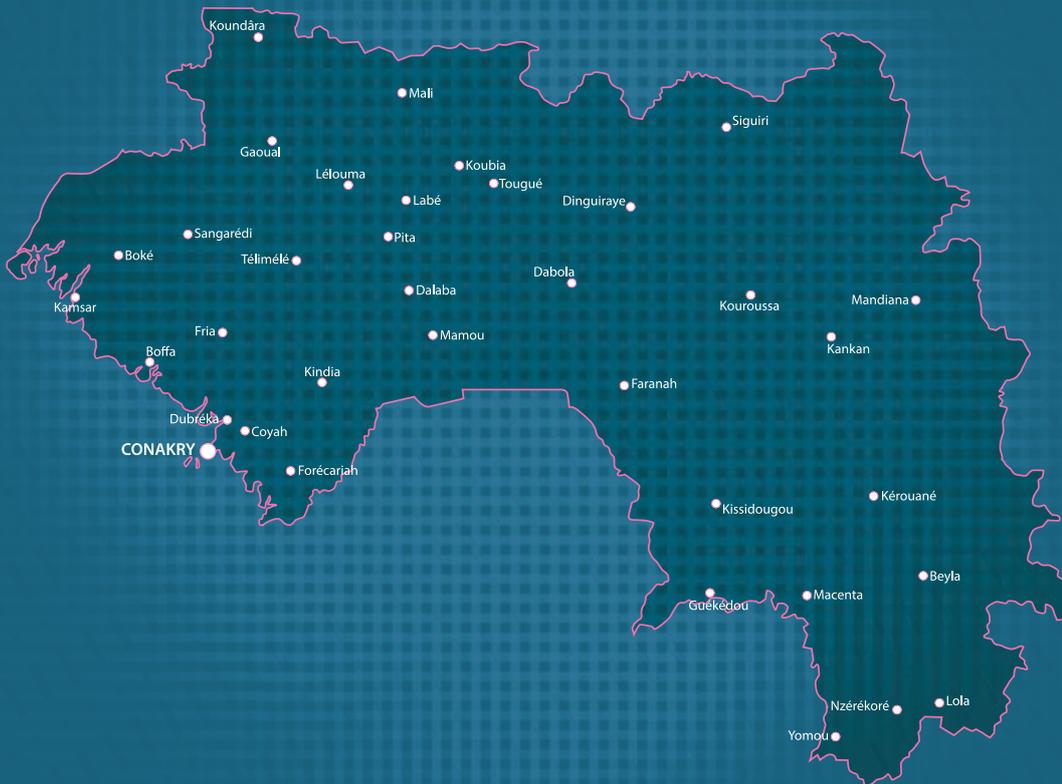


A N A L Y S E

des relevés détaillés des communications (CDR):
RÉPUBLIQUE DE GUINÉE

Rapport final



Analyse des relevés détaillés des communications (CDR): République de Guinée

Ce rapport de l'Union internationale des télécommunications (UIT) a été préparé par le Professeur Ryosuke Shibasaki du CSIS (Center for Spatial Information Science) de l'Université de Tokyo, expert de l'UIT, sous la supervision de la Division des pays les moins avancés, des petits Etats insulaires en développement et des télécommunications d'urgence au sein du Département de l'appui aux projets et de la gestion des connaissances (PKM) du Bureau de développement des télécommunications (BDT).

Les appellations employées dans la présente publication et la présentation des données, cartes comprises, qui y figurent n'impliquent de la part de l'UIT aucune prise de position quant au statut juridique des pays, territoires, villes ou zones, ou de leurs autorités, ni quant au tracé de leurs frontières ou limites.

ISBN:

978-92-61-22281-9 (version papier)

978-92-61-22291-8 (version électronique)

978-92-61-22301-4 (version EPUB)

978-92-61-22311-3 (version MOBI)



Avant d'imprimer ce rapport, pensez à l'environnement.

© ITU 2017

Tous droits réservés. Aucune partie de cette publication ne peut être reproduite, par quelque procédé que ce soit, sans l'accord écrit préalable de l'UIT.

Table des matières

1	Résumé	1
2	Considérations générales	2
3	Communications mobiles en Guinée	3
3.1	Technologie de téléphonie mobile	3
3.2	Etat des lieux des abonnés à la téléphonie mobile en Guinée	4
4	Relevés détaillés des communications (CDR)	4
4.1	Extraction d'informations sur la mobilité humaine et la répartition spatio-temporelle depuis les données CDR	5
4.2	Répercussions sur les secteurs public et privé	5
5	Ensemble de données CDR et statistiques de base CDR	6
6	Données CDR: limites et difficultés	8
7	Méthodologie	9
7.1	Aperçu de l'analyse des données CDR (parties prenantes incluses)	9
7.2	Aperçu des étapes de travail de l'analyse des données CDR	10
7.3	Spécifications des données CDR	11
7.4	Procédure de traitement des données CDR	12
7.5	Interpolation de la spécification du format	15
7.6	Système Hadoop pour le traitement des données	19
7.7	Estimation de la population dynamique	20
7.8	Téléphone mobile et clonage de carte SIM	22
8	Résultats des analyses	23
8.1	Population d'utilisateurs de téléphone mobile par préfecture	23
8.2	Visualisation des flux de population grâce au résultat d'interpolation	24
8.3	Mouvements de population dans une ville (sur 48 heures)	26
8.4	Surveillance des personnes ayant traversé une zone à risque	34
8.5	Tentatives de détection des cas de clonage de téléphones et de cartes SIM	37
8.6	Analyse transfrontière	39
8.6.1	Ensemble de données CDR avec IMEI transfrontière	39
8.6.3	Mouvements de population transfrontières entrants	41
8.6.4	Mouvements de population transfrontières sortants	43
9	Discussions	46
	Annexe 1: Concept global de l'analyse automatisée des données CDR	49
	Annexe 2: Bâtiments POI (points d'intérêt) et réseaux routiers	51

Liste des tableaux, figures et encadrés

Tableaux

Tableau 1: Exemple de données CSV	17
-----------------------------------	----

Figures

Figure 3.1: Base de commutation du système GSM	3
Figure 3.2: Carte administrative de la Guinée	4
Figure 5.1: Emplacement des stations de base en Guinée	7
Figure 5.2: Ensemble de données CDR	7
Figure 5.3: Activité quotidienne des appels	7
Figure 5.4: Utilisateurs uniques par jour	8
Figure 5.5: Nombre total moyen des emplacements uniques (par jour)	8
Figure 7.1: Concept global de l'analyse des données CDR	9
Figure 7.2: Aperçu des étapes de travail de l'analyse des données CDR	11
Figure 7.3: Procédure de traitement des données CDR	12
Figure 7.4: Procédé d'extraction des points de séjour	13
Figure 7.5: Répartition des POI et procédure de réattribution des points de séjour	13
Figure 7.6: Exemple de résultat de recherche du chemin le plus court	14
Figure 7.7: Cadre conceptuel du processus de recherche du chemin le plus court	14
Figure 7.8: Répartition des POI et procédure de réattribution des points de séjour	15
Figure 7.9: Données de la trajectoire et point de séjour	15
Figure 7.10: Points de trajet en mode séquentiel	16
Figure 7.11: Vérification de la trajectoire (usage interne)	19
Figure 7.12: Exemple de cluster Hadoop	20
Figure 7.13: Identification de l'emplacement de domicile avec grossissement de la population	21
Figure 7.14: Agrégation des données journalières aux fins d'une estimation de la population dynamique	22
Figure 8.1: Population d'utilisateurs de téléphone mobile par préfecture	23
Figure 8.2: Emplacements les plus visités par les personnes vivant à Conakry	24
Figure 8.3: Emplacements les plus visités par les personnes vivant à Conakry	24
Figure 8.4a: Flux de population à 6 heures et 9 heures	25
Figure 8.4b: Flux de population à 12 heures et 15 heures	26
Figure 8.5: Flux de population depuis une ville donnée (sur 48 heures)	27
Figure 8.6A: Répartition des personnes à l'intérieur/extérieur de la zone à risque	35
Figure 8.6B: Surveillance des personnes ayant traversé une zone à risque (10 heures plus tard)	35
Figure 8.7A: Où sont-elles allées dans les 48 heures? (12 heures après)	36
Figure 8.7B: Où sont-elles allées dans les 48 heures? (47,5 heures après)	36
Figure 8.8: Détection des cas de clonage (clonage IMEI)	37
Figure 8.9: Détection des cas de clonage (clonage IMSI)	38
Figure 8.10: Ensemble de données CDR avec IMEI transfrontière en Guinée	40
Figure 8.11: Registre des appels journaliers des utilisateurs mobiles guinéens en visite en Sierra Leone	40
Figure 8.12: Registre des appels journaliers des utilisateurs mobiles guinéens en visite au Libéria	40
Figure 8.13: Utilisateurs mobiles guinéens uniques en visite en Sierra Leone (par jour)	41
Figure 8.14: Utilisateurs mobiles guinéens uniques en visite au Libéria (par jour)	41
Figure 8.15: Empreinte des utilisateurs de téléphone mobile vers la Guinée	42
Figure 8.16: Nombre d'utilisateurs mobiles entrant en Guinée	43
Figure 8.17: Empreinte des visiteurs vers la Sierra Leone	44

