 La recherche

1.2.1 Orange Innovation

Orange est, par son origine, une entreprise axée sur la recherche et le développement. Ses activités de recherche sont la résultante du Centre National d'Études des Télécommunications (CNET). Ces laboratoires sont à la pointe de la technologie et de la recherche depuis les années 50. C'est de ces laboratoires

que sont sorties les innovations connues de tous comme le Minitel en 1982. Mais ces laboratoires ont également participé à placer les jalons du réseau internet mondial. Par exemple, en 1962, la première transmission de télévision par satellite à lieu via la station Pleumeur-Bodou PB1 de France Télécom. En 1988, Orange participe au déploiement du premier câble sous-marin en fibre optique. Il relie la France, l'Angleterre et les États-Unis. En 2000, Orange a contribué à la normalisation internationale de l'ADSL qui a permis l'émergence de l'Internet haut débit. Sur ces réseaux, en 2001, une opération chirurgicale est réalisée à distance, entre un chirurgien à New-York et une patiente à Strasbourg. C'est une première mondiale. Orange a fortement contribué à l'apparition et à l'adoption des dernières générations de réseau mobile, notamment la 3G en 2004, puis plus récemment la 4G (2012) et la 5G (2020), et prochainement la 6G. Orange s'implique également dans les groupes régionaux et mondiaux de la normalisation du réseau de télécommunication, comme l'Internet Engineering Task Force, l'European Telecommunication Standards Institute ou encore le 3rd Generation Partnership Project [50].

Mais, Orange a échoué à prendre certains virages technologiques, notamment celui d'internet. Sans pour autant avoir été en retard dans la recherche, Orange n'a pas réussi à prendre le risque de mettre sur le marché des produits et des services en raison d'une peur de cannibaliser des produits et services dit "vache à lait" [30]. Orange a des centres de recherches et développements qui n'ont pas eu l'opportunité de commercialiser leurs produits. Les deux exemples marquants sont les services de vidéo-conférences (comme Microsoft Teams, Zoom...) et la messagerie instantanée (comme WhatsApp, Facebook Messenger...). Ces deux types services ont été développés dans les laboratoires d'Orange au début des années 2010. Mais pour ne pas cannibaliser ses produits et services et/ou part des calculs courts termes financiers, Orange a préféré ne pas commercialiser de tels services. Malheureusement, d'autres sociétés ont développé les mêmes services et les ont commercialisés. Leur succès est une illustration pour un grand groupe, qu'il faut prendre des risques sur des segments "vache à lait" et trouver la capacité de commercialiser ses innovations.

Aujourd'hui Orange Innovation compte 8 000 employés dont 660 chercheurs et 120 doctorants ou post-doctorants qui se penchent sur des sujets diversifiés . Évidemment, le réseau mobile et l'accès à internet sont des domaines de recherches actifs, mais Orange a également des domaines de recherche comme la réalité augmentée et virtuelle, l'intelligence artificielle et le machine learning, ou encore l'internet des objets. Les chercheurs et ingénieurs d'Orange explorent des questions, en lien avec les innovations (business model, procédé, produits et services et usage), techniques et sociétales. Quels sont les impacts d'une innovation sur la société ? Comment rendre cette innovation accessible de manière durable ? La recherche fait également face à des difficultés. Tout d'abord, la plupart des chercheurs que j'ai pu côtoyer se plaignent du cadre législatif français trop contraignant contrairement à d'autres pays [14]. Par exemple, l'utilisation des données d'Orange pour des activités de recherche est trop réglementé, notamment avec la reglementation RGPD [35]. Il est donc plus difficile d'obtenir des autorisations d'utilisation des données pour Orange que certaines autres entreprises internationales. Un autre point qui m'a marqué lors des différentes réunions d'équipes auxquelles j'ai pu assister est la difficulté de monter un projet de recherche ou de contractualiser un doctorant. Les enveloppes budgétaires pour la recherche et le recrutement des doctorants, qui sont de moins en moins importantes, ou la répartition de la propriété intellectuelle, sur laquelle les parties prenantes sont de plus en plus intransigeantes, sont des exemples de point de forte divergence entre les différents partis d'un projet de recherche. Cela rend long et pénible toute initiative de projet.

1.2.2 Valorisation des données Orange

Mon stage s'est déroulé dans le département Orange Sociology and Economics of Networks and Services (SENSE) qui comprend plusieurs chercheurs en ergonomie, sociologie, économie, économétrie et data-science. Par exemple c'est dans ce département qu'ont été menées des recherches en temps réel sur la propagation du Covid-19 en calculant des indicateurs de mobilité à partir des données auxquelles à accès Orange [53, 58]. Ces chercheurs ont également mené des études en lien avec les politiques d'aide au développement sur la digitalisation de certains secteurs, dans des pays comme le Sénégale [23] ou Madagascar [34].

Un projet important chez Orange et dans le département SENSE est le projet de valorisation des données, un projet transversal qui regroupe plusieurs équipes d'Orange Innovation. Ces dernières années avec l'arrivée du Big Data et de l'intelligence artificielle les opérateurs télécoms se sont rendus compte que les données de leur cœur de métier, c'est-à-dire les données du trafic téléphonique, pouvaient servir

à bien d'autres fins que la gestion du réseau et des clients. Par exemple, les travaux effectués durant la crise du Covid-19 s'inscrivent très bien dans ce projet. Cette prise de conscience a donné naissance au NetMob Data Challenge qui publie des données de télécommunications et laisse 6 mois aux participants pour en extraire de l'information nouvelle. Les meilleurs projets sont alors présentés lors d'une colloque, qui prendra place cette année à Madrid début octobre [43]. Le département SENSE est au coeur de ce projet de valorisation des données de télécommunications, dont deux des membres font partis du comité de programme du NetMob challenge.

1.2.3 Une thèse et une collaboration avec Orange Cameroun

Mon stage s'inscrit parfaitement dans cette stratégie de valorisation des données et dans les initiatives de développement du groupe Orange en Afrique. En effet, une thèse a été engagée en collaboration avec Orange Cameroun pour mettre en valeur les données de l'opérateur dans ce pays. Le Cameroun est un pays d'une trentaine de millions d'habitants d'Afrique Centrale, dans lequel Orange est leader en gérant 46% du volume du marché des télécoms. Fin 2022, Orange Cameroun enregistrait 11,3 millions de connexions mobiles [29]. Orange Money existe au Cameroun depuis 2011, mais suite à l'obtention de l'agrément d'Établissement de Paiement en 2022 [11], le groupe est indépendant (d'une banque tiers) dans sa gestion d'Orange Money pour toutes les activités financières de la marque, allant du paiement par mobile à l'échange de crédit téléphonique. La thèse vise à utiliser toutes les données générées par Orange Cameroun, en mobilisant et en développant des recherches scientifiques sur les notions de graphes, hypergraphes et la recherche de communautés dans ces structures.

Cette thèse va démarrer en novembre 2023 et mon stage a donc eu pour objectif d'effectuer une analyse préliminaire des données sur les premiers échantillons envoyés par l'équipe d'Orange Cameroun. Plus précisément, le but est de comprendre la manière dont les données sont enregistrées et stockées dans ce pays. Dans un premier temps, il s'agit de rendre compte de la qualité des données et de procéder à de premières analyses, pour le développement de l'entreprenariat et l'inclusion financière des femmes.

Chapitre 2: Qualité des données

2.1 Les données à dispositions

L'originalité de cette thèse et donc de ce stage est la combinaison de données issues de différents univers. Les données que j'ai eu à ma disposition sont les suivantes :

- les comptes rendus d'appels, téléphoniques et sms
- les comptes rendu d'appels sur l'utilisations des données mobile (3G, 4G)
- les comptes rendus de transactions de Orange Money
- un échantillon de la base de données clients (limité à 200 000 clients)
- la table de références des antennes qui attribue à chaque antenne sa géolocalisation
- les données de l'enquête complémentaire à la quatrième enquête Camerounaise auprès des ménages, réalisée en 2016
- le recensement de la population Camerounaise de 2005

Toutes ces données portaient sur la semaine du 1er au 5 mai 2023. Pour des raisons de sécurité, toutes ces données ont été pseudonymiser. Ainsi les numéros de téléphones qui apparaissent dans ces tables sont hashés et on ne peut pas remonter leur propriétaire. Malheureusement, ce processus d'pseudonymisation doit être effectué par l'équipe d'Orange Cameroun avant d'envoyer les données au département SENSE. De ce fait, s'il y a une erreur dans cette pseudonymisation, il est d'une part difficile de l'identifier, mais il est aussi impossible de la rectifier sans reprendre la pseudonymisation depuis le début. Nous verrons par la suite que cette pseudonymisation a été à l'origine de complications.

2.1.1 La table de référence des antennes

Chaque opérateur télécom identifie ses antennes de manière unique à l'aide d'un numéro, le Global Cell Identity (CGI). Il se décompose lui-même en quatres numéros :

- Mobile Country Code (MCC) : c'est un numéro attribué au pays par les instances de la télécommunication. Par exemple, le MCC de la France est 208 et le Cameroun 624.
- Mobile Network Code (MNC) : c'est un numéro attribué à chaque opérateur du pays. Au Cameroun, Orange possède le numéro 02.
- Local Area Code (LAC) : chaque opérateur divise son réseau mobile en parcelle, et chaque parcelle se voit attribuer un numéro, c'est le LAC.
- Cell Identification (CI) : dans chacune des parcelles l'opérateur peut déployer plusieurs antennes relais, auxquelles il attribue un numéro, le CI.

Le CGI s'écrit MCC-MNC-LAC-CI. Par exemple au Cameroun, le CGI d'une antenne Orange s'écrit 624-02-12345-67891.

2.1.2 Les comptes rendus

Orange Mobile

La table des comptes rendus d'appels (CRA) est la base la plus importante que j'ai eu à traiter (50 millions de lignes et 18 colonnes). C'est dans cette table que sont stockées les informations des appels et des sms. C'est à dire quel numéro a appelé quel autre numéro, quels sont les types de contrat que possède l'appelé et l'appelant. Il y a aussi des informations sur l'antenne relais qui a transmis l'appel entre les deux terminaux, et des données complémentaires comme la date et l'heure de l'appel ainsi que sa durée. Historiquement, cette base de données est utilisée par Orange pour facturer ses clients.

Avant d'analyser la table des comptes rendu d'appels et de comprendre son organisation, une explication des champs que l'on peut trouver :

- Integrated Mobile Subscriber Identifiy (IMSI) : c'est un numéro d'identification unique de la carte SIM.

— International Mobile Equipment Identity (IMEI) : c'est un numéro d'identification unique du téléphone (l'appareil physique).
— Global Cell Identifiy (CGI) : c'est le numéro unique qu'un opérateur associe à son antenne. C'est le même que dans la table de référence des antennes.

La table des CRA se présente normalement (dans toutes les filiales Orange) de la manière suivante. Considérons un client Orange, dont le numéro est A. Il reçoit un appel, d'un numéro B, le 1er mai à 12h00. L'appel dure 15 secondes. Alors est enregistré dans la table, la ligne suivante :

Date	Heure	Direction	Type	Num. client	imsi client	imei client	Antenne client	Souscription client	Autre numéro	Durée
01/05/2023	120000	Entrant	TEL	A	imsi de A	imei de A	624-02-12341-56789	PREP/POST	B	15

TABLE 2.1 – Enregistrement dans la base des CRA d'un appel reçu par A à 12h00.

Cela nous permet de localiser, approximativement, le numéro A au moment de l'appel, mais pas le numéro B. En effet, pour pouvoir localiser le numéro B, il faut qu'il soit lui aussi un client Orange (puisque Orange ne stocke que les informations de ses clients pour créer leurs factures). Si tel est le cas, on aura :

Date	Heure	Direction	Type	Num. client	imsi client	imei client	Antenne client	Souscription client	Autre numéro	Durée
01/05/2023	120000	Sortant	TEL	B	imsi de B	imei de B	624-02-43214-98765	PREP/POST	A	15
01/05/2023	120000	Entrant	TEL	A	imsi de A	imei de A	624-02-12341-56789	PREP/POST	B	15

TABLE 2.2 – Enregistrement dans la base des CRA d'un appel passé entre B à A (tout deux clients Orange), à 12h00.

Ceci réduit donc considérablement la base de données utilisable. En effet, pour pouvoir localiser à la fois l'appelant et l'appelé, il faut que les deux soient clients Orange.

Malheureusement, les données que j'ai reçu ne se présentaient pas du tout sous cette forme. Pour comprendre, analysons un exemple. Voici l'ensemble des communications entre deux clients Orange (ils apparaissent tout deux dans la colonne "Numéro du client") lors du 1er mai 2023.

Date	Heure	Direction	Type	Num. client	imsi client	imei client	Antenne client	Souscription client	Autre numéro	Durée
01/05/2023	95647	Entrant	SMS	A	imsi de A	imei de A	624-02-12341-56789	POST	B	0
01/05/2023	95648	Entrant	SMS	A	imsi de A	imei de A	624-02-12341-56789	POST	B	0
01/05/2023	103141	Sortant	TEL	B	imsi de A	imei de A	624-02-56565-27477	PREP	A	117
01/05/2023	103141	Entrant	TEL	B	imsi de B	imei de B	624-02-30050-23242	POST	A	117
01/05/2023	180243	Sortant	TEL	A	imsi de B	imei de B	624-02-30050-68941	PREP	B	24
01/05/2023	180243	Entrant	TEL	A	imsi de A	imei de A	624-02-56565-53417	POST	B	24
01/05/2023	180326	Entrant	TEL	A	imsi de A	imei de A	624-02-56565-87432	POST	B	131
01/05/2023	180326	Sortant	TEL	A	imsi de B	imei de B	624-02-30050-91345	PREP	B	131

TABLE 2.3 – Échantillon choisi de la table des CRA que j'ai traité.

Intéressons nous à l'appel de 10h31 entre les numéros A et B. La ligne 3 indique que le numéro B a appelé le numéro A puisque la communication est marquée comme sortante. De la ligne 4, nous pouvons lire que le numéro B a reçu un appel du numéro A, la communication est marquée entrante, exactement à la même heure, pour un appel de la même durée. En examinant la localisation des antennes relais utilisées par cette transaction (à l'aide de la table de référence), on remarque que les deux antennes sont espacées de plus de 100 km. Pourtant ces deux lignes correspondent bien au même appel. Il a lieu le même jour, à la même heure, entre les deux mêmes numéros et dure le même temps. Mais alors, qui est l'appelé et qui est l'appelant ? Et à qui devons- nous associer les antennes ? Il y a un problème avec le numéro des clients qui est mal référencé. Le problème peut venir de la pseudonymisation des numéros de téléphone ou de la base de données de laquelle sont issues ces données. Dans les deux cas, c'est une erreur que je ne peux pas corriger, puisque les données sont la propriété d'Orange Cameroun et que la pseudonymisation est faite par leurs équipes avant de nous envoyer les données.

Si l'on regarde les colonnes de l'imsi et de l'imei du client, on se rend compte que ceux-ci *semblent* être correctement référencés. Pour les communications téléphoniques, le tableau se lirait alors : le 1er mai

-2. Cette table est à titre illustrative et ne contient pas toutes les variables.
-1. Toutes les variables ne sont pas représentées.
0. Toutes les variables ne sont pas présente.

2023, à 10 :31 :41, pour un appel qui a duré 117 secondes, l'imsi de A (ligne 3 sortante) a appelé l'imsi de B (ligne 4 entrante). La personne A se trouvait proche de l'antenne 624-02-56565-27477, et la personne B se trouvait proche de l'antenne 624-02-30050-23242. C'est cet algorithme que j'ai utilisé pour reconstituer les appels de la semaine, que je détaille en section 2.2

Orange Money

La table des transactions Orange Money est construites sur un modèle tout à fait différent des compte rendus d'appels. Ici une ligne représente une transaction, et on peut caractériser chaque transaction par sa quantité, ses commissions ou son type.

Num. expéditeur	Num. receveur	Quantité transaction	Com. payées	Com. reçues	Autres com.	Charges reçues	Charges payées	Date-Heure
Numéro A	Numéro B	5000	36	0	0	75	0	2023-05-01 14 :57 :12
Numéro C	Numéro D	1100	1.98	0	0	0	0	2023-05-04 17 :27 :44
Numéro E	Numéro F	102	0	0	0	0	0	2023-05-02 08 :21 :12

TABLE 2.4 – Extrait des comptes rendus des transactions Orange Money [1]

L'intérêt de ces données était double. D'une part, il était possible d'étudier le graphe social des échanges monétaires au Cameroun. D'autre part, conjointement avec la table des comptes rendu d'appels, il est possible de localiser les transactions Orange Money. En effet, chaque transaction Orange Money s'accompagne d'un SMS de confirmation, visible dans les comptes rendus d'appels. Ce SMS permettrait de localiser les transferts d'argent. Malheureusement, puisque dans cette table nous ne disposons que du numéro du client, et que nous savons que celui ci est mal renseigné dans les comptes rendu d'appels, il est impossible de faire une jointure entre ces deux tables.

De plus, en faisant quelques statistiques sur cette table, je me suis rendu compte qu'elle comportait des incohérences. Tout d'abord la table contient 615 443 transactions, ce qui équivaut à 123 089 transactions par jour. Or, le directeur d'Orange Money Cameroun, Emmanuel Tassembedo, annonçait en 2021 "[qu'ils faisaient] environ trois millions de transactions par jour" [1]. Emmanuel Tassembedo, déclarait que Orange Money réalisait plus de 800 milliards de F.Cfa par mois en valeurs cumulées de transactions [1]. La somme de toutes les transactions présentes dans la base de données est de 13,3 milliards de F.Cfa (sur 5 jours donc), soit environ 79,8 milliards de F.Cfa par mois.

Les premières statistiques issues des comptes rendus des transactions Orange Money ne sont pas du tout rassurantes quant à sa fiabilité. C'est pourquoi je n'ai pas traité cette table lors de mon stage, hormis pour vérifier sa qualité.

Données mobiles

Cette table rend compte de l'utilisation des données mobiles (3G, 4G) par les abonnées d'Orange Cameroun. Elle aurait dû servir à caractériser les abonnés Orange par le volume d'octets utilisées et l'utilisation de ces données. Malheureusement, encore une fois, cette table ne contient que les numéros de téléphone des individus. Puisque ce champs est corrompu dans la table des CRA, impossible de l'utiliser comme clé de jointure.

Num. client	Octets reçus	Octets envoyés	Code du service	Antenne	Crédit restant	Volume d'octets restant	Date
Numéro A	493 276	208 236	Data Prepaid WhatsApp	624-02-12345-98765	0	0	2023-05-02
Numéro B	31 271	9 092	Data Browsing MyCanal	624-02-10101-25738	0,12	0	2023-05-05
Numéro C	110 200 482	4 410 626	Data Prepaid Allip	624-02-20202-43176	400,76	252 405 576	2023-05-03

TABLE 2.5 – Extrait de la table des comptes rendu d'utilisation des données [2].

1. Cette table est à titre illustrative et ne contient pas toutes les variables.
2. Toutes les variables ne sont pas représentées.

2.1.3 La base de données client

C'est la base de données d'Orange Cameroun qui regroupe toutes les informations des clients. Malgré notre demande d'avoir accès à toute la base de données, il ne nous a été envoyé qu'un échantillon de 200 000 individus.

Num. client	Type personne	Nom Prénom	Date de naissance	Date activation	Ville	Type d'abonnement	IMEI	Segmentation	Profession	Genre
Numéro A	PP	Dupond Pierre	1990-01-01	2022-01-01	NaN	pre-paid	357854...	B2C	NaN	NaN
Numéro B	Flotte	Dupont Marie	2000-01-01	2015-05-05	NaN	post-paid	352345...	B2B	NaN	NaN
Numéro C	PDV	Dupon Jeanne	1970-12-12	2010-10-10	NaN	hybrid	359287...	B2C	NaN	NaN

TABLE 2.6 – Extrait de la base de données clients [3]

Cette base de données serait très intéressante s'il ne manquait pas autant de données. En effet, les champs "Ville", "Profession" et "Genre" qui sont extrêmement utiles pour notre étude, ne sont pas renseignés. De plus, pour certaines variables (type de personne, segmentation) il y a des modalités que je ne comprenais pas et que les équipes d'Orange Cameroun ne m'ont pas précisées malgré mes demandes.

2.1.4 Données externes à Orange

J'ai utilisé deux sources de données pour compléter les données de télécommunications. La première est le recensement de la population camerounaise de 2005 [10]. Malheureusement il n'existe pas, à ce jour, de recensement au niveau des communes plus récentes. La seconde est l'enquête complémentaire à la quatrième enquête camerounaise auprès des ménages (EC-ECAM) [9]. La quatrième enquête camerounaise auprès des ménages a été réalisée en 2014 et a permis de dégager de nombreux indicateurs sur les conditions de vie des ménages. En s'appuyant sur cette enquête, l'EC-ECAM, réalisée en 2016 présente des indicateurs sur la pauvreté monétaire et des indicateurs d'Objectif au Développement Durable (ODD) au niveau communal. Les données de cette enquête complémentaire sont disponibles sur le site de l'Institut National de la Statistique du Cameroun. Elles sont présentées par monographie régionale. On en compte 12, une par région plus deux, pour les villes de Douala et Yaoundé. Les données brutes sont disponibles dans ces monographies sous le format de tableau qu'il a fallu d'abord exporter du PDF vers une feuille Excel pour pouvoir ensuite les importer sous Python.

J'ai longtemps cherché d'autres données sur les conditions de vie des ménages Camerounais, mais à chaque fois les données étaient payantes ou bien il fallait faire une demande formelle par mail (demande qui m'a été refusée par l'institut). Le site de l'Institut Nationale de la Statistique (INS) met à disposition une base de données ouverte. Mais ces données sont la plupart du temps disponible au niveau du pays ou des régions seulement, et pas au niveau des communes. Les données brutes sont également difficilement téléchargeables. Mais sur une période plus longue, peut-être aurait-il été possible de rentrer en contact avec l'INS pour avoir accès à des données plus facilement réutilisables.

2.2 Nettoyage et interopérabilité

2.2.1 Tirer de l'information des CRA

Comme nous l'avons vu, les comptes rendus d'appels sont mal renseignés. L'alternative a donc été d'identifier les lignes qui correspondent à la même communication à l'aide de la date et de l'heure des appels. Puisque nous allons utiliser la durée de l'appel, nous devons exclure les SMS. Ensuite, nous groupons les lignes suivant la date et l'heure de l'appel. Dans chacun de ces groupes, nous séparons les appels "Sortant" des appels "Entrant". On sait que la probabilité d'avoir deux appels différents à la même date, à la même heure et de la même durée est très faible (mais pas nulle). Ainsi, nous pouvons réaliser une jointure sur les tables "Sortant" et "Entrant", avec comme clé la date, l'heure et la durée de l'appel. Je ne garde de cette jointure que les correspondances uniques, c'est-à-dire que si il y a eu deux appels à la même heure et de la même durée, je ne peux pas correctement déterminer les deux bouts de l'appel. En résulte une table, dont chaque ligne représente un appel, avec le numéro IMSI des deux interlocuteurs ainsi que leur antenne respective, la date, l'heure et la durée de l'appel.

3. Toutes les variables ne sont pas présente.

Imsi emetteur	Imsi récépteur	Antenne emetteur	Antenne récépteur	Durée	Date	Heure
Imsi A	Imsi B	624-02-12345-67891	624-02-12345-11234	121	01-05-2023	11 :10 :23
Imsi C	Imsi D	624-02-98765-12345	624-02-61334-98467	461	03-05-2023	18 :23 :16
Imsi B	Imsi E	624-02-12345-34521	624-02-72791-11837	43	04-05-2023	07 :08 :34

TABLE 2.7 – Exemple de la table construite à partir des CRA qui permet de construire le graphe social (les données sont fictives).

Bien sûr, je suis conscient que ces données sont incomplètes. Mais compte tenu de la difficulté à obtenir des données mises à jour lors de mon stage, je me suis contenté de pouvoir reconstituer une partie des appels de la semaine. Finalement, sur les 50 millions de lignes dans la table des CRA, je retrouve 5 172 033 appels, pour lesquels je sais qui est l'appelant, qui est l'appelé et quelles sont les antennes relais, respective, qui ont servies à transmettre l'appel.

2.2.2 Interopérabilité

Un des points les plus intéressant de toutes ces données était leur interopérabilité. Par exemple, l'idée d'analyser les transferts Orange Money d'un point de vue géographique, en retrouvant un SMS de confirmation de la transaction dans la table des CRA. Malheureusement cela n'a pas été possible pour plusieurs raisons, dont la mauvaise pseudonymisation des numéros de téléphone, qui sont une partie essentielle de la clé de jointure entre les deux tables. Pour la même raison, je n'ai pas pu utiliser la base de données clients.

J'ai tout de même réussi à utiliser différentes sources de données ensemble. En effet, les données externes à Orange (recensement et EC-ECAM) ont vocation à enrichir le graphe sociale à différents niveaux d'aggrégation, notamment à caractériser les noeuds de ce graphe. Cela permet ensuite de caractériser les communautés que l'on peut trouver dans le graphe, comme nous le verrons dans le chapter 4.

Chapitre 3: Approche exploratoire

3.1 Analyse du réseau Orange

Une table de données essentielle est la table de référence des antennes. D'autant plus que cette base de données est extrêmement sensible aux changements. En effet, dans une optique d'amélioration du réseau, les antennes sont réparées et changées régulièrement. La base de données doit donc être mise à jour, en enlevant l'ancienne antenne (qui n'est plus utilisée) et en ajoutant la nouvelle. Cela implique d'avoir un historique précis des antennes, de leur position géographique mais surtout de leur date de mise en service et de mise hors service. Il ne faut pas utiliser la même base de référence des antennes pour toutes les données de communication. Lors de mon stage, les données que j'ai eu à manipuler s'étalent sur une période d'une semaine et je n'ai donc pas eu ce problème. Malgré cela, la base de données de référence que j'avais à ma disposition était incomplète et certaines données étaient incohérentes. L'équipe d'Orange Cameroun connaît les lacunes de cette table et travaille actuellement à la rendre plus fiable. Son exploration permet tout de même de comprendre le fonctionnement du réseau Orange au Cameroun et peut compléter les interprétations du graphe sociale.

La base de données compte très peu de valeurs manquantes (moins de 0,0006%). En revanche, certaines antennes apparaissent plusieurs fois dans la base. Sont donc attribuées à ces antennes deux coordonnées différentes. Pour analyser ces différences j'ai calculé la distance, pour toutes ces antennes, entre leurs deux coordonnées qui sont données. J'obtiens alors le graphique suivant :

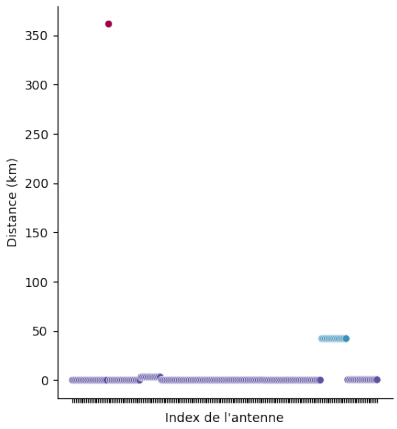


FIGURE 3.1 – Distance entre les deux coordonnées pour les antennes ayant deux coordonnées différentes.

On peut voir que pour la plupart des antennes l'écart entre les deux coordonnées est de quelques mètres, ou au maximum de quarante kilomètres. Pour toutes ces antennes je n'ai gardé dans la base qu'une seule coordonnées (la première, dans l'ordre d'apparition dans la table). J'ai considéré que pour des écarts jusqu'à 40 kilomètres, l'erreur était négligeable et que les deux coordonnées se trouvaient toutes les deux dans la même commune. Pour l'unique antenne qui possède deux coordonnées distante de plus de 300 kilomètres, je l'ai retiré de la base de données.

Ensuite, pour vérifier la véracité des données nous pouvons également utiliser le code LAC des antennes. En effet, on sait que toutes les antennes partageant le même code LAC doivent être proches. J'ai donc calculé, pour chaque code LAC, la distance moyenne entre toutes les antennes. J'ai trouvé trois codes LAC, pour lesquels la distance moyenne était anormalement élevée. En représentant ces antennes sur la carte du Cameroun on voit bien qu'elles sont réparties dans tout le pays. Ces antennes ont donc un défaut de référencement dans la table, je les ai donc retiré de la base de donnée.

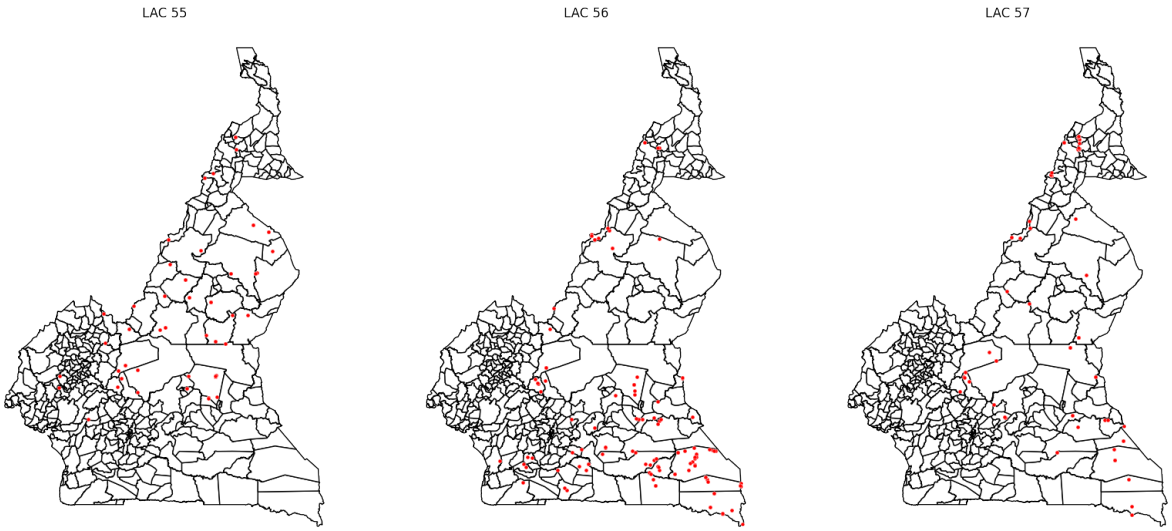


FIGURE 3.2 – Représentation des antennes sur la carte du Cameroun pour les trois code LAC les plus absurde.

Maintenant que les antennes mal référencées ont été retirées de la base de données, je peux représenter la base de données sur une carte. Cela permet de voir si les coordonnées des antennes ont un sens. On

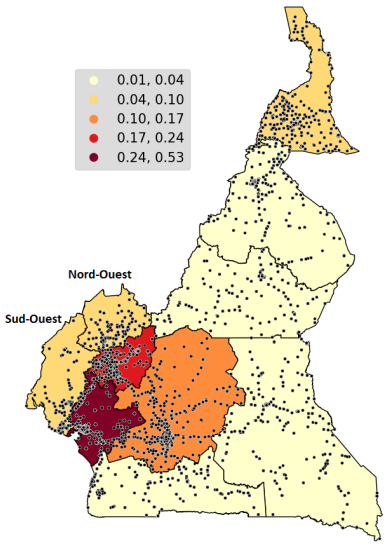


FIGURE 3.3 – Représentation des antennes sur la carte du Cameroun, ainsi que la densité d'antennes par région, en nombre d'antenne par km^2.

voit que toutes les antennes de la base de données sont bien sur le territoire Camerounais. On remarque également une concentration des antennes dans les villes et le long des principales routes du pays. Les régions rurales du pays ressortent, avec les densités d'antennes les plus faibles du pays. On peut noter tout de même une plus faible concentration des antennes dans les régions du Nord-Ouest et du Sud-Ouest. En effet, ces régions sont historiquement anglophones et l'opérateur principal est le Sud-Africain MTN. Le réseau Orange y est moins développé, car il y a moins de clients potentiels.

Une idée pour vérifier la cohérence de la base de données des antennes, est en la couplant avec les comptes rendus d'appels, de calculer la vitesse de déplacement des individus. En utilisant l'API Open Source Routing Machine (OSRM) [52], il est possible de calculer la distance qui sépare deux positions en simulant un trajet en voiture, à vélo ou à pied. Mais je n'ai pas réussi à correctement utiliser cette

API. En effet, j'ai obtenu des résultats différents lors de mes appels à l'API à travers Python et lorsque je testais le service sur le web. N'arrivant pas à résoudre le problème, j'ai préféré avancer sur le reste du stage.