Figure 8.18: Empreinte des visiteurs vers le Libéria	45
Figure 8.19: Origine et destination de la population transfrontière vers la Sierra Leone	45
Figure 8.20: Origine et destination de la population transfrontière vers le Libéria	46
Figure A.1.1: Concept global de l'analyse des données CDR	49
Figure A.2.1: Bâtiments POI en Afrique de l'Ouest	51
Figure A.2.2: Données du réseau routier en Afrique de l'Ouest	52
Figure A.2.3: Evaluation du réseau routier en Afrique de l'Ouest	52
Figure A.2.4: Extraction des données du réseau routier en Afrique de l'Ouest	53

1 Résumé

La téléphonie mobile est l'une des technologies les plus utilisées dans le monde moderne. Elle n'est pas seulement un outil de communication, elle est aussi un vecteur d'informations dans des secteurs tels que l'administration, la finance, l'éducation, l'agriculture et la santé. Elle joue un rôle particulièrement important dans le domaine de la gestion des situations d'urgence et des catastrophes en fournissant des estimations cruciales sur le nombre et les mouvements de population avant ou pendant de tels événements.

L'expérience de la Guinée, fortement touchée par l'épidémie d'Ebola au même titre que ses voisins le Libéria et la Sierra Leone, montre bien que la politique d'intervention visant à enrayer le virus n'aurait pas pu être efficace sans une compréhension de la mobilité des populations. Parce que l'Ebola est une maladie épidémique, les autorités doivent comprendre la mobilité humaine pour pouvoir construire des modèles essentiels à la prise de décisions, élaborer des politiques efficaces et planifier des interventions pour endiguer et combattre la maladie.

Les données des relevés détaillés des communications (CDR) fournissent des informations sur les activités du réseau de communication mobile ainsi que sur la mobilité humaine dans son ensemble, ce qui permet une action rapide contre la maladie. Le présent rapport montre comment l'analyse des données CDR peut aider à résoudre certaines problématiques spécifiques liées à l'épidémie d'Ebola, en estimant les trajectoires dynamiques et la répartition spatio-temporelle des populations.

Aux fins de ce rapport, une analyse des flux humains a été menée à deux niveaux différents: une analyse des mouvements de ville en ville en Guinée et une analyse des mouvements transfrontières entre trois pays (Guinée, Sierra Leone et Libéria). Les résultats de l'analyse de ville en ville ont montré une forte corrélation entre la répartition des utilisateurs de téléphonie mobile et les populations réelles au niveau des districts, permettant ainsi d'estimer quantitativement la répartition spatio-temporelle et les mouvements des personnes.

Bien que le rapport ne tienne compte que de la population couverte par les données CDR, les résultats de l'analyse fourniront au gouvernement et aux autorités locales une meilleure compréhension quantitative de la structure des flux de population dans le temps (à court et long termes) et lors d'événements particuliers. En cas d'épidémie, ils pourront ainsi extraire les données relatives à la migration de ville à ville pour les analyser et comprendre comment et où la maladie se propage de manière à la combattre.

L'analyse des données liées aux mouvements transfrontières montre par ailleurs plus de mobilité dans les villages voisins le long de la frontière que depuis et vers les zones centrales dans les pays frontaliers. Le suivi des flux depuis et vers les régions centrales n'est donc pas suffisant pour enrayer une épidémie. A noter cependant que les données analysées ont été enregistrées pendant la phase d'endiguement de l'épidémie.

Le présent rapport montre comment les flux migratoires et la répartition des utilisateurs de téléphone mobile au niveau spatio-temporel peuvent être utilisés pour analyser cette question et d'autres problèmes de société. Il présente l'utilisation de la technologie mobile et de technologies évoluées en matière de collecte et de diffusion des informations. Il montre également comment les données fournies par les opérateurs de téléphonie mobile cellulaire peuvent, à moindres coûts, fournir des cartes précises et détaillées de la répartition de la population dans une zone donnée. Enfin, ce rapport décrit comment la technologie du téléphone mobile est généralement utilisée, traite des systèmes GSM (Global System for Mobile Communication) et des futures technologies et présente diverses applications destinées à une utilisation dans les secteurs de l'agriculture, de la santé, de la finance, de la gouvernance, de la planification en prévision des catastrophes ainsi que des interventions et des activités de rétablissement en cas de catastrophe.

2 Considérations générales

Les données spatio-temporelles de la population peuvent être considérées comme un élément clé de l'intervention des pouvoirs publics pour lutter contre l'épidémie d'Ebola, ce en raison de la nature transmissible de la maladie et de son mode de propagation qui est fortement influencé par la mobilité humaine. La mondialisation des économies a accru le volume et la rapidité de la mobilité humaine, diminué le coût des transports et facilité les déplacements humains à grande échelle. Aussi, les données CDR relatives aux relevés détaillés des communications (données sur les téléphones mobiles) ont-elles attiré l'attention des décideurs et des chercheurs dans divers domaines compte tenu de leur capacité à restituer les tendances et les trajectoires des mouvements de population.

Les données CDR comprennent des étiquettes d'identité aléatoires, des données horodatées de communications mobiles et des localisations approximatives de communications représentées par les coordonnées géographiques des antennes radio. La localisation des antennes ne nous permet pas de repérer avec précision l'emplacement des utilisateurs de téléphone mobile mais la localisation proxy des dispositifs de téléphonie mobile peut être un outil puissant pour dégager les tendances générales des mouvements de population.

Traditionnellement, les travaux de suivi et d'analyse à l'échelle de la ville reposent sur une localisation fixe et une quantité considérable de données statistiques, ce qui ne permet pas d'identifier les événements multi-temporels à plus grande échelle. Cependant, au-delà des aspects quantitatifs de la mobilité humaine (par exemple, le volume de la population et sa rapidité de déplacement), la compréhension des aspects qualitatifs des personnes dans les données CDR (par exemple, le sexe, l'âge et la profession) devrait accroître l'utilisation des données CDR pour diverses applications. Les journaux d'utilisation des appareils mobiles serviront, le cas échéant, de support pour la collecte des données.

Les relevés détaillés des communications sont des données qui témoignent de l'activité sur le réseau de communication mobile et qui, lorsqu'elles sont agrégées, peuvent servir à estimer la densité de population d'une région donnée.

Dans le cadre d'un suivi à grande échelle, les données CDR des stations de base de réseaux mobiles fournissent d'excellents schémas de configuration spatiale qui reflètent la vie urbaine et ses dynamiques temporelles. Ces schémas pourraient même devenir, à l'avenir, un outil d'extraction et d'identification de problèmes moins évidents. L'analyse des données CDR rend certes compte des flux humains, mais il est à noter que la plupart des analyses n'utilisent les données que d'un seul pays ou d'une partie d'une région d'un pays, car il est difficile pour les opérateurs de réseaux mobiles de suivre les mouvements par-delà les frontières.

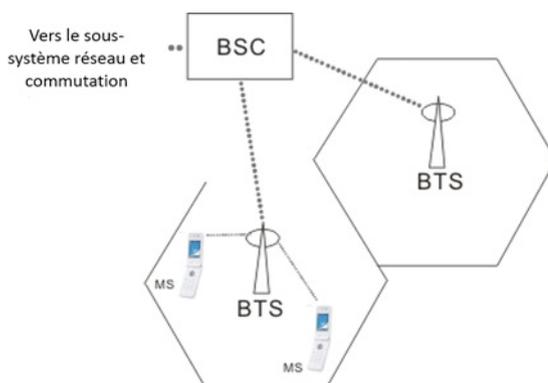
La dernière épidémie d'Ebola en Afrique de l'Ouest a décimé plus de 11 000 personnes. Les autorités sanitaires nationales de l'Afrique de l'Ouest sont intervenues pour contenir le virus, en essayant notamment d'estimer la population atteinte dans la zone à risque et de comprendre les schémas épidémiologiques propres à chaque région. Aujourd'hui, la simulation des mouvements de population repose sur des données statistiques et ne permet pas à ce titre d'obtenir une vision fiable et dynamique des flux de population dans une zone où un foyer épidémique a été déclaré. Elle ne permet pas non plus de prédire où éclatera la prochaine épidémie. L'utilisation des relevés détaillés des communications (CDR) générés par les données de la téléphonie mobile devrait précisément améliorer ce type de simulation. Les données CDR contiennent des informations temporelles et spatiales relatives aux services d'appels vocaux, de messagerie et de données utilisés sur chaque combiné - collectées à des fins de facturation - et peuvent également être utilisées pour estimer la dynamique actuelle et temporelle des mouvements de population pour, à terme, renforcer la préparation et le soutien et sauver davantage de vie.

3 Communications mobiles en Guinée

3.1 Technologie de téléphonie mobile

Le premier réseau mobile a concrètement vu le jour il y a près de 35 ans. Il ne cesse de croître depuis et intègre aujourd'hui les services modernes large bande que l'on connaît.

Figure 3.1: Base de commutation du système GSM



Source: UIT

En 1982, la Conférence européenne des administrations des postes et des télécommunications (CEPT) a créé un «Groupe spécial mobile (GSM)», parce que les systèmes analogiques existants ne pouvaient pas, du point de vue de l'abonné, offrir un service acceptable. Un système GSM est fondamentalement conçu comme une combinaison de trois grands sous-systèmes: le système de commutation de réseau (NSS), le système de station de base (BSS) et le système d'appui à l'exploitation (OSS).¹ Les données sont généralement collectées depuis l'unité de commande de station de base (BSC), qui fait partie du sous-système radioélectrique. Lorsqu'un usager émet un appel, le téléphone mobile se connecte à la station émettrice/réceptrice de base la plus proche (BTS).²

Actuelles et prochaines générations de la technologie de téléphonie mobile

- Les réseaux 2G ou deuxième génération de la technologie de téléphonie mobile, introduits pour la première fois en 1992, ont été les premiers à utiliser le chiffrement numérique des conversations et à offrir des services de données avec de la messagerie texte (SMS).
- Les réseaux 3G, qui ont succédé aux réseaux 2G avec des taux de transfert plus élevés, ont été les premiers à permettre des appels vidéo. Ils sont donc particulièrement appréciés des smartphones modernes qui nécessitent des vitesses de transmission élevées et constantes pour faire tourner leurs applications.
- La 4G est la quatrième génération des standards pour la téléphonie mobile. Elle succède à la 3G et fournit un accès Internet à très haut débit pour les appareils mobiles. Compte tenu de ces importants débits de transfert de données, les réseaux 4G sont aujourd'hui utilisés pour les modems sans fil des ordinateurs portables de même que pour l'accès à l'Internet à domicile.

¹ GSM: <https://www.itu.int/osg/spu/ni/3G/casestudies/GSM-FINAL.doc>

² T. Horanont. CSIS Discussion Paper. 2012. A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources No. 115.

3.2 Etat des lieux des abonnés à la téléphonie mobile en Guinée

La République de Guinée se compose de quatre régions naturelles: la Guinée maritime, la Moyenne-Guinée, la Haute-Guinée et la Guinée forestière. La Guinée est divisée en huit régions administratives, à leur tour subdivisées en 33 préfectures au total. Conakry est la capitale ; c'est aussi la plus grande ville de Guinée et son centre économique. La population de la Guinée en 2017 devrait atteindre plus de 11,7 millions de personnes³.

Figure 3.2: Carte administrative de la Guinée

La République de Guinée

- 8 Régions administratives:
 - Région de Conakry
 - Région de Nzérékoré
 - Région de Kankan
 - Région de Kindia
 - Région de Boké
 - Région de Labé
 - Région de Faranah
 - Région de Mamou
- Capitale:
 - Conakry, la plus grande ville
- Population (projections 2017):
11 702 692 habitants



En 2014, 72,1% de la population environ était couverte par des réseaux d'opérateurs mobiles. Il existe quatre grands opérateurs de téléphonie mobile en Guinée: Cellcom, Intercel, MTN et Orange. Ils fournissent des services GMS qui fonctionnent à la fois sur la base du prépaiement et du postpaiement. L'Autorité de Régulation des Postes et Télécommunications (ARPT) règlemente le secteur des télécommunications en Guinée.

4 Relevés détaillés des communications (CDR)

Les données des relevés détaillés des communications contiennent des informations de base sur l'utilisation des téléphones mobiles, telles que les stations de base auxquelles les téléphones de l'appelant et du destinataire étaient connectés au moment de l'appel, les identités des sources (points d'origine), les identités des destinations (points d'extrémité), la durée de chaque appel, le montant facturé pour chaque appel, le temps total d'utilisation sur la période de facturation, le temps libre total restant sur la période de facturation et le coût total facturé sur la période de facturation. L'opérateur connaît les emplacements des stations de base et les données CDR peuvent ainsi servir à estimer l'emplacement approximatif des deux parties, à des fins de géolocalisation. L'espacement des stations de base, et donc la précision dans la détermination de l'emplacement de l'appelant, varie selon le trafic attendu et le terrain. Les stations de base sont généralement espacées de 2 à 3 kilomètres dans les zones rurales et de 400 à 800 mètres dans les zones à forte densité de population. Ces données géospatiales sont extrêmement utiles dans le cadre des applications humanitaires et de développement.

³ Institut national de la statistique: www.stat-guinee.org/

4.1 Extraction d'informations sur la mobilité humaine et la répartition spatio-temporelle depuis les données CDR

Les emplacements séquentiels enregistrés dans les données CDR peuvent fournir une vue partielle des déplacements de population car les données CDR sont spécifiques à chaque individu, ce qui permet de tracer des mouvements individuels. Avec l'agrégation de ces mouvements individuels dans les données CDR, il est possible de déceler des tendances et de suivre la répartition spatio-temporelle de toute la population. En croisant ces informations avec des données secondaires, il est aussi possible d'analyser les relations entre la mobilité humaine et d'autres facteurs sociétaux.

4.2 Répercussions sur les secteurs public et privé

Répercussions sociales (au niveau national)

Cette section montre comment les informations extraites des données CDR peuvent être utilisées pour faire face à certains problèmes de société:

- **Interventions en cas de catastrophe:** Les catastrophes naturelles donnent lieu à des situations d'urgence, eu égard au caractère impérieux de fournir les informations vitales qui permettront de mobiliser rapidement les ressources pour faciliter l'intervention et la procédure de rétablissement. En 2010, une équipe de recherche dirigée par l'Institut Karolinska⁴ en Suède a montré que les données CDR pouvaient être utilisées pour diriger l'aide d'urgence grâce à l'analyse des données enregistrées dans les téléphones mobiles lorsque les personnes fuient les catastrophes naturelles.
- **Santé:** Les chercheurs ont également utilisé des données CDR avec un modèle simple de transmission de la maladie fondé sur les données de prévalence de l'infection et, ce faisant, ont pu cartographier les voies de dispersion des maladies. En analysant les profils de déplacements régionaux de millions d'abonnés mobiles, les chercheurs ont pu cartographier les endroits spécifiques où la probabilité de propagation de la maladie était la plus forte.
- **Développement socio-économique:** Les données CDR peuvent constituer un indicateur indirect pour évaluer le recensement de la population de même que les niveaux de pauvreté dans les régions et augmenter de manière appréciable les résultats des enquêtes nationales en estimant les changements associés au développement de l'économie.
- **Transports:** L'urbanisation rapide dans les pays en développement a accru la pression sur les infrastructures telles que les réseaux routiers. Les routes et les systèmes de transport public deviennent saturés et les gens perdent beaucoup de temps sur la route entre leur domicile et leur travail, ce qui a un coût économique collectif. En analysant les données CDR, les scientifiques peuvent tracer de nouvelles routes pour décongestionner les axes saturés et réduire les temps de déplacement.