3.2 Analyse de la base client

Étant donné que la thèse a comme objectif l'inclusion des femmes dans le pays, il est primordial de pouvoir identifier le genre des individus. Cette donnée se trouve de toute évidence dans la base de données client. Malencontreusement, à la réception des données nous nous sommes rendu compte que cette variable de genre manquait. Mais les prénoms et noms des clients étant renseignés, nous pouvons estimer une probabilité de genre à partir de ces informations. Et pourquoi pas utiliser les données opérateurs pour ensuite identifier le genre d'un individu à partir de son utilisation du réseau. La première méthode mise en place a été d'utiliser un package Python (`gender-extractor` [59]) qui infère le genre à partir du prénom. Mais l'algortihme renvoyait dans 50% des cas la réponse "ambiguous". Du fait des prénoms Camerounais contenu dans la base de données Orange, qui ne sont probablement pas bien représenté dans la base d'apprentissage de ce package Python, les résultats n'étaient pas probants (biais d'apprentissage [5]). En première intention, pour inférer le sexe sur les 50% restant j'ai donc mis en place un modèle de régression logistique avec comme base d'apprentissage les individus dont le sexe a été inféré par le premier algortihme. La variable expliquée est le sexe, et les variables explicatives sont les données clients disponible de la base Orange. Les résultats sont très mauvais, Figure 3.4. Au total,

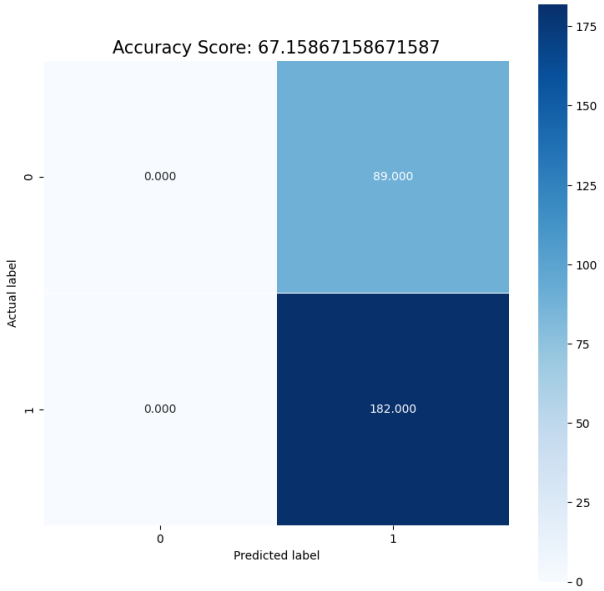


FIGURE 3.4 – Matrice de confusion du premier modèle

le modèle nous donne 80% d'hommes pour 20% de femmes, toutes les femmes ayant été détéctées par l'algortihme `gender-extractor`. La totalité des sexes inférés par le modèle logit sont masculins. Cela peut s'expliquer premièrement par la très faible quantité de données, et deuxièmement par le fait que le type d'abonnement téléphonique n'a probablement aucun lien avec le genre d'une personne. Une solution pour palier à ce problème de données aurait été d'ajouter les informations d'utilisations du téléphone que l'on récupère dans les compte rendus d'appels, qui sont des données au niveau des individus. Ces données-là sont potentiellement plus discriminante au regard du genre de l'individu que les données de la base client. Mais, comme expliqué dans la chapter 2, la variable qui devait servir de clé de jointure pour les tables (le numéro de téléphone) a été mal pseudonymiser. Enfin, le modèle choisi (la régression logistique) n'était peut être pas le meilleur modèle à utiliser dans ce cas là. J'aurais pu, par exemple, réaliser une Analyse Factorielle Discriminante avant de faire la régression logistique, ou bien entraîner un réseau de neurones.

Finalement, j'ai choisi de mobilisé un autre service, `Namsor` [41], qui fournit une probabilité de genre à partir du nom, prénom et pays d'un individu. Ce service propose une API qui permet de faire les

requêtes directement depuis Python. Il est difficile d'évaluer la pertinence des résultats ni de les comparer aux précédents en l'abscence de données réelles sur le genre. En revanche `Namsor` donne 60% d'hommes pour 40% de femmes, ce qui est plus proche de la répartition du pays qui est de 50% pour chaque genre. On peut également analyser les probabilités d'attribution de genre. Même si la plupart des attributions sont faites avec une grande probabilité, il reste 12% des données qui ont une probabilité inférieur à 0.80. Malheureusement le nombre de requêtes gratuites est limité à 5000. Cette solution est donc viable pour

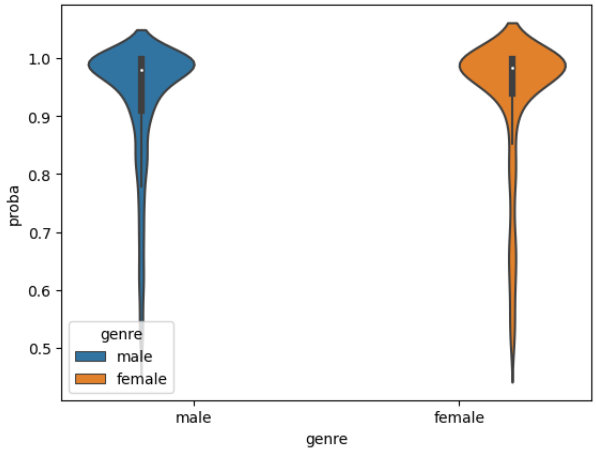


FIGURE 3.5 – Distribution des probabilités d'affectation par genre

mon échantillon de données, mais il faudra trouver autre chose pour passer à du Big Data ou payer le service !

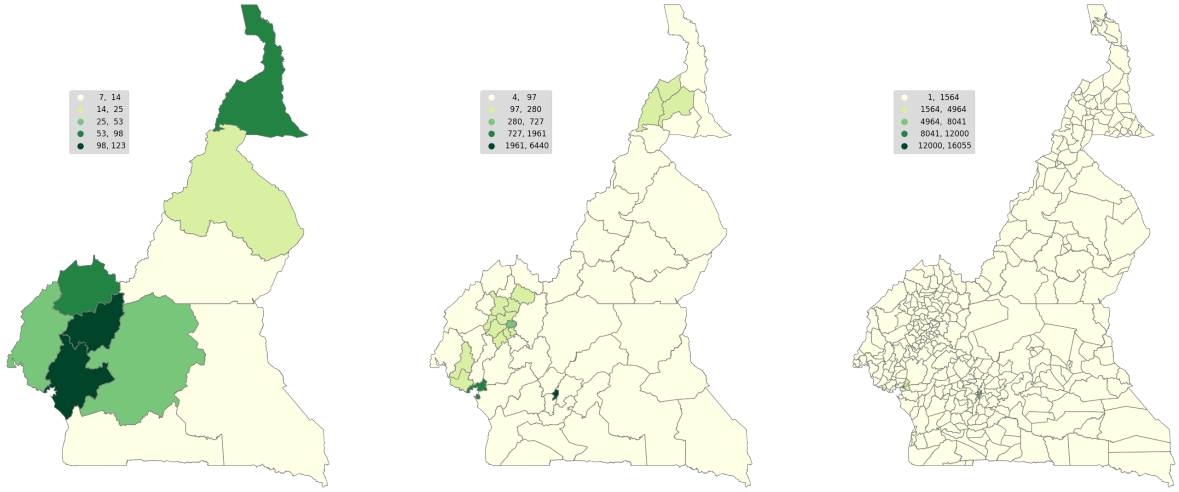
3.3 Analyse des données externes

Ces données externes vont servir, dans le chapter 4, à qualifier les noeuds pour comprendre la structure du réseau. Il est donc important d'en faire une analyse préalable, celle-ci permet en plus de mieux comprendre la dynamique du territoire camerounais.

3.3.1 Le recensement de 2005

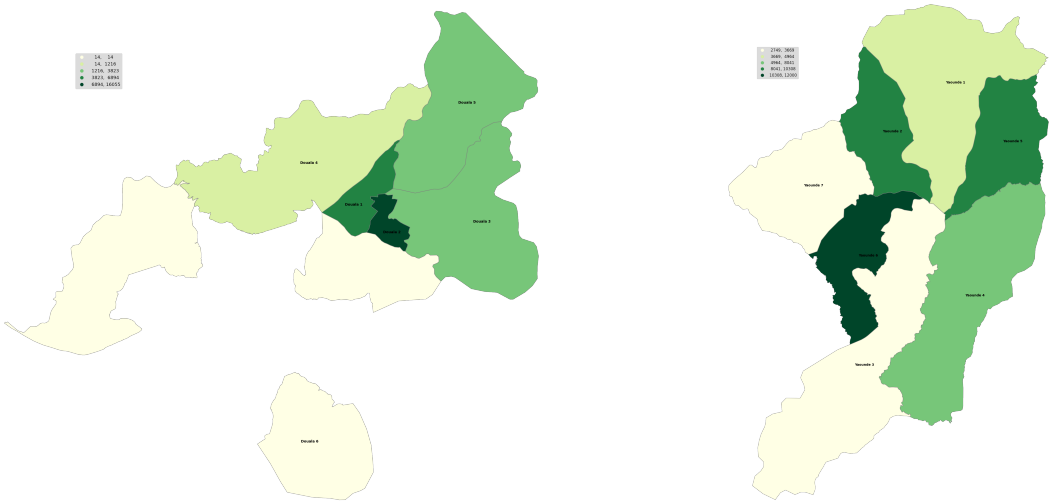
Le dernier recensement fait au Cameroun est ancien puisqu'il date de 2005. Certaines données sont plus récentes mais aucune n'atteint la précision de ce recensement. En effet, c'est le seul qui donne une estimation de la population au niveau des communes du pays.

On peut voir sur la Figure 3.6a que la population se concentre dans les régions autour de la capitale administrative Yaoundé et de la capitale économique Douala, ainsi que autour de la ville de Bamenda dans la région du Nord-Ouest. La carte des départements accentue cette observation, alors que la carte des communes est plus difficile à lire. En effet, les villes de Douala et Yaoundé sont beaucoup plus peuplées que les autres communes. Les deux villes comptent chacune plus de quatre millions d'habitants (en 2023), ce qui représente presque un tiers de la population totale du pays. Le sixième arrondissement de la ville de Douala est constitué d'une île et de deux rives. Une grande partie de l'arrondissement est constituée de l'écorégion des mangroves d'Afrique Centrale. Ce qui explique la très faible densité dans cet arrondissement.



(a) Régions (b) Départements (c) Communes

FIGURE 3.6 – Densité de la population Camerounaise, en habitants/km², suivant trois niveaux d'administration (2005).



(a) Douala (b) Yaoundé

FIGURE 3.7 – Densité de la population Camerounaise, en habitants/km², dans les deux villes les plus peuplées (2005).

On peut également analyser le déploiement du réseau mobile d'Orange Cameroun avec la densité de population du pays. Nous utilisons la carte régionale pour avoir une plus grande lisibilité. On remarque,

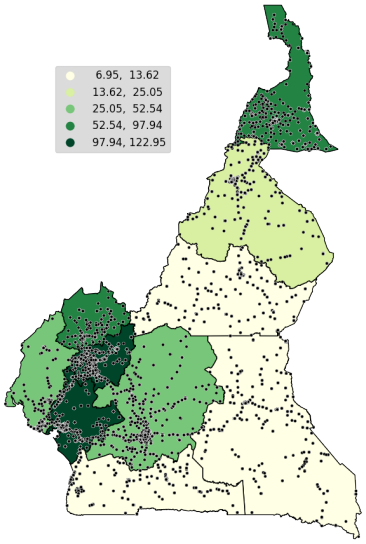


FIGURE 3.8 – Réseau d'antennes d'Orange Cameroun superposé à la densité de population, habitants par km^2, des régions (2005).

comme on peut s'y attendre, que les antennes sont concentrées autour des villes les plus densément peuplées. Notamment, autour des capitales Yaoundé et Douala, et des zones urbaines qui se développent dans la région de l'Ouest vers la région du Littoral.

3.3.2 Les données de l'enquête auprès des ménages

L'enquête complémentaire à la quatrième enquête auprès des ménages a rendu public l'accès à 42 indicateurs socio-économiques des communes, tels que le taux d'emploi, le taux de pauvreté, le taux d'alphabétisation ou encore le taux de scolarisation. Parmi toutes les données que j'ai réussi à récupérer, la plupart des indicateurs comptaient des données manquantes. J'ai commencé par inférer les données manquantes. Pour ne pas inférer des données à partir de trop peu de données je n'ai gardé que les indicateurs qui comportaient moins de 5% de données manquantes. C'est-à-dire les variables pour lesquelles les données étaient disponibles pour au moins 342 communes (sur 360). J'ai utilisé une méthode d'inférence spatiale, le kriging [24], pour interpoler ces données manquantes. C'est une méthode univariée, qui détermine à partir de l'autocorrélation spatiale de la variable, les poids à attribuer à chaque commune. Les poids sont également fonction de la proximité géographique des communes à la commune dont la valeur est manquante. Enfin, on attribue aux communes manquantes une valeur, qui est la moyenne pondérée des valeurs des autres communes. Ainsi, on obtient 33 indicateurs socio-économiques, définis pour toutes les communes du pays.

On peut alors calculer la matrice de corrélation de tous ces indicateurs, Figure 3.9. On observe, par exemple, que les taux d'emploi et taux de chômage (BIT et élargi) ne sont pas corrélés aux taux de scolarisation des enfants de 6 à 18 ans. En revanche, les indicateurs de pauvreté (taux de pauvreté, profondeur de la pauvreté et taux de pauvreté chronique) sont corrélés négativement aux taux de scolarisation des enfants de 6 à 18 ans. On remarque que les taux d'alphabétisation et les taux de scolarisation sont positivement corrélés, comme on peut s'y attendre. Et donc, également négativement corrélé aux indicateurs de pauvreté. À l'inverse, les indicateurs d'emploi et de chômage sont décorrélés des indicateurs de scolarisation et de pauvreté. On peut également noter une faible corrélation entre le pourcentage de la population de 15 ans et plus qui possède un mobile et le taux de scolarisation dans le secondaire (), le taux d'alphabétisation des 15-64 ans (), et le pourcentage de la population qui utilise l'électricité comme principale source d'éclairage (). Cela renvoie à l'étude [23] qui montre que l'accès à l'électricité peut être un moteur important pour l'inclusion digitale en Afrique. Enfin, on peut noter que les indicateurs de richesse (pourcentage de la population qui a accès à des installations sanitaires améliorées et non par-

tagées, utilise l'électricité comme source principale d'éclairage, utilise une source améliorée d'eau) sont corrélés négativement aux indicateurs de pauvreté (en moyenne).