Répercussions sur le secteur privé (opérateur de téléphonie mobile)

- **Evaluation de l'utilisation des téléphones:** Les données CDR révèlent plusieurs schémas d'utilisation, comme en attestent les caractéristiques des appels en fonction des groupes démographiques. Sont ici considérées les données spatiales et temporelles comme le nombre d'appels pour chaque période, la moyenne des appels, la durée des appels pour les groupes d'utilisateurs identifiés (homme, femme, étudiant, actif, femme au foyer, etc.), l'emplacement (là où les personnes du même groupe ou de groupes différents ont tendance à émettre des communications vocales et de données), le nombre de minutes par utilisateur en moyenne, le pourcentage de communications locales, le pourcentage de communications longue distance, le pourcentage de communications d'itinérance, le pourcentage de communications locales en heures creuses, le pourcentage de communications longue distance en heures creuses et le

⁴ <http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001083>

pourcentage de communications d'itinérance en heures creuses. Les données CDR anonymisées ont pour obligation de maintenir le lien avec les informations du groupe d'utilisateurs.

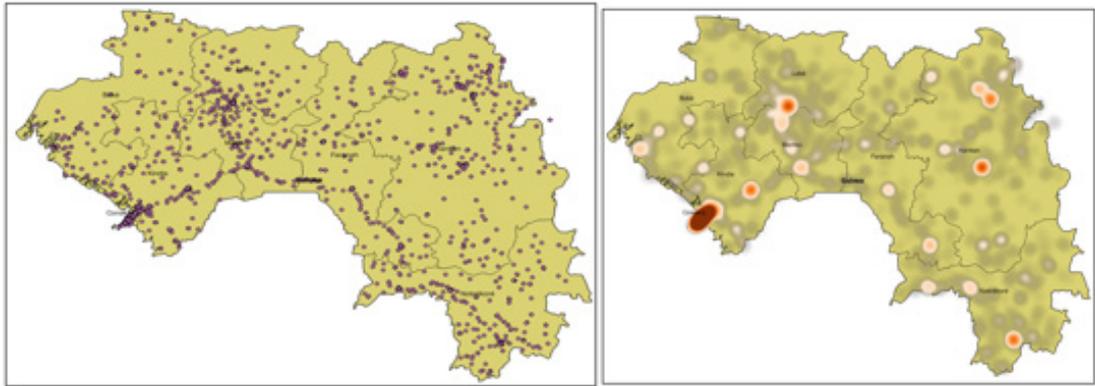
Ces informations pourraient aider les opérateurs à adapter leur service à des groupes spécifiques de personnes ainsi qu'à des emplacements spécifiques, à maintenir la satisfaction de la clientèle par le biais de mesures incitatives et à cibler le soutien à la clientèle, par exemple pour ceux qui adoptent des services et des fonctionnalités spécifiques.

- **Etablissement de profils clients:** Les opérateurs peuvent établir des profils clients à partir des schémas d'utilisation et élaborer sur cette base une structure tarifaire visant à optimiser l'efficacité. Qu'elles soient issues des données de l'utilisateur mobile ou des appels du service à la clientèle, les données disponibles dans les relevés détaillés des communications peuvent aider à améliorer les services et à promouvoir des possibilités ou façons de raccourcir le temps de garde.
- **Prévision des ventes:** Les clients émettent des appels ou utilisent des services, et les opérateurs peuvent améliorer ces services de même que l'efficacité du réseau par la prévision de l'utilisation desdits services. L'analyse de l'utilisation peut servir à déterminer une stratégie d'obsolescence programmée ou trouver des services complémentaires. La prévision, qui porte également sur le nombre de clients et la part de marché, permet de définir l'utilisation et d'estimer les revenus des groupes de clients.
- **Identification des cas de fraude ou d'utilisation abusive du téléphone mobile:** En utilisant les informations existantes des CDR, les opérateurs peuvent détecter des comportements anormaux voire abusifs chez leurs abonnés, limiter ou arrêter la fourniture des services associés et alerter les clients.
- **Planification de changements en matière de besoins en ressources:** Les données extraites des CDR relatives à l'utilisation peuvent aider à la planification de changements en ce qui concerne les besoins en ressources (par exemple, ajout d'une passerelle ou d'un serveur média sur un site donné). Elles peuvent être utilisées pour mesurer la qualité de service via la saisie d'informations sur les pertes de paquets, la latence et la gigue. Elles peuvent aussi être utilisées dans le cadre de la planification du réseau, par exemple en cas d'événements susceptibles de poser des problèmes en termes de capacité pour les stations de base.

5 Ensemble de données CDR et statistiques de base CDR

Dans le cadre du projet guinéen, deux mois de données CDR (de juin à juillet 2015) ont été collectés en Guinée pour montrer comment les mouvements de population dynamiques pouvaient être estimés. Les données ont été préparées par les opérateurs de réseaux mobiles de la Guinée: Cellcom, Intercel, MTN et Orange. La population à l'étude incluait par conséquent une majorité d'abonnés à la téléphonie mobile. Un ensemble d'informations anonymes a été créé pour tenir compte des besoins de confidentialité des données relatives aux abonnés. Le numéro d'identification unique a été remplacé par un nouveau numéro généré aléatoirement au moyen d'un algorithme générique de cryptographie par hachage et il n'existe aucun moyen de rétablir les identités d'origine. La Figure 5.1 recense les stations de base et leurs zones de couverture. La carte de la densité à droite indique une forte concentration de stations de base dans la ville de Conakry.

Figure 5.1: Emplacement des stations de base en Guinée



Source: UIT

La Figure 5.2 liste les informations de base issues des données CDR des opérateurs de réseaux mobiles. Les spécifications préparées à l'intention des opérateurs demandaient à ce que les données soient fournies dans un format spécifique. Tous les opérateurs n'ont pas préparé ces données dans le format spécifié et, pour ces cas particuliers, un traitement préalable des données a été nécessaire. En outre, le numéro de l'appelant a été utilisé en l'absence de code IMSI dans les données Intercell.

Figure 5.2: Ensemble de données CDR

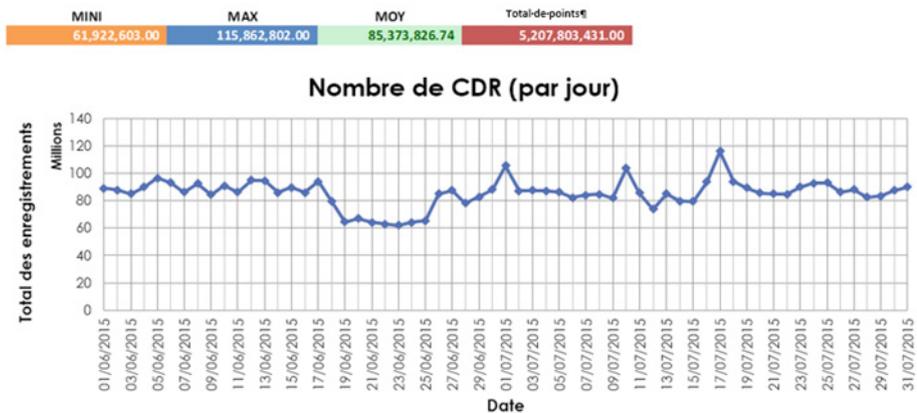
NO	Pays	MNO	Voix SMS									
			Taille des données	Période	Total des enregistrements	IMEI (appelant)	IMSI (appelant)	Nb d'appelants	Horo datage	LAC/ cell ID	Durée	Appel entrant/ sortant
1	Guinée	MTN	303 G	Juin-juillet 2015 (sauf 19-25 juin)	1,003,898,498	0 (6,979,708)	0 (2,699,233)	0 (2,660,893)	0	0	0	-
2		Orange	1,100 G	Juin-juillet 2015	3,915,651,949	0 (10,386,715)	0 (5,265,388)	0 (10,285,506)	0	0	0	-
3		Celcom	53 G	Juin-juillet 2015	279,441,946	0 (3,379,974)	0 (1,862,923)	-	0	0	-	-
4		Intercell	2.1 G	Juin-juillet 2015	8,811,038	-	-	0 (99,531)	0	0	0	-

Source: UIT

Statistiques de base CDR

Les données CDR ont été collectées depuis les opérateurs de réseaux mobiles et l'activité totale a généré en moyenne plus de 85,4 millions d'enregistrements par jour. La Figure 5.3 illustre l'activité quotidienne des appels.