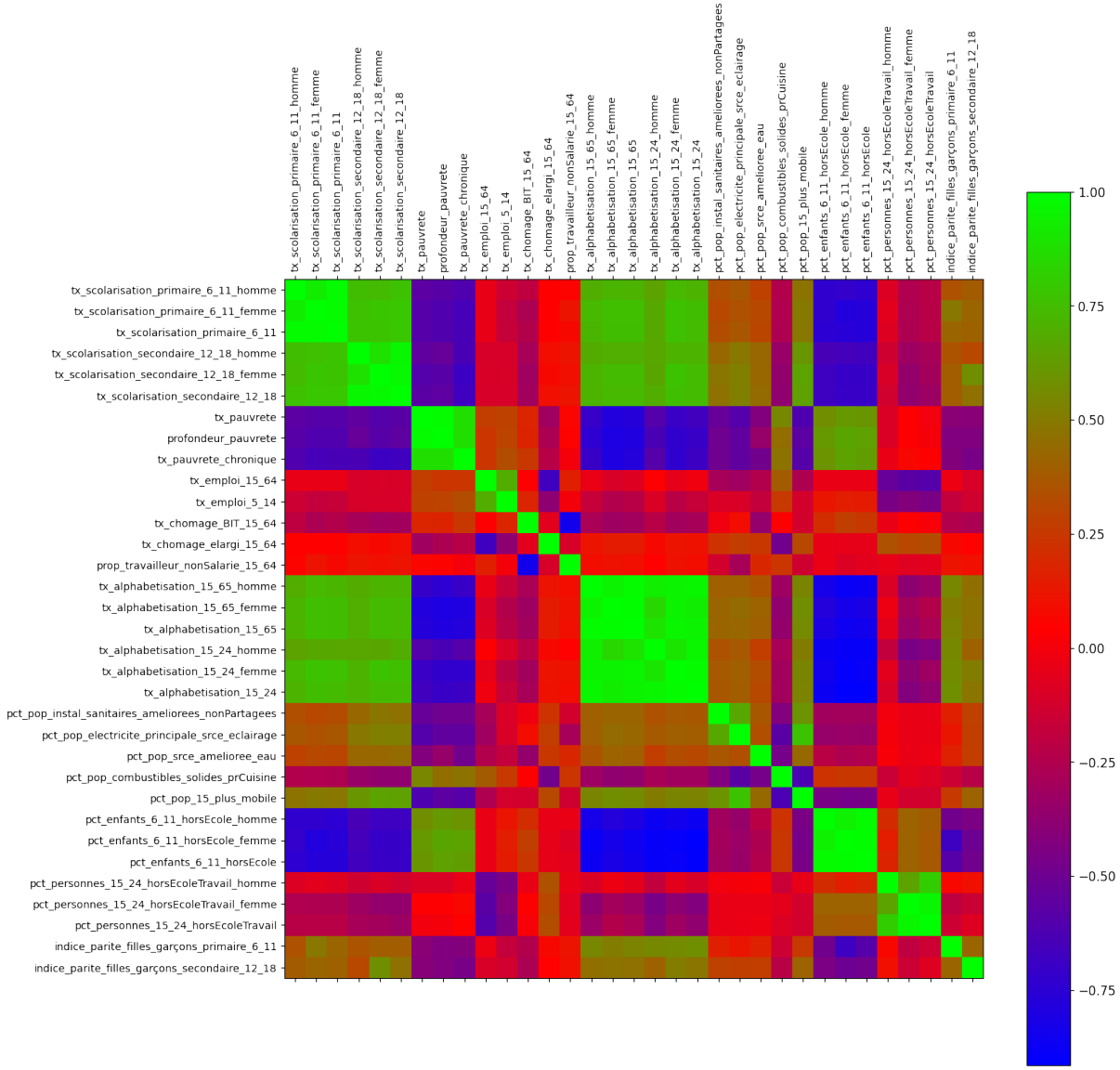


FIGURE 3.9 – Matrice de corrélation des variables de l'enquête auprès des ménages.

Nous pouvons aussi évaluer l'auto-corrélation spatiale de ces variables. C'est à dire, est ce que les communes ayant un fort taux d'emploi (par exemple) sont aléatoirement répartis dans le pays ou bien se concentrent elles dans des zones spécifiques. Cette analyse spatiale, univariée, des variables a été réalisée à l'aide du package `pysal` [54]. Avant toute analyse spatiale, il faut déterminer les relations de voisinage que l'on considère entre les communes. On peut représenter ces relations par un graphe. Deux communes sont reliées dans le graphe si elles s'influencent. En suivant la première loi de Tobler "everything is related to everything else, but near things are more related than distant things", nous allons considérer que deux communes s'influencent si elles sont géographiquement proche, si elles font parti du même département et si elles sont limitrophes. Ce qui nous donne le graphe d'influence des communes à la Figure 3.10. Ici

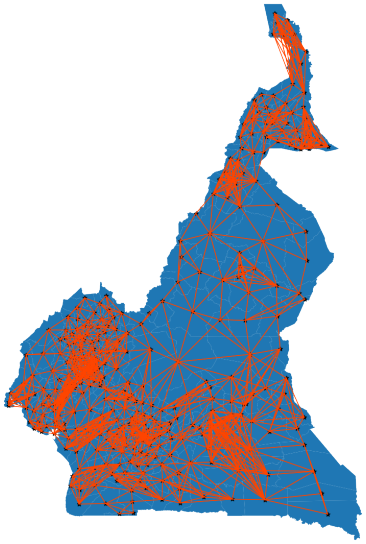


FIGURE 3.10 – Carte des poids spatiaux attribués aux communes

toutes les relations ont le même poids. Je ne présente ici que l'étude spatiale du taux d'alphabétisation des 15-64 ans, mais une analyse similaire est possible pour toutes les variables.

Taux d'alphabétisation des 15-64 ans

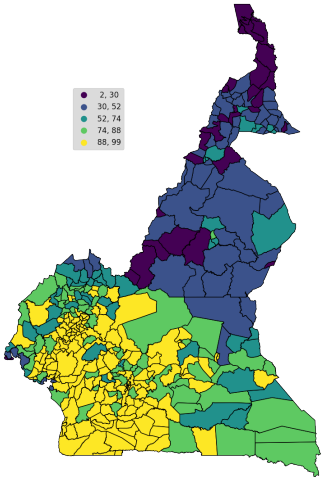
Commençons par regarder une carte de la répartition et la distribution du taux d'alphabétisation dans le pays. On peut voir sur la carte une ségrégation assez marquée entre le nord et le sud du pays. Et la Figure 3.11b montre une répartition bimodale du taux d'alphabétisation, avec un groupe comprenant un taux d'alphabétisation des 15-64 ans de plus de 50%, et l'autre moins de 50%.

Pour déterminer s'il y a une corrélation spatiale, nous pouvons regarder la différence de taux d'alphabétisation entre une commune et ses voisins (déterminés par les poids spatiaux) et entre une commune et toutes les autres, et comparer leur densité, Figure 3.12. Si les deux densité sont similaires, c'est que pour une commune donnée, ses communes voisines n'ont pas un taux d'alphabétisation similaire. En revanche, si la densité des différences entre une commune et ses voisins est bien distincte de la densité des différences entre une commune et toutes les autres, et que l'on observe un pic en zéro. C'est-à-dire que pour beaucoup de communes, ses voisins ont un taux d'alphabétisation similaire. Et on peut supposer une corrélation spatiale de la variable observée. C'est effectivement ce que l'on observe. On voit bien que la densité rouge, qui correspond à la différence de taux d'alphabétisation entre une commune et ses voisins, présente un pic en zéro distinct de la seconde densité.

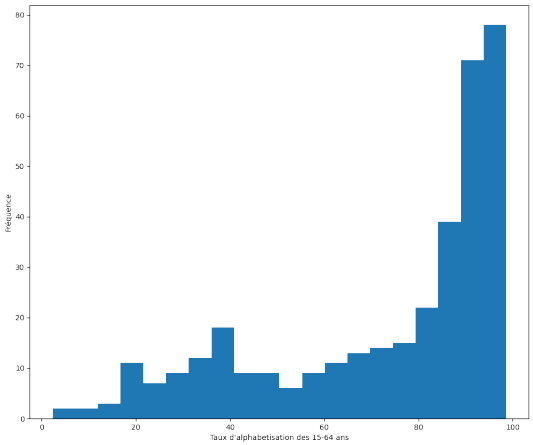
Une fois que l'auto-corrélation spatiale semble bien présente nous pouvons la mesurer grâce à l'indice de Moran [39] :

Définition 3.3.1 (Indice de Moran)**.**

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(x_i - \overline{x}\right) \left(x_j - \overline{x}\right)}{\sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2}$$



(a) Taux d'alphabétisation par commune au Cameroun.



(b) Histogramme du taux d'alphabétisation par commune.

FIGURE 3.11 – Répartition du taux d'alphabétisation des 15-64 ans, EC-ECAM, 2016.

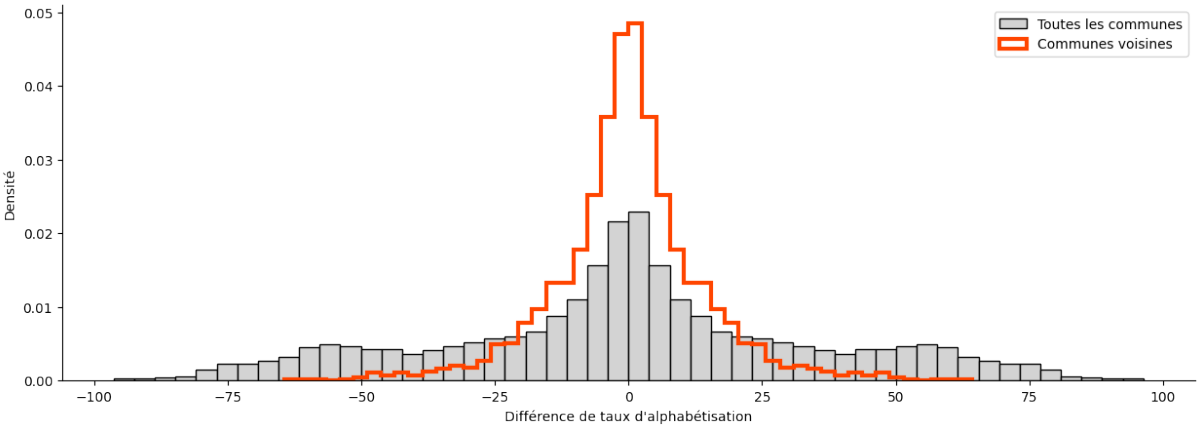


FIGURE 3.12 – Comparaison entre la densité des différences entre une commune et ses voisins (rouge) et la densité des différences entre une commune et toutes les autres (gris).

Avec :

- N : le nombre d'unité spatiale
- $W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$
- x : la variable d'interêt
- w_{ij} : l'élément ij d'une matrice de poids spatiaux avec des zéros sur la diagonale

Cette statistique est comprise entre -1 et 1. Pour $I = 0$, la variable est aléatoirement dispersée, pour $I = -1$, la variable est parfaitement dispersée et pour $I = 1$, la variable est parfaitement corrélée. L'indice de Moran peut également s'interpréter géométriquement. Il faut tout d'abord définir le décalage spatial (spatial lag). À chaque commune on attribue comme valeur de décalage spatial du taux d'alphabétisation, la moyenne des taux d'alphabétisation de tous ses voisins, définit par la matrice des poids spatiaux 3.10. L'indice de Moran correspond alors à la pente de la régression linéaire entre la variable d'intérêt (ici le taux d'alphabétisation) et sa variable de décalage spatial associée. On voit bien sur la Figure 3.13 qu'il y a une forte auto-corrélation spatial pour le taux d'alphabétisation des 15-64 ans, avec un indice de Moran élevé 0,83 associé à une faible p-valeur de 0.001. Cela veut dire que les communes ayant un taux d'alphabétisation similaire ont tendance à se trouver géographiquement proche entre elles. La tendance globale, est pour les communes ayant un taux d'alphabétisation élevé, d'être entouré par des communes ayant un fort taux d'alphabétisation, et inversement pour les communes avec un faible taux d'alphabétisation.

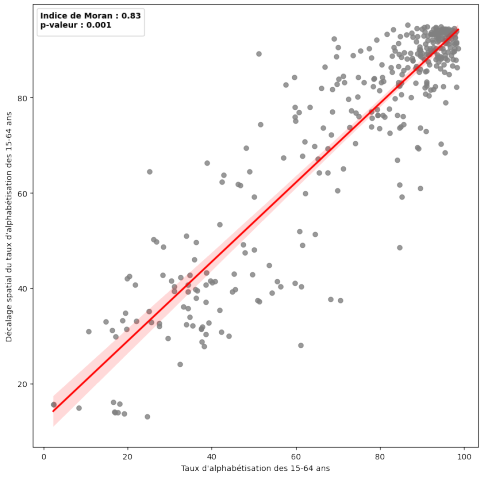


FIGURE 3.13 – Régression linéaire entre le taux d'alphabétisation et sa variable de décalage spatial associée.

Cette analyse spatiale participe à de mieux comprendre les dynamiques socio-économiques du pays, et ainsi aide à interpréter le graphe sociale, que ce soit sa structure ou sa dynamique.

Chapitre 4: L'étude d'un graphe social

Les données de télécommunications se représentent très bien sous la forme d'un graphe social ou les noeuds sont les individus et les arêtes les communications. C'est cette représentation des données que nous allons utiliser dans cette partie.

4.1 Social Network Analysis

Il convient de définir la notion de graphe, de certains indicateurs et de communautés dans un graphe avant d'appliquer ce modèle à nos données.

4.1.1 Définition

Un graphe se définit mathématiquement par un ensemble de noeuds et d'arêtes, $G = \{V, E\}$. V est l'ensemble des noeuds (*vertices*) et E est l'ensemble des arêtes (*edges*).

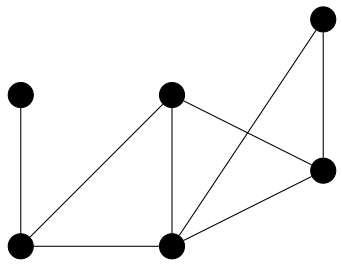


FIGURE 4.1 – Un exemple de graphe à 6 noeuds et 8 arêtes.

Les données de télécommunications permettent de savoir qui a créé l'appel et qui a reçu l'appel. Cette information se traduit sur le graphe par une direction des arêtes. Nous pouvons également pondérer ces

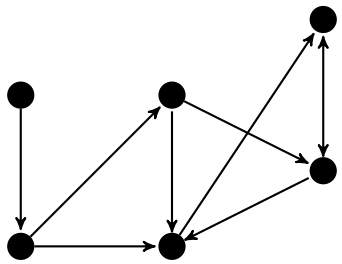


FIGURE 4.2 – Un exemple de graphe dirigé à 6 noeuds et 8 arêtes.

arêtes en ajoutant des poids et créer des arêtes qui bouclent sur des noeuds, Figure 4.3. Cela sera utile lorsque nous agrégerons les communications au niveau des communes dans la section 4.2. C'est un graphe pondéré, dirigé et qui accepte les boucles que nous allons utiliser.

4.1.2 Statistiques

Nous pouvons également définir plusieurs statistiques sur un graphe pondéré et dirigé. Les premières statistiques sont globales, elles concernent le graphe dans son ensemble.

L'*ordre* d'un graphe est le nombre total de noeud que possède ce graphe. L'ordre du graphe de la Figure 4.3 est 6. La *taille* d'un graphe est le nombre d'arêtes que possède ce graphe. La taille du graphe de la Figure 4.3 est 9. Enfin, la *taille totale pondérée* d'un graphe est la somme des poids de toutes les

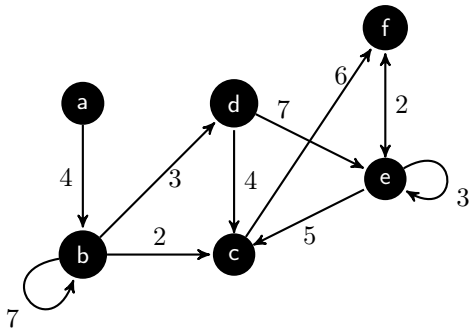


FIGURE 4.3 – Un exemple de graphe pondéré et dirigé avec des boucles.

arêtes. La taille totale pondérée du graphe de la Figure 4.3 est 43.

Des statistiques sur les noeuds sont, par exemple, le *degré*, qui représente le nombre de noeuds auxquels un noeud est connecté. Par exemple, le degré du noeud b de la Figure 4.3 est 3 (a, c, d). Et le *poids* d'un noeud est la somme des poids de toutes les arêtes d'un noeud. Le noeud b a un poids de 16 ($7+4+3+2$).

Pour un graphe qui accepte les boucles, nous pouvons distinguer les arêtes *intra-noeud* et *inter-noeuds*. Par exemple le noeud b possède une arête intra-noeud de poids 7, un degré inter-noeuds de (a, c, d) et un poids inter-noeuds de 9 ($4 + 3 + 2$).

Parmi les arêtes inter-noeud et pour un graphe dirigé, nous pouvons faire la distinction entre les arêtes *entrantes* et *sortantes*. Ainsi, pour le noeud b par exemple, nous avons un degré entrant de 4 et un degré sortant de 5 ($3 + 2$). Nous pouvons également définir le *poids entrant* et le *poids sortant* d'un noeud.

4.1.3 Indicateurs

Les indicateurs de centralité sont des statistiques calculées sur les noeuds qui indiquent leurs niveau d'importance dans l'ensemble du graphe. Nous nous concentrons sur trois de ces indicateurs. On considère une graphe G, d'orde n.

Définition 4.1.1 (Centralité de degré)**.**

$$\forall v \in V, C_d(v) = \frac{k}{n-1}$$

Avec k le degré du noeud v.

La centralité de degré détermine la connectivité d'un noeud au reste du réseau. C'est le ratio de son nombre de voisins par son nombre de voisins possible dans ce graphe. Cette métrique défini les noeuds qui sont les plus connecté au reste du réseau, les plus populaire. Ce sont ces personnes qui possèdent le plus d'information ou qui peuvent se connecter rapidement avec une grande partie du réseau.

Définition 4.1.2 (Centralité d'intermédiarité - *betweenness*)**.**

$$\forall v \in V, C_b(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Avec $\sigma(\cdot,\cdot)$, la fonction qui calcule le nombre de plus courts chemin entre deux noeuds, et $\forall v \in V, \sigma(\cdot,\cdot|v)$, la fonction qui calcule le nombre de plus court chemin entre deux noeuds passant par v.

Les noeuds avec un degré d'intermédiarité élevé jouent le rôle de vecteur de propagation. Ces noeuds sont les plus à même de propager l'information à une grande partie du réseau. Ils représentent souvent des 'ponts' entre des communautés. Cette métrique est couteuse en temps de calcul puisqu'il il faut calculer tous les plus courts chemins entre tous les noeuds du réseau.

Définition 4.1.3 (Centralité de proximité - *closeness*)**.**

$$\forall v \in V, C_c(v) = \frac{1}{\frac{1}{n-1}\sum_{u=1}^{a-1} d(u,v)}$$

Avec $d(\cdot,\cdot)$ la longueur du plus court chemin entre deux noeuds et $a-1$ le nombre de noeud atteignable depuis v, c'est à dire le nombre de noeud pour lesquels il existe un chemin qui mène de v à ce noeud.

Les noeuds avec une centralité de proximité élevé sont les noeuds qui sont capable d'influencer rapidement les noeuds auxquels ils sont connectés.

Ces trois mesures de centralités permettent d'identifier des acteurs clés du graphe. Avec les données de télécommunications, on peut, en théorie, identifier les individus qui sont, par exemple, le plus à même de connecter deux communautés.

4.1.4 Communautés

La notion de communautés repose uniquement sur les caractéristiques du graphe. Une communauté se définit par un ensemble de noeuds qui sont densément interconnectés et relativement moins intensément connectés avec le reste du réseau. Pour mesurer si un groupe de noeuds est plus densément connectés entre eux qu'avec le reste du graphe, nous allons utiliser la modularité, introduite par Newman [21].

Définition 4.1.4 (Modularité - pour un graphe pondéré, non dirigé)**.**
On considère un graphe $G = \{V, E\}$, d'ordre n.
On note A la matrice des poids.
On note également pour tout noeud i, $k_i = \sum_{j=1}^{n} A_{ij}$, le poids du noeud i.
Et enfin, on note $m = \frac{1}{2}\sum_{i,j\in[1,n]^2} A_{ij}$, la moitié de la taille totale pondérée du graphe.
La modularité s'écrit alors :

$$Q = \frac{1}{2m}\sum_{i,j\in[1,n]^2}\left[A_{ij} - \gamma\frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

Où $\forall i \in V, c_i$ est la communauté à laquelle appartient i. Et $\delta(x, y) = 1$ si $x = v$, 0 sinon. γ est le paramètre de résolution [44].

La modularité mesure la densité des liens à l'intérieur des communautés, comparativement aux liens entre les communautés. Elle permet donc de calculer pour un graphe donné et une partition en communautés donnée, si cette partition est une bonne partition ou non. La valeur de la modularité est comprise entre $[-1, 1]$.

Leicht et Newman [36] adapte la définition de la modularité aux graphes orientés en se basant sur l'observation suivante : Considérons deux noeuds u et v. u a un degré sortant élevé et un faible degré entrant. Inversement v a un degré entrant élevé et un faible degré sortant. Alors, la probabilité qu'il existe une arête de u vers v est plus élevée qu'une arête de v vers u. La prise en compte de cette observation donne la définition :

Définition 4.1.5 (Modularité - pour un graphe pondéré et orienté)**.**
On considère un graphe $G = \{V, E\}$, d'ordre n.
On note A la matrice des poids.
On note également pour tout noeud i, $k_i^{in/out} = \sum_{j=1}^{n} A_{ij}$, le poids du noeud entrant (in) ou sortant (out) du noeud i.
Et enfin, on note $m = \frac{1}{2}\sum_{i,j\in[1,n]^2} A_{ij}$, la moitié de la taille totale pondérée du graphe.
La modularité s'écrit alors :

$$Q = \frac{1}{m}\sum_{i,j\in[1,n]^2}\left[A_{ij} - \gamma\frac{k_i^{in} k_j^{out}}{m}\right]\delta(c_i, c_j)$$

Où $\forall i \in V, c_i$ est la communauté à laquelle appartient i. Et $\delta(x, y) = 1$ si $x = v$, 0 sinon. γ, le paramètre de résolution.

Ainsi, la modularité définit une fonction qui permet de mesurer la qualité d'une partition en communautés d'un graphe. Elle peut également servir de fonction à maximiser dans la recherche de communautés dans un graphe. C'est le principe sur lequel s'appuie de nombreux algorithme de détection de communautés.

Algorithme de Louvain

Cet algorithme a été publié en 2008 [7]. Il tire son nom de l'université de Vincent Blondel, l'Université catholique de Louvain. Cet algorithme à quelque peu révolutionné la détection de communautés, en effet sa rapidité a permis la détection de communautés dans des graphes de plus en plus important. Par exemple, l'algorithme de Louvain termine en deux minutes sur un graphe comprenant 2,6 millions de noeuds et 6,3 millions d'arêtes.
L'algorithme repose sur deux étapes, qui sont répétées itérativement.

- On commence par attribuer à chaque noeud une communauté différente. Il y a donc autant de noeud que de communautés. Puis, pour chaque noeud i, on considère tous ses voisins, j. On calcule alors le gain en modularité que l'on obtiendrait à passer i de sa communauté à la communauté de j. On place alors i dans la communauté qui maximise la modularité. Cette première étape s'arrête lorsque plus aucun changement de communauté n'améliore la modularité.
- La deuxième partie de l'algorithme consiste alors à créer un nouveau graphe, dont les noeuds sont les communautés détectées à la première étape, et les arêtes sont la somme des arêtes reliant deux communautés. Sur ce nouveau graphe, il est alors possible de répéter la première étape, de maximisation de la modularité. L'algorithme s'arrête lorsque plus aucun noeud ne change de communauté, on atteint un maximum de modularité.

Les performances de cet algorithme, en qualité et en temps, en ont fait une référence. Il est encore très utilisé aujourd'hui. Cet algorithme peut également s'utiliser sur des graphes orientés en utilisant la définition 4.1.5 [16].

Algorithme de Leiden

Cet algorithme s'appuie sur le principe de l'algorithme de Louvain. Il a été publié en 2019 par Vincent Traag et al [57]. Il reprend les phases une et deux de l'algorithme de Louvain, que sont le mouvement des noeuds et l'agrégation du graphe au niveau des communautés. Cependant, il ajoute une étape entre les deux.

Une fois que la partition $\mathcal{P}$ est trouvée à la fin de la première étape, au lieu de tout de suite générer le graphe aggrégé en s'appuyant sur cette partition, l'algorithme de Leiden va chercher à améliorer la partition $\mathcal{P}$ en une partition $\mathcal{P}_a$. Cette amélioration de la partition $\mathcal{P}$ se fait de la manière suivante. On commence par attribuer à chaque noeud sa propre communauté. On rapproche deux noeuds dans une même communauté s'ils sont déjà dans la même communauté dans la partition $\mathcal{P}$, et s'ils sont suffisamment bien connectés entre eux par rapport au reste de leur communauté dans la partition $\mathcal{P}$. Cette amélioration de la partition s'appuie également sur la modularité, mais au lieu de chercher les changements de communautés qui maximise la modularité comme l'étape 1, les changements de communautés sont aléatoirement choisis parmi tous ceux qui augmentent la modularité, avec une probabilité d'autant plus élevé que si le gain en modularité est élevé. Cet aléatoire permet d'explorer un peu plus l'ensemble des partitions possibles. La partition $\mathcal{P}_a$ divise la partition $\mathcal{P}$ en sous communautés. L'étape d'agrégation se fait alors sur la partition $\mathcal{P}_a$, chaque communauté de cette partition devient un noeud. Et la communauté à laquelle appartient ces noeuds (agrégé) est déterminé par la partition $\mathcal{P}$. Une illustration de cette amélioration est présentée dans la Figure 4.4.

L'algorithme de Leiden implémente également une queue de suivie lors de l'étape 1. Cela sert à ne pas tester tous les noeuds lorsque l'on maximise la modularité, mais seulement les noeuds qui peuvent potentiellement augmenter la modularité. Cette petite différence participe à améliorer les performances de l'algorithme de Louvain, qui est déjà rapide.

Il est également possible, comme dans l'algorithme de Louvain, d'utiliser la modularité orientée. Cet algorithme semble garantir plusieurs caractéristiques sur les communautés qui assure que ces dernières ont une réelle signification. Plus de détails sont disponible dans le papier [57] et les annexes.

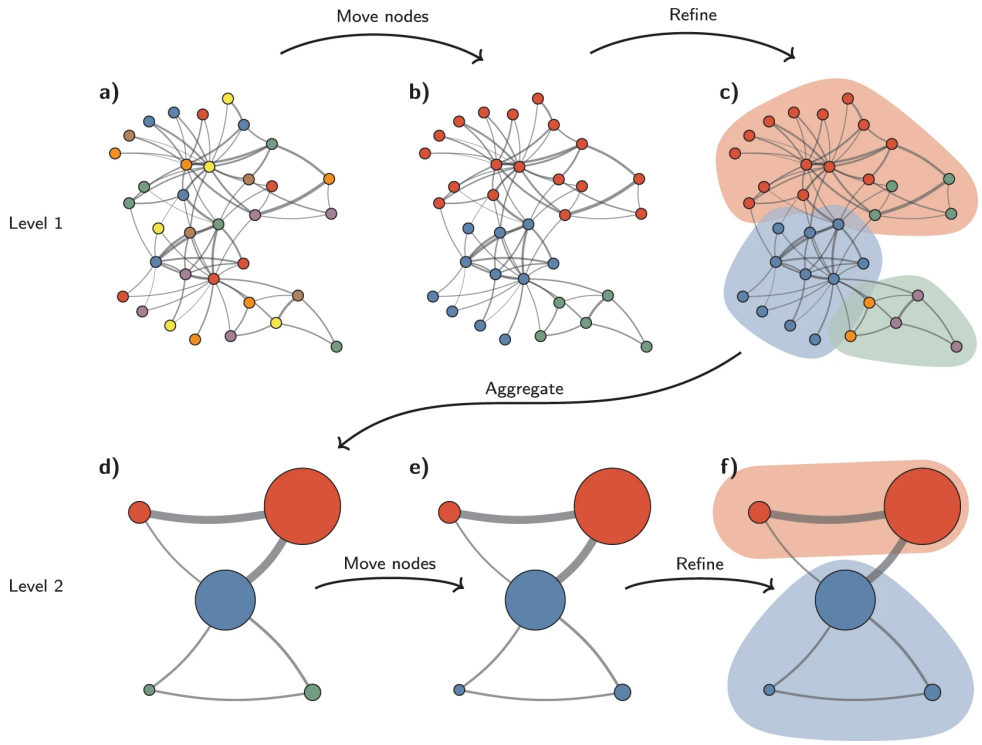


FIGURE 4.4 – Illustration de l'algorithme de Leiden, Figure 3 du papier *From Louvain to Leiden : guaranteeing well-connected communities* de Traag, Waltman et van Eck [57].

4.2 Création du graphe

À partir des données au format de la Table 2.7 je peux construire deux types de graphe sociaux. Le premier porte sur les individus eux même. En effet, la granularité de ces données est très fine. Cela à l'avantage de pouvoir travailler sur des données très précise, mais également de pouvoir travailler, s'il le faut, avec un niveau hiérarchique plus élevé. C'est le cas ici, ou l'on va travailler au niveau des communes du pays. Il faut donc aggréger les communications au niveau des communes. Les noeuds du graphe social représenterons alors les communes et les arêtes les appels téléphoniques qui ont été passé depuis une commune vers une autre. J'ai choisi de me concentrer sur un graphe qui utilise les communes comme noeud, à la fois parce que cela est plus précis qu'au niveau des départements ou des régions, et parce que les données de l'enquête auprès des ménages Camerounais sont au niveau des communes. Ce niveau hiérarchique m'a semblé un bon compromis entre une granularité suffisamment fine, ainsi qu'une quantité et une diversité des données mobilisées satisfaisante.

4.2.1 Préparer les données

Pour réaliser ce graphe il faut d'abord déterminer pour chaque appel la commune depuis laquelle a été émis l'appel et la commune qui a reçu l'appel. À partir des identifiants des antennes il est possible de déterminer ces communes. J'ai commencé par attribuer, pour chaque antenne de la Table 2.7 qui est aussi présente dans la table de référence des antennes, ses coordonnées (longitude et latitude). Ensuite, il restait quelques antennes que je ne retrouvais pas dans la table référence. Cela représente tout de même plus d'un million de lignes sur les 5 millions. Pour les prendre en compte, j'ai utilisé le code LAC de ces antennes. En effet, ce code permet de déterminer une zone relativement précise de la localisation de l'antenne, sans avoir sa localisation exacte. Pour ces antennes, j'ai sélectionné toutes les antennes de la table de référence qui partageaient le même code LAC, et attribuer le barycentre de ces antennes, comme localisation. Cela m'a permis de récupérer la localisation de 1 millions d'appels en plus.

À partir de ces coordonnées je suis en mesure de déterminer dans quelle commune du Cameroun se trouvait les individus au moment de leur appel. En regroupant les appels par le couple commune de l'émetteur, commune du récepteur, j'obtiens la matrice des arêtes, Table 4.1. Ces matrcies représente à elle seule toutes les informations du graphe social, représenté Figure 4.5 On remarque tout de suite que les villes qui sont les plus importantes sont Douala et Yaoundé. Ensuite viennent les communes urbaines

Commune de l'émetteur	**Commune du récepteur**	**Poids de l'arête**
Douala 1	Yaoundé 6	3 653
Maga	Yaoundé 3	172
Douala 5	Loum	784
Yaoundé 6	Douala 1	2 943

TABLE 4.1 – Exemple de la matrice des arêtes qui constitue le graphe social.

comme Maroua 1 ou Ngaoundere 2.

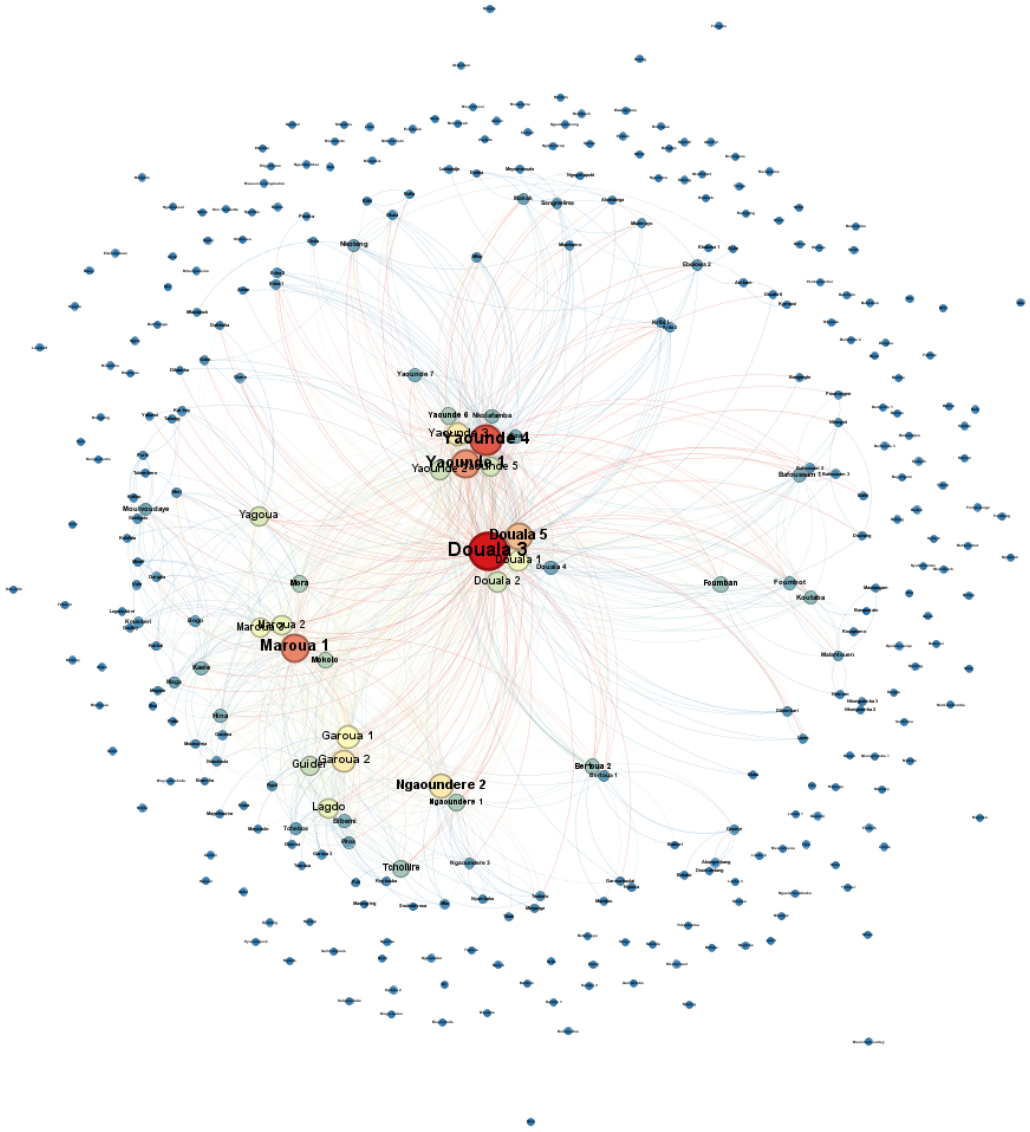


FIGURE 4.5 – Visualisation du graphe social aggrégé au niveau des communes.