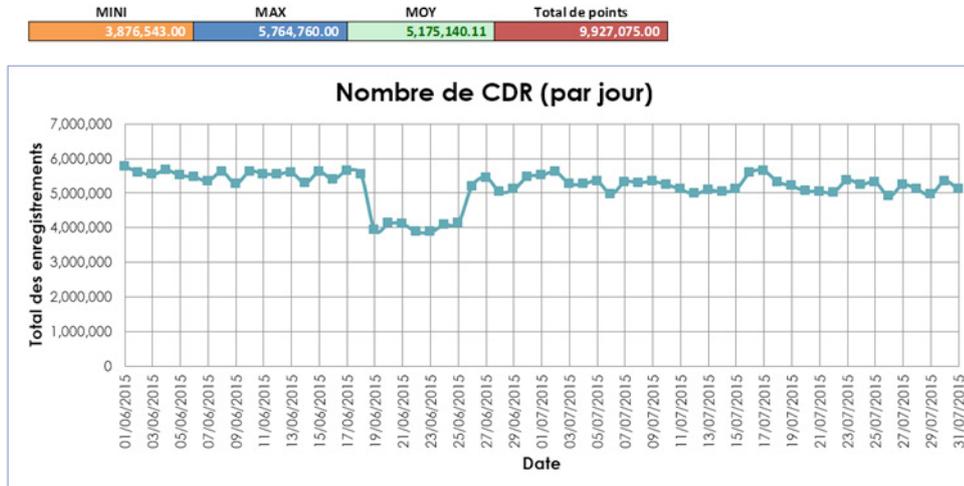
Figure 5.3: Activité quotidienne des appels



Source: UIT

Comme le montre la Figure 5.4, le nombre d'appels est plus élevé certains jours et moins élevé d'autres jours, en fonction des événements. L'activité quotidienne avoisine les 5 175 140 utilisateurs en moyenne et les données CDR par utilisateur et par jour se montent à 16,47 en moyenne.

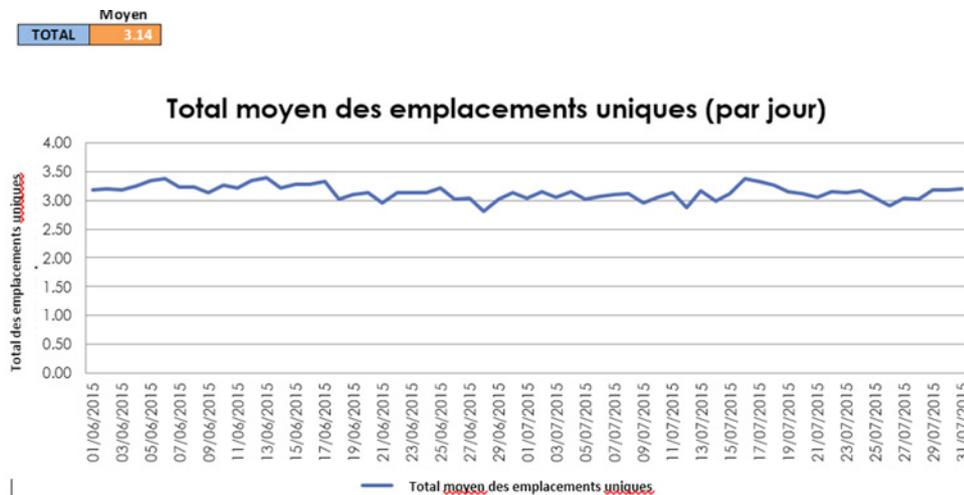
Figure 5.4: Utilisateurs uniques par jour



Source: UIT

La Figure 5.5 montre que le nombre total des emplacements uniques est de 3,14 en moyenne, ce qui signifie que les utilisateurs se rendent en moyenne sur trois emplacements différents chaque jour.

Figure 5.5: Nombre total moyen des emplacements uniques (par jour)



Source: UIT

6 Données CDR: limites et difficultés

Même si l'utilisation des données CDR peut apporter de nombreux avantages dans divers secteurs et pour diverses applications, comme décrit précédemment à la section 4, celle-ci se heurte à certains obstacles et difficultés.

- **Questions de confidentialité:** Lorsque les téléphones mobiles transmettent de la voix, ils recueillent et enregistrent d'autres informations inhérentes aux CDR, mais aussi des informations

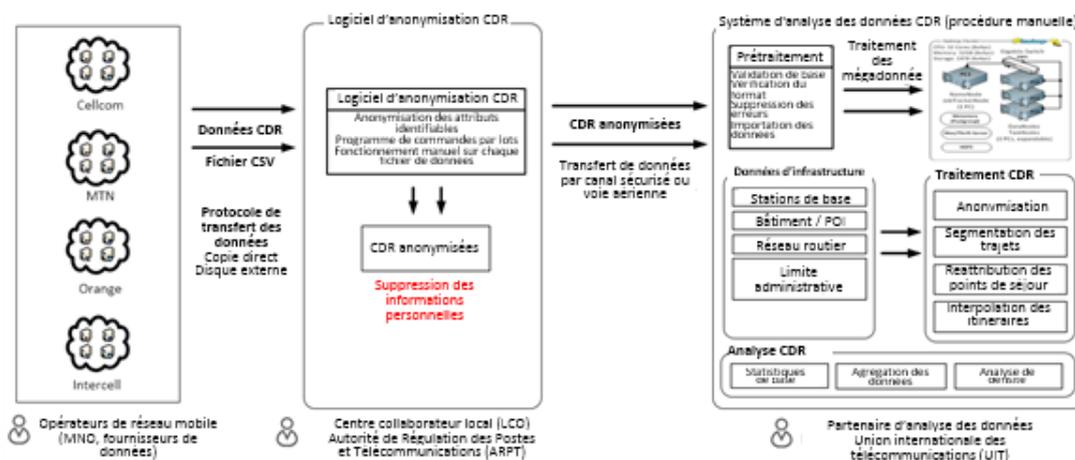
personnelles et spécifiques à l'utilisateur, ce qui crée de nouveaux défis en termes de conflit entre l'innovation technologique et le droit à la vie privée.

- Précision:** La précision de la géolocalisation et de la position estimée des appareils mobiles est un facteur important qui pourrait influencer les progrès continus de l'analyse locale sur ces appareils mobiles. Presque toutes les données CDR provenant des réseaux de télécommunication utilisent l'emplacement des stations de base pour estimer l'emplacement géographique des appareils. Dans la majorité des cas, la précision de cette méthode ne varie que de 50 à 300 mètres dans les environnements urbains de forte densité. Pour atténuer ce problème lié à l'exactitude, le rapport applique une méthode d'estimation des emplacements de séjour et des itinéraires de déplacement, qui s'appuie sur des données cartographiques numériques incluant les points d'intérêt (POI) et les données relatives au réseau de transport.
- Disponibilité des données:** Dans la plupart des pays, la recherche et l'étude des données sur les téléphones mobiles se limitent à la disponibilité des données des opérateurs. Bien que des ensembles de données soient disponibles ces dernières années et aient ouvert la possibilité aux chercheurs d'effectuer une analyse d'impact urbain et social à grande échelle, le soutien du secteur de la téléphonie mobile et la disponibilité des données sont encore très limités.
- Discontinuité des données:** Les relevés détaillés des communications sont générés lorsque les utilisateurs utilisent leur téléphone mobile, ce qui crée un manque de continuité au niveau des points d'accès consécutifs dans les CDR et, partant, un problème de localisation de l'utilisateur lors de l'analyse des données. La méthode d'estimation des emplacements de séjour et des itinéraires de déplacement mentionnée ci-avant est utilisée pour remédier à ce problème de discontinuité. En outre, et malgré les normes de formatage adoptées dans le secteur, la variété des formats de données nécessite un important travail de traitement préalable et de nettoyage des données avant analyse.