(a) Image réalisée avec Gephi [20]. La couleur et la taille des noeuds représentent le degré total des noeuds. Seul 2,56% des arêtes les plus importantes sont visibles. La représentation du graphe a été obtenu avec l'algorithme Force-Atlas 2, disponible dans Gephi.

4.2.2 Enrichir les noeuds du graphe

Tout l'intérêt de réaliser le graphe à une granularité plus grande, les communes, est de pouvoir caractériser les noeuds (les communes) du graphe avec des informations externes. Les premières données que nous pouvons utiliser comme attribut des noeuds sont les coordonnées des barycentres des communes. Cela est nécessaire pour représenter le graphe sur une carte grâce au package `osmnx` [51]. Nous pouvons ensuite utiliser toutes les statistiques (indicateurs de centralité, nombre d'appels intra-commune, les degrés entrant, sortant...) produites sur les noeuds. De plus, nous pouvons caractériser chaque noeud par la commune à laquelle il appartient. Enfin, nous pouvons utiliser toutes les données externes issues de l'enquête auprès des ménages Camerounais ainsi que le recensement.

Pouvoir caractériser les noeuds à la fois par leurs communautés et des critères socio-économiques, va nous permettre de spécifier les caractéristiques de certaines communautés.

4.3 Analyse du graphe

4.3.1 Statistiques exploratoires

Nous commençons par regarder quelques statistiques simples du graphe que l'on étudie. Tout d'abord le graphe possède 318 noeuds. Nous n'avons pas l'ensemble des 360 communes du pays, parce que toutes les communes n'ont pas été à l'origine ni n'ont été la destination d'un appel, dans ma fenêtre de données. Ensuite, la taille du graphe est de 45159, et la taille totale pondérée, cela équivaut au nombre d'appels, est de 4979799. Nous pouvons également calculer la proportion d'appels qui se font inter-communes et intra-communes.

On a 62% des appels qui ont été réalisés entre deux communes différentes (inter-communes). Donc, 37% des appels ont été fait à l'intérieur d'une même commune (intra-commune).

Commune de l'émetteur	Commune du récepteur	Poids de l'arête
Douala 5	Douala 3	28863
Douala 3	Douala 5	27890
Garoua 2	Garoua 1	24054
Garoua 1	Garoua 2	23865
Ngaoundere 2	Ngaoundere 1	20416
Ngaoundere 1	Ngaoundere 2	19588

(a) inter-communes.

Commune de l'émetteur	Commune du récepteur	Poids de l'arête
Douala 3	Douala 3	116 638
Yaoundé 4	Yaoundé 4	86 500
Maroua 1	Maroua 1	85 208
Douala 5	Douala 5	70 251
Garoua 1	Garoua 1	68 706
Garoua 2	Garoua 2	66 338

(b) intra-communes.

TABLE 4.2 – Listes des communes ayant le plus d'appels sortant...

On vérifie une fois de plus que ce sont bien les communes urbaines qui communiquent le plus, avec les communes de Douala 3 et 5 comme principales communes.

En moyenne une commune à un degré de 284 et un poids de 31 319. Une commune est donc en contact, en moyenne avec 284 autres communes, et a communiqué 31 319 fois avec une autre commune. La Table 4.3 présente les trois communes avec les plus grands et les plus petits degré et poids. À noter que

Commune	Degré
Douala 3	606
Yaoundé 4	600
Yaoundé 1	593
Fungom	3
Bum	2
Menchum-valley	1

(a) Degré

Commune	Poids
Douala 3	547 676
Yaoundé 4	461 566
Maroua 1	358 699
Fungom	3
Bum	2
Menchum-valley	1

(b) Poids

TABLE 4.3 – Listes des communes classées par degré et par poids.

le degré d'un noeud peut être supérieur à 317 puisque le degré compte à la fois le degré entrant et sortant.

Nous pouvons distinguer ces deux cas. En moyenne, une commune possède un degré entrant de 142, c'est à dire qu'elle reçoit des appels de 142 communes différentes. Le poids entrant moyen est lui de 15 659. Enfin, en moyenne, une commune possède un degré sortant de 142, c'est à dire qu'elle appelle en

Commune	Degré
Douala 3	302
Yaoundé 1	301
Yaoundé 4	298
Fungom	1
Menchum-valley	1
Bum	0

(a) Degré entrant

Commune	Poids
Douala 3	273 619
Yaoundé 4	229 721
Maroua 1	182 868
Fungom	1
Bum	1
Menchum-valley	0

(b) Poids entrant

TABLE 4.4 – Listes des communes classées par degré et poids entrant.

moyenne 142 communes différentes. Le poids sortant moyen est lui de 15 659.

Commune	Degré
Douala 3	304
Yaoundé 4	302
Yaoundé 1	292
Nkum	1
Menchum-valley	1
Muyuka	0

(a) Degré sortant

Commune	Poids
Douala 3	274 057
Yaoundé 4	231 845
Maroua 1	180 909
Fungom	1
Bum	1
Menchum-valley	0

(b) Poids sortant

TABLE 4.5 – Listes des communes classées par degré et poids sortant.

Intéressons nous maintenant aux indicateurs de centralité. On présente à la Table 4.6 les trois indicateurs de centralités, à chaque fois les trois valeurs les plus élevées et les trois les plus faibles. La centralité de degré indique si un noeud est très connecté, de manière général, au reste du graphe. Comme on peut s'y attendre les communes qui possèdent des degré et poids élevé (totaux, entrants et sortants), possèdent également une centralité de degré élevée. Ce sont ces communes-là qui possèdent le plus d'information. Si l'on veut pouvoir savoir tout ce qu'il se passe dans le reste du pays, il est préférable de se trouver dans une de ces communes. La centralité de proximité nous indique également que ce sont ces même communes qui influencent le plus leurs communes voisines. Leurs communications sortantes élevées, Table 4.5b, renforcent cette idée.

La centralité d'intermédiarité, Table 4.6b indique quelque chose de nouveau. En effet, les trois communes ayant une centralité d'intermédiarité plus élevée que les autres sont Oku, Kribi et Doumaintang. Ces valeurs sont difficiles à interpréter car, premièrement ce ne sont pas des communes remarquables. Il n'y a pas une population particulièrement importante par exemple. Pour pouvoir mieux interpréter ces valeurs, il faudrait sûrement une meilleure connaissance historique, géographique, économique, géopolitique et culturelle du pays - par exemple, la place des tribus dans l'organisation sociale [55]. La commune de Oku se trouve au bord de la zone de guerre dans les régions du Nord-Ouest et du Sud-Ouest, en raison de la crise anglophone qui dure depuis 2017. Un surplus d'appels de toutes les communes pour prendre des nouvelles de la guerre pourrait expliquer pourquoi cette commune, en particulier, joue le rôle de vecteur de propagation ou de ponts entre plusieurs communautés. De par sa localisation, cette commune serait une plaque tournante de la propagation d'information au Cameroun ? Une analyse plus poussée est nécessaire pour appréhender les dynamiques à l'oeuvre.

4.3.2 Analyse de l'homophilie

L'homophilie, *assortativity* en anglais, est un concept introduit dans les graphes sociaux par Newman en 2003 [45]. L'homophilie mesure si les noeuds portant une certaine caractéristique sont plus à même

Commune	Degré
Douala 3	1.91
Yaoundé 4	1.89
Yaoundé 1	1.87
Nkum	0.009
Menchum-valley	0.006
Muyuka	0.003

(a) Centralité de degré

Commune	Degré
Oku	0.0096
Kribi 2	0.0089
Doumaintang	0.0085
Muyuka	0
Bum	0
Menchum-valley	0

(b) Centralité d'intermédiarité

Commune	Poids
Douala 3	0.95
Yaoundé 1	0.94
Yaoundé 4	0.93
Fungom	0.45
Menchum-valley	0.41
Bum	0

(c) Centralité de proximité

TABLE 4.6 – Indicateurs de centralité pour les 3 communes les plus élevé et les 3 les plus faible.

de plus communiquer entre eux qu'avec des noeuds qui ne partage pas cette caractéristique. La valeur de l'homophilie varie de -1 à 1. Une homophilie de 1 pour une variable X signifie que les noeuds ayant une valeur similaire de X entre eux ont tendance à être très connecté, tandis que les neouds qui ont une valeur de X différente ont tendance à moins communiquer. À l'inverse, une homophilie de -1 indique que les noeuds avec une valeur de X élevée ont tendance à être plus connecté à des noeuds ayant une valeur de X faible, et inversement.

Cette homophilie peut se calculer aussi bien sur des variables quantitatives que sur des variables qualitatives. La Table 4.7 donne l'homophilie pour quelques attributs des noeuds. Par exemple, les communes

Attribut	Homophilie
Taux d'alphabétisation des 15-64 ans	0.075
Taux de scolarisation au secondaire des 12-18 ans	0.066
Part de la population qui utilise l'électricité comme source principale d'éclairage	0.0071
Part de la population qui a accès à une source améliorée d'eau	0.0031
Degré sortant	-0.312
Degré entrant	-0.313
Centralité de degré	-0.3145

TABLE 4.7 – Valeur de l'homophiliepour quelques attributs choisis.

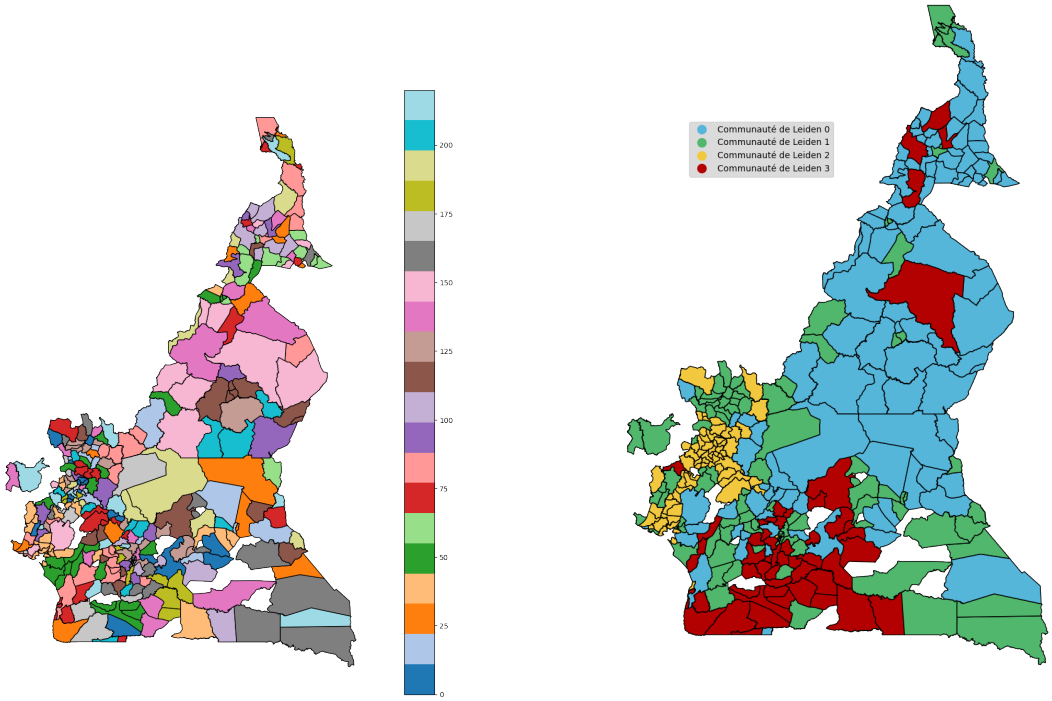
ayant une grande part de la population qui a accès à une source améliorée d'eau ou qui utilise l'électricité comme principale source d'éclairage, n'ont pas tendance à appeler d'avantage les communes qui ont un taux similaire ou différent, leur homophilie est proche de zéro (0,007 et 0,003 respectivement). Ces deux variables ne sont donc pas des variables discriminantes dans l'étude des interactions des communes. En revanche, les degré sortant, entrant et la centralité de degré ont tous trois une forte homophilie négative. L'homophilie de la centralité de degré (-0,3145) indique que les communes ayant une centralité de degré élevées ont tendance à plus communiquer avec les communes qui ont une faible centralité de degré, qu'avec les autres. En effet, les communes avec une grande centralité de degré vont plus communiquer avec beaucoup de communes ayant une moindre importance, pour recevoir et distribuer de l'information. Enfin, l'homophilie des degrés sortant et entrant nous indique un phénomène similaire. L'homophilie étant négative, elle s'interprète de la manière suivante. Les communes qui lancent les appels, les lancent vers des

communes ayant un degré sortant faible. Et les communes qui reçoivent beaucoup d'appels, les reçoivent de la part de commune qui en reçoivent moins. Finalement, en schématisant, il y a une tendance qui est que certaines communes ne font presque qu'appeler, tandis que d'autres communes ne font que recevoir des appels. Typiquement on peut faire l'hypothèse que les communes rurales attendent de recevoir des appels provenant des grandes villes pour téléphoner. Sans doute, les populations dans ces communes sont moins riches et peuvent donc moins se permettre d'appeler. Les dynamiques à l'oeuvre doivent être approfondies.

4.3.3 Étude des communautés

Commençons par l'étude des communautés issues de l'algorithme de Louvain, présentées Figure 4.6a.

Cet algorithme détecte 220 communautés. Cela pose tout d'abord un problème de représentation. En effet, il est assez difficile de représenter ces 220 communautés avec 220 couleurs distinctes. Un dégradé ne fonctionne pas non plus, puisque l'algorithme ne garantit en rien que les numéros des communautés aient un rapport avec leurs caractéristiques. De plus, étudier 220 communautés sur 318 noeuds revient presque à étudier chaque noeud séparément, ce qui annihile tout l'intérêt de détecter des communautés. Regardons alors les communautés de Leiden, Figure 4.6b.



(a) Communautés de Louvain.

(b) Communautés de Leiden.

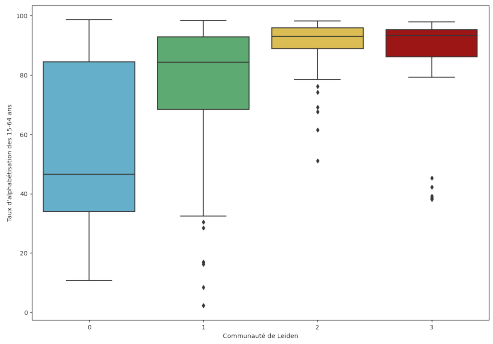
FIGURE 4.6 – Détection de communautés avec deux algorithmes différents.

L'algorithme de Leiden, révèle quant à lui, quatre communautés. La communauté 0, en bleu, compte 102 communes qui se trouvent plutôt au nord du pays, dans les régions de l'Extrême-Nord, du Nord et d'Adamaoua. La communauté 1, en vert, compte 100 communes. On y retrouve plusieurs grandes villes, comme Douala, Yaoundé ou encore Garoua et Ngaoundéré. La communauté 2, en jaune, compte 65 communes, qui se concentrent dans la région de l'Ouest, autour des villes de Bamenda, Bafoussam, Bafang ou encore Mbouda. Enfin, la communauté 3, en rouge, compte 51 communes, qui se trouvent, on le voit sur la carte, autour de la ville de Yaoundé.