7 Méthodologie

7.1 Aperçu de l'analyse des données CDR (parties prenantes incluses)

Figure 7.1: Concept global de l'analyse des données CDR



Source: UIT

Le processus de l'analyse CDR comporte un certain nombre d'étapes pouvant être automatisées, comme l'utilisation du logiciel d'anonymisation développé et distribué aux fournisseurs de données pour retirer les données d'identification personnelles de leurs sets de données et conserver l'anonymat

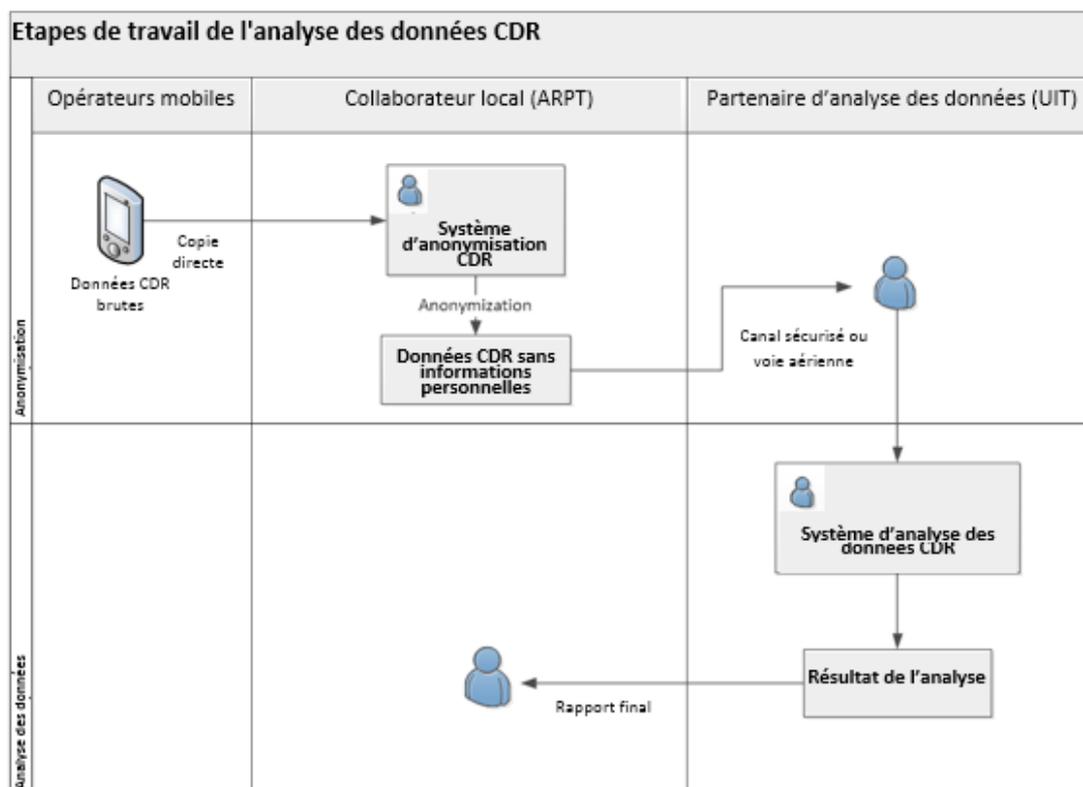
des personnes. Les données sont aussi transférées manuellement en ligne au moyen de canaux sécurisés et hors ligne au moyen de disques externes. Les détails de chaque étape et de chaque module sont décrits ci-dessous:

- **Opérateur de réseau mobile (MNO):** Les données sont collectées par les opérateurs. L'opérateur mobile recueille les données de localisation de l'appareil mobile enregistré sur son serveur, lesquelles données sont ensuite exportées et transférées vers les centres collaborateurs locaux (LCO) pour traitement ultérieur, généralement sous la forme d'un fichier CSV au format compressé. Dans le cadre du projet, les données sont transférées par copie directe.
- **Centre collaborateur local (LCO):** Cette unité propre aux pays stocke et assainit les données (pour supprimer le risque de divulgation des données personnelles). Le régulateur ou le fournisseur de licence de téléphonie mobile assume généralement cette fonction au même titre que le transfert des données anonymisées vers un partenaire d'analyse de données désigné.
- **Partenaire d'analyse des données (DAP):** Le DAP a pour rôle de conserver toutes les données CDR assainies qui ont été transmises par les opérateurs via le LCO et de procéder au traitement et à l'analyse de ces données.
 - Anonymisation des données CDR: Cette fonction gère la procédure d'anonymisation des données CDR. Le LCO reçoit les données CDR brutes au format CSV et procède manuellement à leur traitement pour supprimer toutes les informations liées à la confidentialité. Les données CDR peuvent ensuite être transférées au partenaire d'analyse des données. Pour ce projet, un macro-programme a été exécuté (une application en ligne de commande) après avoir défini certains paramètres tels que le chemin d'entrée, le chemin de sortie, les données de départ et l'analyseur de format CDR.
 - Analyse des données CDR: Cette fonction collecte toutes les données CDR auprès du LCO, procède à une vérification manuelle de ces données et les importe sur une plate-forme de mégadonnées avant de lancer une analyse plus poussée. Cette analyse réalisée par le DAP comprend plusieurs modules dont le prétraitement, le traitement des mégadonnées et le traitement des données CDR. Le traitement des données CDR est expliqué en détail à la section suivante.

7.2 Aperçu des étapes de travail de l'analyse des données CDR

La Figure 7.2 présente les étapes de travail de la procédure d'anonymisation des données CDR brutes jusqu'à l'analyse.

Figure 7.2: Aperçu des étapes de travail de l'analyse des données CDR



Source: UIT

7.3 Spécifications des données CDR

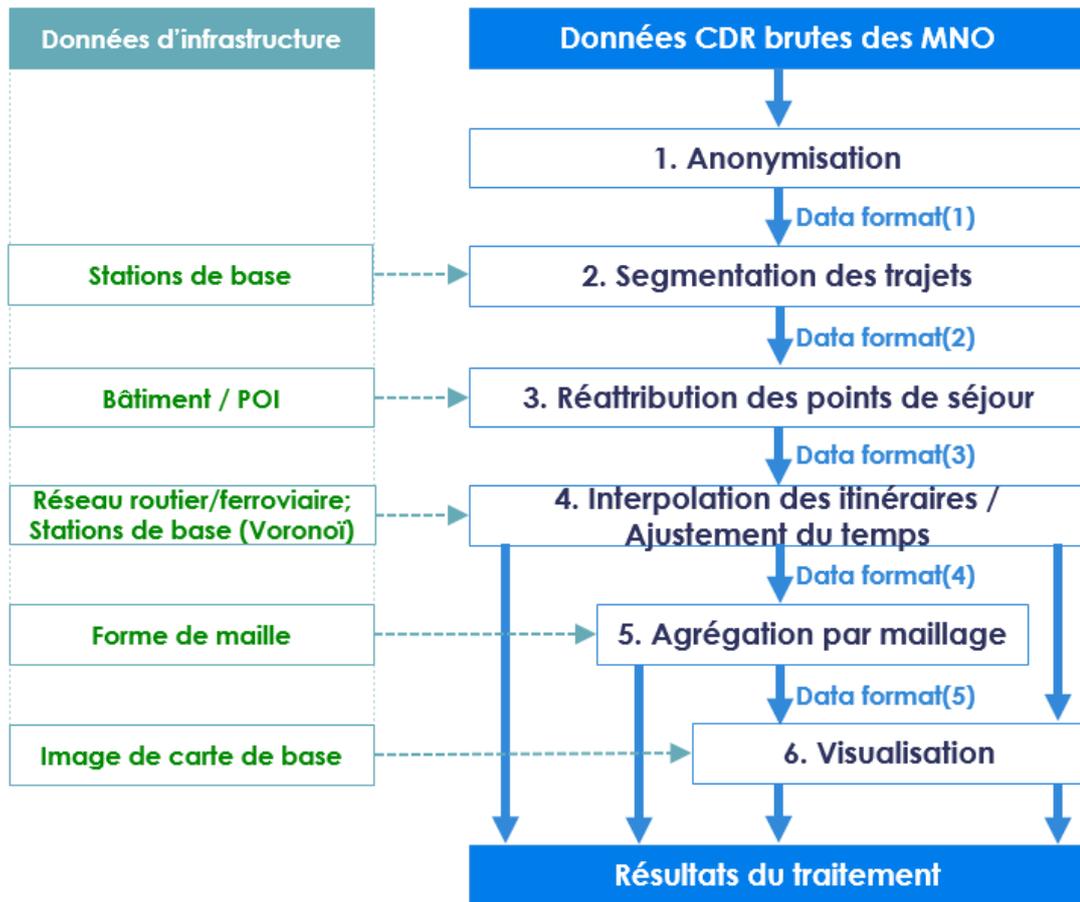
L'opérateur est invité à fournir les informations standardisées suivantes:

- l'identité d'équipement mobile internationale (IMEI) de l'appelant et de l'appelé:
 - unique pour chaque appareil;
 - cryptée de manière irréversible (au moyen d'une fonction de hachage);
- l'identité internationale d'abonné mobile (IMSI) de l'appelant et de l'appelé:
 - unique pour la carte SIM (module d'identification de l'abonné);
 - cryptée au moyen d'une fonction de hachage (de manière irréversible);
- l'horodatage des formats de début et de fin d'appel: AAAA-MM-JJ HH24:mm:ss (par exemple 2016-07-16 12:05:22);
- l'identité de l'appelant et les coordonnées de la station de base:
 - LAC, identité de cellule;
 - longitude, latitude;
- l'identité de l'appelé et les coordonnées de la station de base:
 - LAC, identité de cellule;
 - longitude, latitude;
- le numéro de téléphone mobile de l'appelant: crypté de manière irréversible au moyen d'une fonction de hachage;

- le type d'activité:
 - voix, SMS, données;
 - 2G, 3G, LTE.

7.4 Procédure de traitement des données CDR

Figure 7.3: Procédure de traitement des données CDR



Source: UIT

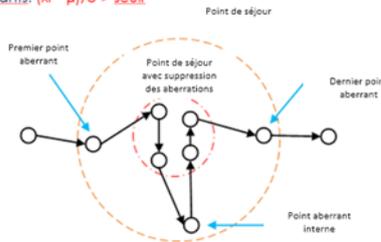
- 1 Anonymisation (hachage⁵):** Tous les attributs identifiables sont anonymisés au moyen d'une fonction de hachage cryptographique (SHA-256) pour protéger la confidentialité des utilisateurs. Les attributs d'origine sont retirés et remplacés par une clé de hachage générée aléatoirement.
- 2 Segmentation des trajets:** Extraction des points de séjour depuis les données CDR anonymisées et segmentation en composantes séjour et déplacement. La Figure 7.4 présente le procédé d'extraction des points de séjour appliquant des paramètres seuils aux données CDR.

⁵ Le hachage consiste à utiliser un programme pour extraire du texte et le convertir en une autre valeur unique mais aléatoire.

Figure 7.4: Procédé d'extraction des points de séjour

Extraction des points de séjour

- Paramètres seuils des points de séjour
 - Durée minimale: 15 minutes
 - Distance maximale: 300 mètres
- Détection des points aberrants
 - En utilisant la moyenne et l'écart-type
 - Points aberrants: $(x_i - \mu) / \sigma > \text{seuil}$



Source: UIT

3 Réattribution des points de séjour: Cette étape a pour but de réattribuer les points de séjour (origine et destination des trajets) sur les points d'intérêt (POI) à proximité – avec un degré de probabilité – et de combler les trous entre les segments séjour/déplacements. Un POI est considéré comme étant à proximité d'une station de base s'il est plus proche que les autres de ladite station de base (tessellation de Voronoï).

Une réattribution s'avère nécessaire dans la mesure où les données CDR de localisation se fondent sur l'emplacement de la station de base, à savoir que tous les utilisateurs d'une même zone sont localisés au même endroit.

La réattribution des données de la population sur les POI – plutôt que sur les emplacements des stations de base – permet d'obtenir une répartition plus réaliste reflétant les lieux où les utilisateurs sont le plus susceptibles de s'être rendus (par exemple, zones commerciales, bâtiments résidentiels et villages).

Un nouvel ensemble de données POI a été mis sur pied dans le cadre de cette procédure via la collecte de données issues de la répartition des bâtiments depuis les données Internet en libre accès (voir Annexe 2). La Figure 7.5 montre comment les POI sont répartis dans la ville. Les zones en bleu représentent les bâtiments POI avec les points de séjour extraits (là où les informations de localisation – initialement fondées sur les emplacements des stations de base – sont réaffectées).

Figure 7.5: Répartition des POI et procédure de réattribution des points de séjour

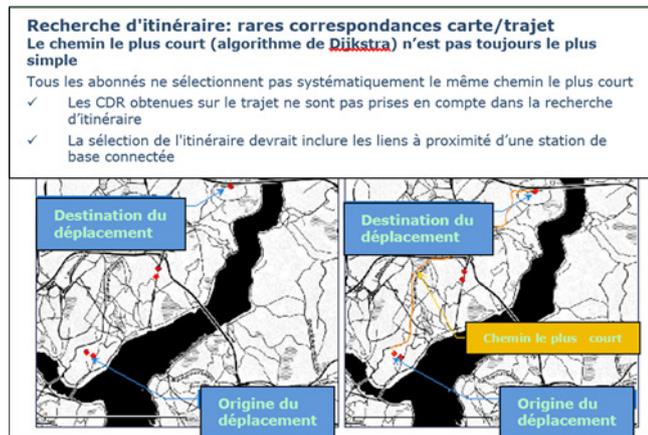


Source: UIT

4 Interpolation des itinéraires: Les itinéraires entre chaque paire de points de séjour réattribués (bâtiments POI) sont interpolés en utilisant le chemin le plus court entre les POI (origine et destination des déplacements). Sachant que le journal des heures de chaque enregistrement de données CDR ne coïncide pas avec le début et la fin du trajet d'un utilisateur de téléphone mobile, l'heure du voyage est modifiée en fonction d'une certaine distribution de probabilité. Dans le cadre de cette procédure, les données du réseau routier ont donc été constituées sur la base des données Open Street Map (OSM), en tenant compte du temps de déplacement sur le chemin le plus court.

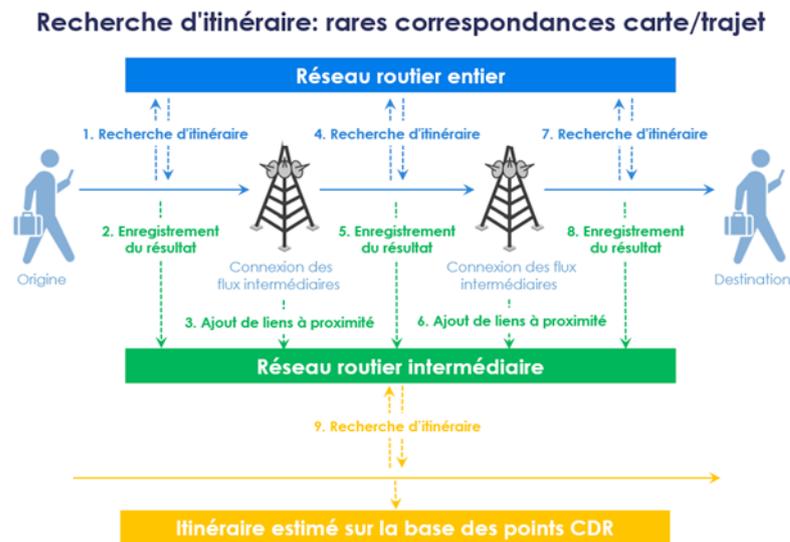
La Figure 7.6 montre comment trouver le chemin le plus court entre l'origine et la destination du déplacement. Les points rouges représentent l'emplacement des points de séjour réattribués et les joints jaunes le chemin le plus court entre l'origine et la destination. Certains points intermédiaires, qui ne sont pas des points de séjour, ne sont pas sur le chemin le plus court car seules l'origine et la destination du trajet sont affichées. Un algorithme sélectionne le chemin le plus court entre l'origine et la destination en tenant compte également des points intermédiaires. La Figure 7.7 présente le cadre du processus de recherche du chemin le plus court tandis que la Figure 7.8 affiche un itinéraire passant par des points intermédiaires.