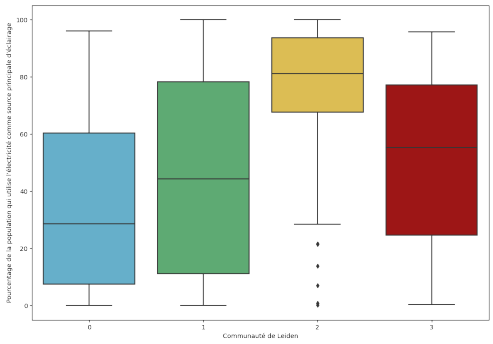
La carte fait clairement apparaitre au travers de la communautés 2, la zone anglophone à l'Ouest. De plus selon une étude de l'université de Laval, "[D]ans l'Ouest, on trouve surtout les Bamouns et les Bamilékés, un peuple qui doit sa notoriété à son dynamisme économique et à son expansion spatiale" [55]. Nous pourrons voir plus loin que cette phrases se vérifie dans les données.

Mais commençons par analyser la population à l'intérieur de ces communautés. Comme l'on pouvait s'y attendre, c'est bien la communauté 1 qui compte la plus grande population avec 6 600 000 individus, soit 40% de la population Camerounaise. Rappelons que les chiffre concernant la population datent de 2005. La communauté 0 est la deuxième plus peuplée avec 26% de la population, viennent ensuite les communautés 2 et 3 avec respectivement 21% et 11% de la population.

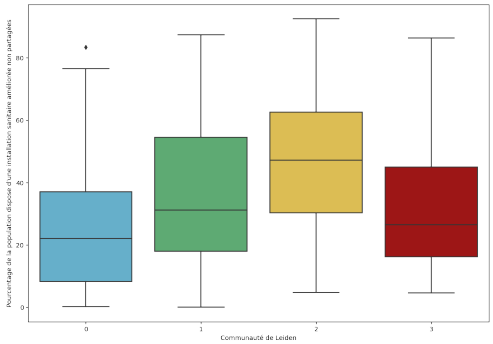
Maintenant, nous pouvons utiliser les données de l'enquête auprès des ménages pour caractériser ces communautés. Regardons, Figure 4.7a, le taux d'alphabétisation par communautés. Nous observons que la communauté 0, en bleu, possède un taux d'alphabétisation nettement plus faible que les autres communautés. Cela vient appuyer notre analyse spatiale de cette variable faites en Figure 3.3.2. En effet, la communauté 0 se trouve plutôt au nord du pays, comme les communes ayant un faible taux d'analphabétisation. En supposant que les communes les plus riches sont plus alphabétisées que les communes plus pauvres, on peut dire que l'algorithme de Leiden sépare les communes par niveaux de richesse. Avec les communautés 2 et 3 les plus riches. Voyons si cette analyse se confirme en regardant analysant les variables d'accès à l'électricité, des sanitaires améliorées et au mobile. On voit bien que la communautés



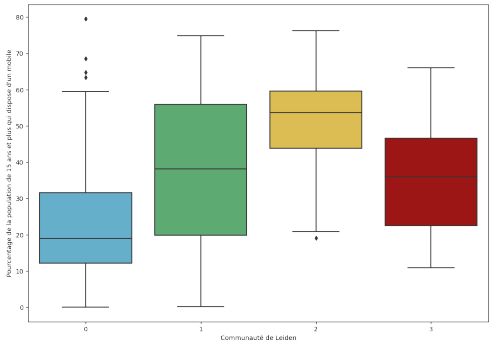
(a) Taux d'alphabétisation



(b) Accès à l'électricté



(c) Accès à des sanitaires améliorés



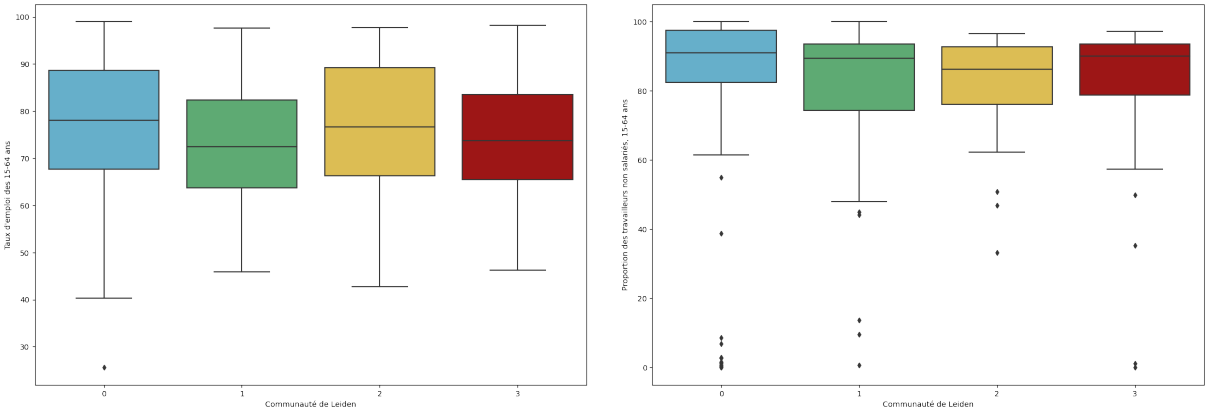
(d) Accès à un téléphone mobile

FIGURE 4.7 – Répartition, par communautés de Leiden, de quatres variables qui informent sur le niveau de richesse.

2 possède un taux d'utilisation de l'électricité comme source d'énergie principale bien plus élevée que les autres. On observe également, bien que cela soit moins marqué, que en moyenne, les communes de la communauté 0 possède un taux d'utilisation de l'électricité comme source d'énergie principale plus faible que toutes lescommunautés 1 et 3. Les même observations peuvent être faites sur la Figure 4.7c, bien que les différences soient moins marquées. Nous pouvons également regarder la proportion de la population de 15 ans et plus qui possède un mobile, Figure 4.7d. On voit bien que la communauté 0 est bien moins équipée pour utiliser le réseau télécoms que la communauté 3. En sachant que le coût d'un mobile au Cameroun est élevé, cette variable renforce l'idée que les communautés de Leiden discrimine les communautés par niveau de richesse.

Cependant, si l'on regarde la Figure 4.8a, on voit que le taux d'emploi des communes ne rentre pas en compte dans la détermination de leur richesse. Une explication de ce phénomène est que la population pauvre doit travailler pour survivre, et essentiellement dans le secteur informel. Le taux d'emploi n'est pas une variable adéquat pour déterminer la richesse des communes, du fait du travaille informel très

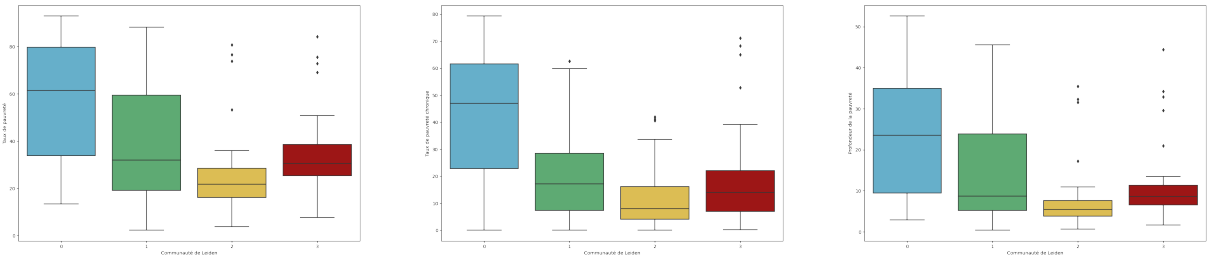
répandu dans ce pays. Une autre variable que l'on peut examiner, est la proportion dans population des



(a) Taux d'emploi par communauté de Leiden. (b) Taux de non salariée, par communautés de Leiden.

FIGURE 4.8 – Répartition, par communautés, du taux d'emploi et du taux de non salarié.

15-64 ans des non salarié. On voit sur la Figure 4.8b que la communauté 0 possède une proportion légèrement plus élevée que celle des autres communautés, mais cela n'est pas remarquable. L'enquête auprès des ménages mets à nous fournit rois variables, la pauvreté, la pauvreté chronique ainsi que la profondeur de la pauvreté. On remarque bien que la communauté 0 est bien plus pauvre que les autres sur les trois

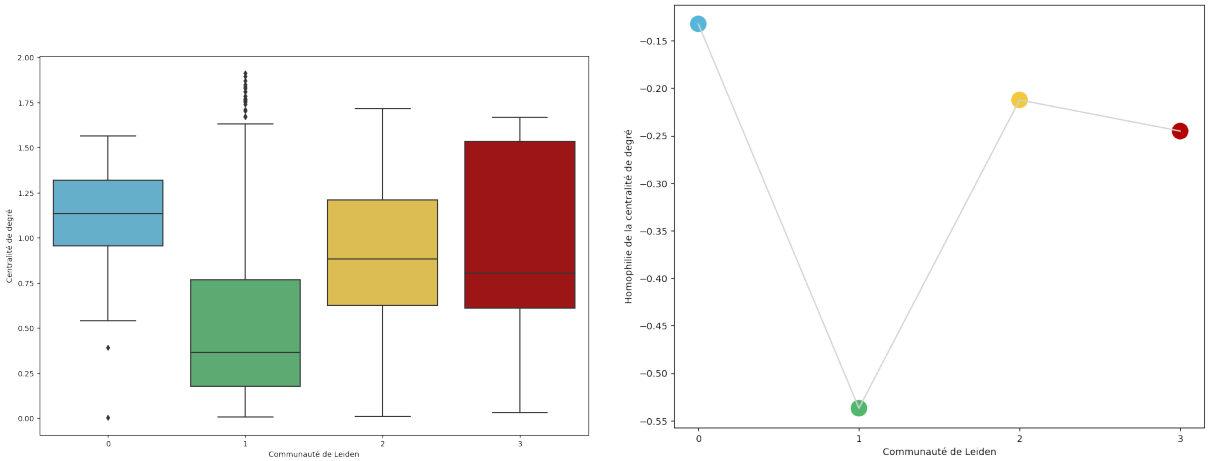


(a) Taux de pauvreté (b) Taux de pauverté chronique (c) Profondeur de la pauvreté

FIGURE 4.9 – Étude de la pauvreté par communautés de Leiden.

indicateurs. On remarque également que la communauté 2, en jaune, semble la moins pauvre. Cela rejoint les propos de Lionel Jean, de l'université de Laval, déjà évoqués plus haut. La communauté 2, qui se situe dans l'Ouest du pays, est bien une communauté riche, voir très riche. On sait que cette communauté est majoritairement anglophone. L'algorithme de Leiden a permis de détecter des communautés sur leurs richesse mais également sur la langue majoritairement utilisée dans les communes.

Enfin, nous pouvons examiner l'indicateur de centralité de degré par communautés, Figure 4.10a. Cet indicateur ne semble pas discriminer les communautés 0, 2 et 3. On remarque tout de même que la communauté 1 possède une centralité de degré nettement inférieur aux autres communautés. Elle possède de nombreuses valeurs extrêmes. On sait que la communauté 1 regroupe un grand nombre de commune, dont les communes qui constituent les villes de Yaoundé et Douala. Or ces deux villes concentrent presque un tier de la population totale du pays, comme vu dans la subsection 3.3.1. Finalement, la centralité de degré ne permet pas de caractériser les communautés. Mais l'étude d'une communauté peut nous donner des renseignements. En effet, si l'on regarde l'homophilie de la centralité de degré par communautés, Figure 4.10b, on voit que la communauté 1 possède une homophilie de -0,5. L'analyse globale faîtes dans



(a) Centralité de degré

(b) Homophilie de la centralité de degré

FIGURE 4.10 – Analyse de la cenralité de degré par communautés de Leiden.

la subsection 4.3.2 est renforcée par l'analyse par communautés. Cela veut dire que les petites communes communiquent essentiellement avec les villes de Douala et Yaoundé, et ces deux villes communiquent plus avec les communes rurales qu'entre elles. On a un effet de "hub" de ces deux villes, à la fois au niveau du pays, mais encore plus avec les communes avec lesquelles Yaoundé et Douala sont le plus en contact.

Conclusion

Ce stage avait deux objectifs principaux. Le premier était de vérifier la qualité et l'intégrité des données. Cela s'est avéré très utile puisque je me suis rendu compte que les données envoyées par l'équipe d'Orange Cameroun comprenaient plusieurs défauts. D'abord, les données ne semblaient pas complètes. Les chiffres que l'on peut tirer de la base de données Orange Money semblent loin de ceux publiés dans la presse. La table de référence des antennes ne référence pas la totalité des antennes en activité du réseau. Cette table est très souvent modifiée. Récupérer une version fiable et à jour s'avère complexe. Ensuite, la principale base de données, les comptes rendus d'appels, semble mal renseignée les numéros de téléphone. Cela a fortement compliqué la compréhension des données et l'extraction des données dans un format utilisable pour réaliser un graphe social. Enfin, la disponibilité de données complémentaires, comme les données de recensement et les données de l'enquête complémentaire à la quatrième enquête auprès des ménages Camerounais, est rare ou bien les données sont anciennes.

Mais ces données permettent tout de même d'analyser la structure du pays. Il y une forte séparation entre le nord, plus pauvre, et le sud. Les capitales économiques et politiques que sont Douala et Yaoundé sont tellement importantes qu'elles "noient" les autres communes lors des analyses. Malgré la mauvaise spécification des données, j'ai tout de même réussi à dégager une partie de ces données pour faire un graphe social. L'utilisation de l'algorithme de Leiden pour détecter des communautés a permis de séparer les communes en quatre groupes. Ceux-ci semblent, lorsque l'on croise ces données avec celles de l'enquête auprès des ménages, discriminer les communes suivant leur niveau de richesse. Un autre résultat qui vient appuyer l'impression de hub des villes de Douala et Yaoundé, est le fait que les communes qui communiquent beaucoup, communiquent avec beaucoup de petites communes. Et les communes qui communiquent moins, ne communiquent qu'avec des communes importantes.

Avec des données plus importantes et plus précises, une analyse plus fine des communications à l'intérieur de ces villes peut être très informative dans la dynamique générale des communications du pays. Je n'ai pas eu le temps, mais j'aurais bien aimé utiliser l'algorithme d'OSLOM, qui permet de détecter des communautés avec différents niveaux hiérarchiques et des nœuds qui appartiennent à différentes communautés à la fois. Une idée, que je n'ai pas eu le temps de mettre en œuvre, est ensuite d'utiliser les hypergraphes pour modéliser ces communautés et les étudier. Enfin, il aurait été intéressant de vérifier si les différences observées entre les communautés de Leiden sont statistiquements significatives à l'aide de tests.

Finalement, ce stage dans une équipe de recherche m'a permis de voir à quoi peut ressembler le métier de chercheur et l'environnement de recherche d'une thèse. À la fois les points positifs mais aussi les difficultés que l'on peut rencontrer, comme les problèmes de financement ou les problèmes d'acquisition de données propres. À la fin de ce stage, je suis convaincu de vouloir poursuivre mes études en faisant une thèse. Sans avoir arrêté de domaine de recherche dans lequel je souhaiterais travailler, je suis ouvert à la découverte de nouveaux domaines et à la démarche scientifique, comme j'ai pu l'aborder lorsque j'ai étudié la théorie des graphes sociaux lors de ce stage.

Bibliographie

[1] Digital Business Africa. *Orange Money au Cameroun c'est 800 milliards F.Cfa de transactions par mois et 07 millions d'abonnés*. fr-FR. Juill. 2021. URL : https://www.digitalbusiness.africa/orange-money-au-cameroun-cest-800-milliards-f-cfa-de-transactions-par-mois-trois-millions-de-transactions-par-jour-et-07-millions-dabonnes/ (visité le 07/09/2023).

[2] Federico Battiston et al. "Networks beyond pairwise interactions: Structure and dynamics". en. In : *Physics Reports*. Networks beyond pairwise interactions: Structure and dynamics 874 (août 2020), p. 1-92. ISSN : 0370-1573. DOI : 10.1016/j.physrep.2020.05.004. URL : https://www.sciencedirect.com/science/article/pii/S0370157320302489 (visité le 05/07/2023).

[3] Federico Battiston et al. "The physics of higher-order interactions in complex systems". en. In : *Nat. Phys.* 17.10 (oct. 2021). Number: 10 Publisher: Nature Publishing Group, p. 1093-1098. ISSN : 1745-2481. DOI : 10.1038/s41567-021-01371-4. URL : https://www.nature.com/articles/s41567-021-01371-4 (visité le 04/07/2023).

[4] Austin R. Benson, David F. Gleich et Desmond J. Higham. *Higher-order Network Analysis Takes Off, Fueled by Classical Ideas and New Data*. arXiv:2103.05031 [physics]. Mars 2021. DOI : 10.48550/arXiv.2103.05031. URL : http://arxiv.org/abs/2103.05031 (visité le 06/07/2023).

[5] *Bias–variance tradeoff*. en. Page Version ID: 1163535785. Juill. 2023. URL : https://en.wikipedia.org/w/index.php?title=Bias%E2%80%93variance_tradeoff&oldid=1163535785 (visité le 15/09/2023).

[6] Christian Bick et al. *What are higher-order networks?* arXiv:2104.11329 [nlin, stat]. Juill. 2022. DOI : 10.48550/arXiv.2104.11329. URL : http://arxiv.org/abs/2104.11329 (visité le 06/07/2023).

[7] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In : *J. Stat. Mech.* 2008.10 (oct. 2008), P10008. ISSN : 1742-5468. DOI : 10.1088/1742-5468/2008/10/P10008. URL : https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008 (visité le 30/06/2023).

[8] S. Boccaletti et al. "The structure and dynamics of multilayer networks". In : *Physics Reports* 544.1 (nov. 2014). arXiv:1407.0742 [nlin, physics:physics], p. 1-122. ISSN : 03701573. DOI : 10.1016/j.physrep.2014.07.001. URL : http://arxiv.org/abs/1407.0742 (visité le 06/07/2023).

[9] INS Cameroun. *Enquête Complémentaire à la quatrième Enquête Camerounaise Auprès des Ménages (EC-ECAM 4) : MONOGRAPHIE DE LA REGION DU NORD – Institut National de la Statistique du Cameroun*. fr-FR. URL : https://ins-cameroun.cm/statistique/enquete-complementaire-a-la-quatrieme-enquete-camerounaise-aupres-des-menages-ec-ecam-4-monographie-de-la-region-du-nord/ (visité le 06/09/2023).

[10] INS Cameroun. *Répertoire actualisé des villages du Cameroun : Troisième Recensement Général de la Population et de l'Habitat du Cameroun – Institut National de la Statistique du Cameroun*. fr-FR. URL : https://ins-cameroun.cm/document/repertoire-actualise-des-villages-du-cameroun-troisieme-recensement-general-de-la-population-et-de-lhabitat-du-cameroun/ (visité le 07/09/2023).

[11] *Cameroun : 'Orange Money S.A n'est pas une banque'*. fr. Juill. 2022. URL : https://www.camerounweb.com/CameroonHomePage/business/Cameroun-Orange-Money-S-A-n-est-pas-une-banque-670685 (visité le 18/09/2023).

[12] Aaron Clauset, M. E. J. Newman et Cristopher Moore. "Finding community structure in very large networks". In : *Phys. Rev. E* 70.6 (déc. 2004). arXiv:cond-mat/0408187, p. 066111. ISSN : 1539-3755, 1550-2376. DOI : 10.1103/PhysRevE.70.066111. URL : http://arxiv.org/abs/cond-mat/0408187 (visité le 17/09/2023).

[13] Daniel Cohen. *Trois leçons sur la société post-industrielle*. Français. Paris : SEUIL, sept. 2006. ISBN : 978-2-02-085170-1.

[14] *Défis et progrès à l'ère des données et de l'intelligence artificielle*. fr. URL : https://www.sofrecom.com/news-insights/defis-et-progres-a-lere-des-donnees-et-de-lintelligence-artificielle.html (visité le 18/09/2023).

[15] Christian DEPOVER et al. *Perspectives pour la formation des maîtres en Francophonie - Collectif - Autrement - ebook (ePub) - Paris Librairies*. fr. URL : https://www.parislibrairies.fr/ebook/9782080277398-perspectives-pour-la-formation-des-maitres-en-francophonie-collectif/ (visité le 04/09/2023).

[16] Nicolas DUGUÉ et Anthony PEREZ. "Directed Louvain : maximizing modularity in directed networks". en. report. Université d'Orléans, nov. 2015. URL : https://hal.science/hal-01231784 (visité le 30/06/2023).

[17] Fabrice DURAND et Clément ROULLEAU. "L'essentiel sur les grands opérateurs dans le monde". fr. In : (2022).

[18] John Kenneth GALBRAITH, Maurice Le NAN et Jean-Louis CRÉMIEUX-BRILHAC. *Le Nouvel État industriel: Essai sur le système économique américain*. Français. Gallimard, 1989. ISBN : 978-2-07-071528-2.

[19] Laetitia GAUVIN, André PANISSON et Ciro CATTUTO. "Detecting the Community Structure and Activity Patterns of Temporal Networks: A Non-Negative Tensor Factorization Approach". en. In : *PLOS ONE* 9.1 (jan. 2014). Publisher: Public Library of Science, e86028. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0086028. URL : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0086028 (visité le 22/08/2023).

[20] *Gephi - The Open Graph Viz Platform*. URL : https://gephi.org/ (visité le 18/09/2023).

[21] M. GIRVAN et M. E. J. NEWMAN. "Community structure in social and biological networks". en. In : *Proc. Natl. Acad. Sci. U.S.A.* 99.12 (juin 2002), p. 7821-7826. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.122653799. URL : https://pnas.org/doi/full/10.1073/pnas.122653799 (visité le 30/06/2023).

[22] Petter HOLME et Jari SARAMÄKI. "Temporal Networks". In : *Physics Reports* 519.3 (oct. 2012). arXiv:1108.1780 [physics], p. 97-125. ISSN : 03701573. DOI : 10.1016/j.physrep.2012.03.001. URL : http://arxiv.org/abs/1108.1780 (visité le 06/07/2023).

[23] Georges V. HOUNGBONON, Erwan LE QUENTREC et Stefania RUBRICHI. "Access to electricity and digital inclusion: evidence from mobile call detail records". en. In : *Humanit Soc Sci Commun* 8.1 (juill. 2021). Number: 1 Publisher: Palgrave, p. 1-11. ISSN : 2662-9992. DOI : 10.1057/s41599-021-00848-0. URL : https://www.nature.com/articles/s41599-021-00848-0 (visité le 31/08/2023).

[24] *How Kriging works—ArcGIS Pro | Documentation*. URL : https://pro.arcgis.com/en/pro-app/latest/tool-reference/3d-analyst/how-kriging-works.htm (visité le 18/09/2023).

[25] Iacopo IACOPINI, Márton KARSAI et Alain BARRAT. *The temporal dynamics of group interactions in higher-order social networks*. arXiv:2306.09967 [physics]. Juin 2023. DOI : 10.48550/arXiv.2306.09967. URL : http://arxiv.org/abs/2306.09967 (visité le 07/07/2023).

[26] Iacopo IACOPINI et al. "Simplicial models of social contagion". en. In : *Nat Commun* 10.1 (juin 2019). Number: 1 Publisher: Nature Publishing Group, p. 2485. ISSN : 2041-1723. DOI : 10.1038/s41467-019-10431-6. URL : https://www.nature.com/articles/s41467-019-10431-6 (visité le 05/07/2023).

[27] Observatoire des INÉGALITÉS. *Près de 800 millions d'adultes analphabètes dans le monde*. fr. URL : https://www.inegalites.fr/Pres-de-800-millions-d-adultes-analphabetes-dans-le-monde (visité le 03/09/2023).

[28] *Institut National de la Statistique du Cameroun – Des statistiques de qualité pour un meilleur accompagnement du Cameroun vers son émergence*. fr-FR. URL : https://ins-cameroun.cm/ (visité le 18/09/2023).

[29] GSMA INTELLIGENCE. *GSMA Intelligence*. en-GB. 2023. URL : https://www.gsmaintelligence.com/ (visité le 29/08/2023).

[30] Kevin KELLER, Philip KOTLER et Delphine MANCEAU. *Marketing Management 16e édition (Redesign) + FastTrack*. Français. 16e édition. PEARSON, 2022. ISBN : 978-2-326-00293-7.

[31] Mikko KIVELÄ et al. "Multilayer networks". In : *Journal of Complex Networks* 2.3 (sept. 2014), p. 203-271. ISSN : 2051-1310. DOI : 10.1093/comnet/cnu016. URL : https://doi.org/10.1093/comnet/cnu016 (visité le 04/07/2023).

[32] Renaud LAMBIOTTE, Martin ROSVALL et Ingo SCHOLTES. *Understanding Complex Systems: From Networks to Optimal Higher-Order Models*. arXiv:1806.05977 [cond-mat, physics:physics]. Juin 2018. URL : http://arxiv.org/abs/1806.05977 (visité le 06/07/2023).

[33] Andrea LANCICHINETTI et al. "Finding Statistically Significant Communities in Networks". en. In : *PLoS ONE* 6.4 (avr. 2011). Sous la dir. d'Eshel BEN-JACOB, e18961. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0018961. URL : https://dx.plos.org/10.1371/journal.pone.0018961 (visité le 30/06/2023).

[34] Erwan LE QUENTREC et al. *Retour d'expérience de l'utilisation du mobile dans le dispositif de formation continue des enseignants du primaire à Madagascar*. Research Report. Agence Universitaire de la Francophonie, 2013, [154 p.] URL : https://hal.science/hal-01715633 (visité le 04/09/2023).

[35] *Le règlement général sur la protection des données - RGPD*. fr. URL : https://www.cnil.fr/fr/reglement-europeen-protection-donnees (visité le 18/09/2023).

[36] E. A. LEICHT et M. E. J. NEWMAN. "Community structure in directed networks". In : *Phys. Rev. Lett.* 100.11 (mars 2008). arXiv:0709.4500 [physics], p. 118703. ISSN : 0031-9007, 1079-7114. DOI : 10.1103/PhysRevLett.100.118703. URL : http://arxiv.org/abs/0709.4500 (visité le 17/09/2023).

[37] LEMONDE. "Orange, victime du succès de son plan de départ pour les seniors". fr. In : *Le Monde.fr* (jan. 2023). URL : https://www.lemonde.fr/economie/article/2023/01/31/chez-orange-un-couteux-dispositif-temps-partiel-senior_6159957_3234.html (visité le 04/09/2023).

[38] Gianluca MANZO. "Complex Social Networks are Missing in the Dominant COVID-19 Epidemic Models". en. In : *Sociologica* 14.1 (mai 2020). Number: 1, p. 31-49. ISSN : 1971-8853. DOI : 10.6092/issn.1971-8853/10839. URL : https://sociologica.unibo.it/article/view/10839 (visité le 18/09/2023).

[39] P. A. P. MORAN. "Notes on Continuous Stochastic Phenomena". In : *Biometrika* 37.1/2 (1950). Publisher: [Oxford University Press, Biometrika Trust], p. 17-23. ISSN : 0006-3444. DOI : 10.2307/2332142. URL : https://www.jstor.org/stable/2332142 (visité le 15/09/2023).

[40] Federico MUSCIOTTO, Federico BATTISTON et Rosario N. MANTEGNA. "Detecting informative higher-order interactions in statistically validated hypergraphs". en. In : *Commun Phys* 4.1 (sept. 2021). Number: 1 Publisher: Nature Publishing Group, p. 1-9. ISSN : 2399-3650. DOI : 10.1038/s42005-021-00710-4. URL : https://www.nature.com/articles/s42005-021-00710-4 (visité le 07/07/2023).

[41] *Namsor | Name checker for Gender, Origin and Ethnicity determination*. en. URL : https://namsor.app/ (visité le 18/09/2023).

[42] United NATIONS. *Population | Nations Unies*. fr. Publisher: United Nations. URL : https://www.un.org/fr/global-issues/population (visité le 03/09/2023).

[43] NETMOB. *NetMob Challenge 2023*. URL : https://www.netmob.org/ (visité le 04/09/2023).

[44] M. E. J. NEWMAN. "Equivalence between modularity optimization and maximum likelihood methods for community detection". en. In : *Phys. Rev. E* 94.5 (nov. 2016), p. 052315. ISSN : 2470-0045, 2470-0053. DOI : 10.1103/PhysRevE.94.052315. URL : https://link.aps.org/doi/10.1103/PhysRevE.94.052315 (visité le 30/06/2023).

[45] M. E. J. NEWMAN. "Mixing patterns in networks". In : *Phys. Rev. E* 67.2 (fév. 2003). Publisher: American Physical Society, p. 026126. DOI : 10.1103/PhysRevE.67.026126. URL : https://link.aps.org/doi/10.1103/PhysRevE.67.026126 (visité le 03/07/2023).

[46] M. E. J. NEWMAN. "The Structure and Function of Complex Networks". en. In : *SIAM Rev.* 45.2 (jan. 2003), p. 167-256. ISSN : 0036-1445, 1095-7200. DOI : 10.1137/S003614450342480. URL : http://epubs.siam.org/doi/10.1137/S003614450342480 (visité le 30/06/2023).

[47] M. E. J. NEWMAN et M. GIRVAN. "Finding and evaluating community structure in networks". en. In : *Phys. Rev. E* 69.2 (fév. 2004), p. 026113. ISSN : 1539-3755, 1550-2376. DOI : 10.1103/PhysRevE.69.026113. URL : https://link.aps.org/doi/10.1103/PhysRevE.69.026113 (visité le 30/06/2023).

[48] Mark Newman. *Networks, 2nd Edition*. eng. Second edition. OCLC: 1043555621. Oxford : Oxford University Press, 2018. ISBN : 978-0-19-252749-3.

[49] Orange. *Printemps 2023 - L'afrique, un marché d'opportunités et de défis.pdf*. 2023.

[50] Orange. *Toute l'histoire - Histoire*. 2023. URL : https://histoire.orange.com/fr/histoire-des-telecoms/toute-histoire/ (visité le 29/08/2023).

[51] *osmnx · PyPI*. URL : https://pypi.org/project/osmnx/ (visité le 18/09/2023).

[52] *Project OSRM*. URL : http://project-osrm.org/ (visité le 18/09/2023).

[53] Giulia Pullano et al. "Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study". English. In : *The Lancet Digital Health* 2.12 (déc. 2020). Publisher: Elsevier, e638-e649. ISSN : 2589-7500. DOI : 10.1016/S2589-7500(20)30243-0. URL : https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30243-0/fulltext#articleInformation (visité le 31/08/2023).

[54] *PySAL*. URL : http://pysal.org/ (visité le 18/09/2023).

[55] *République du Cameroun*. URL : https://www.axl.cefan.ulaval.ca/afrique/cameroun.htm (visité le 18/09/2023).

[56] RFI. *Afrique: ces coups d'États qui ont secoué le continent depuis trois ans*. fr. Section: afrique. Août 2023. URL : https://www.rfi.fr/fr/afrique/20230830-afrique-ces-coups-d-%C3%A9tats-qui-ont-secou%C3%A9-le-continent-depuis-trois-ans (visité le 03/09/2023).

[57] V. A. Traag, L. Waltman et N. J. Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". en. In : *Sci Rep* 9.1 (mars 2019), p. 5233. ISSN : 2045-2322. DOI : 10.1038/s41598-019-41695-z. URL : https://www.nature.com/articles/s41598-019-41695-z (visité le 17/09/2023).

[58] Eugenio Valdano et al. "Highlighting socio-economic constraints on mobility reductions during COVID-19 restrictions in France can inform effective and equitable pandemic response". In : *Journal of Travel Medicine* 28.4 (mai 2021), taab045. ISSN : 1708-8305. DOI : 10.1093/jtm/taab045. URL : https://doi.org/10.1093/jtm/taab045 (visité le 31/08/2023).

[59] Witold Wydmański. *Gender extractor*. original-date: 2021-07-12T10:54:20Z. Mai 2022. URL : https://github.com/wwydmanski/gender-extractor (visité le 15/09/2023).

[60] Jean-Gabriel Young, Giovanni Petri et Tiago P. Peixoto. "Hypergraph reconstruction from network data". en. In : *Commun Phys* 4.1 (juin 2021). Number: 1 Publisher: Nature Publishing Group, p. 1-11. ISSN : 2399-3650. DOI : 10.1038/s42005-021-00637-w. URL : https://www.nature.com/articles/s42005-021-00637-w (visité le 07/07/2023).