Figure 7.6: Exemple de résultat de recherche du chemin le plus court



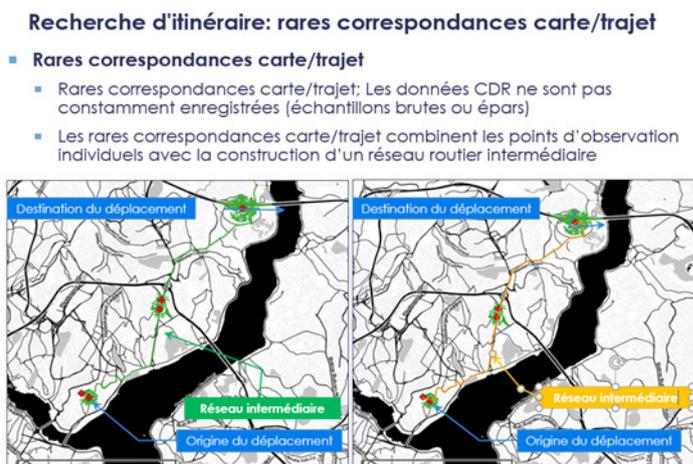
Source: UIT

Figure 7.7: Cadre conceptuel du processus de recherche du chemin le plus court



Source: UIT

Figure 7.8: Répartition des POI et procédure de réattribution des points de séjour



Source: UIT

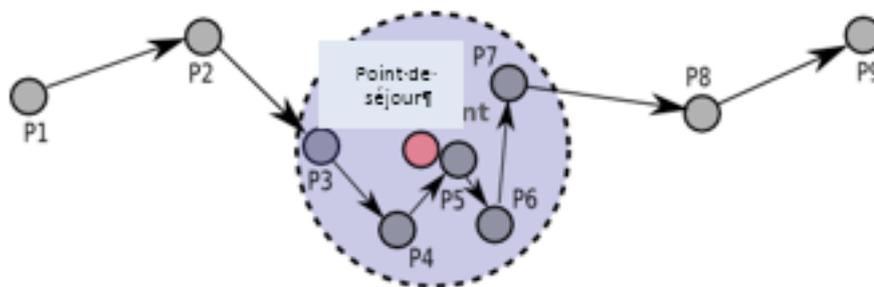
- 5 **Agrégation par maillage:** Il s'agit ici de la mesure globale de la densité de population dans une grille donnée (1 km) et dans une période de temps donnée (1 heure).
- 6 **Visualisation:** Création d'une projection animée (film généré par Mobmap⁶) des mouvements agrégés/désagrégés de population (hors ligne).

7.5 Interpolation de la spécification du format

Cette section fournit quelques définitions de base concernant les données de la trajectoire et présente la terminologie appropriée pour l'exploration des données relatives à la trajectoire.

- **Données de la trajectoire:** séquence de points horodatés, $P = (p_1, p_2, \dots, p_n)$, où $p = (\text{ID}, \text{heure}, \text{latitude}, \text{longitude})$ et $n = \text{nombre total de points}$.

Figure 7.9: Données de la trajectoire et point de séjour



Source: UIT

- **Point de séjour:** Référence géographique indiquant l'endroit où un utilisateur a séjourné pendant une durée seuil (tt) et à l'intérieur d'une distance seuil (dt). Pour une trajectoire donnée, le point de séjour se caractérise par une série de points consécutifs $P = \{p_m, p_{m+1}, \dots, p_n\}$, où $\forall m < i \leq n$, $\text{Distance}(p_m, p_i) \leq dt$, $\text{Distance}(p_m, p_{n+1}) > dt$ et $\text{Intervalle de temps}(p_m, p_n) \geq tt$. Par conséquent, $s = (x, y, t_a, t_l)$, où x et y correspondent à la localisation centroïde de ces points dans le point de séjour.
- **Historique de localisation:** Il s'agit ici de l'historique de localisation individuelle (h) représentée par une séquence des points de séjour visités pendant les périodes de transition correspondantes:

⁶ <http://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/>

$$h = \langle s_0 \xrightarrow{\Delta t_1} s_1 \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_{n-1}} s_n \rangle$$

où s_i est un point de séjour et Δt_i l'intervalle de temps entre deux points de séjour.

- **Emplacements importants:** Les emplacements importants sont pour chaque individu les lieux fréquemment fréquentés au quotidien tels que le domicile, le lieu de travail et le supermarché.
- **Point de trajet:** Enregistrement contenant les propriétés du trajet, qui représente une séquence de points GPS ayant la même activité (par exemple STAY ou MOVE, voir la Figure 7.10). Les propriétés incluent la durée, la distance, la catégorie, le mode de transport, etc.
- **Sous-trajet:** Représente un groupe de points de trajet séquentiels entre un point de séjour et le prochain point de séjour (par exemple stay-> walk-> train-> walk-> stay). Décrit également l'activité d'un utilisateur (par exemple, l'utilisateur sort de chez lui pour se rendre au travail).
- **Trajet:** Ensemble de points de trajet ou de sous-trajets dans une journée en mode séquentiel. Un trajet doit comporter au moins un sous-trajet. Il correspond à l'activité séquentielle globale dans la journée de l'utilisateur. En règle générale, un trajet comporte plusieurs points de trajet STAY ou MOVE. Un exemple est fourni à la Figure 7.10.

Figure 7.10: Points de trajet en mode séquentiel

Trajet	S o u s - t r a j e t	1, TripPoint [name=STAY, mode=STAY, Time=00:00:15-08:03:50
		2, TripPoint [name=MOVE, mode=WALK, Time=08:03:50-08:18:30
		3, TripPoint [name=MOVE, mode=TRAIN, Time=08:18:30-08:48:38
		4, TripPoint [name=MOVE, mode=WALK, Time=08:48:38-10:56:15
		5, TripPoint [name=STAY, mode=STAY, Time=10:56:15-21:16:29
		6, TripPoint [name=MOVE, mode=WALK, Time=21:16:29-23:20:25
		7, TripPoint [name=MOVE, mode=TRAIN, Time=23:20:25-23:35:57
		8, TripPoint [name=MOVE, mode=WALK, Time=23:35:57-23:45:58
		9, TripPoint [name=STAY, mode=STAY, Time=23:45:58-23:55:25

Source: UIT

Le tableau 1 fournit un exemple de données de trajectoire incluant la terminologie de base ainsi que les informations requises et le format utilisé dans l'exploration des données de trajectoire. Les résultats de l'interpolation sont conditionnés séparément sous la forme de données de trajet par utilisateur et par date. Les résultats comportent 11 colonnes: Identité d'utilisateur, Date, Séquence du trajet, Type de mobilité, Mode de transport, Distance totale, Durée totale, Date de début, Date de fin, Total des points et Listes de points. Ils sont exportés au format CSV (valeurs séparées par des virgules) et chaque colonne est renseignée.

Tableau 1: Exemple de données CSV

Résultats	Contenu des colonnes au format CSV
1 Identité d'utilisateur	<ul style="list-style-type: none"> Nom de la colonne: UID Unique pour chaque appareil Cryptage au moyen d'une fonction de hachage (irréversible)
2 Date	<ul style="list-style-type: none"> Nom de la colonne: DATE Format de la date: aaaa-MM-jj Exemple: 2015-12-31
3 Séquence du trajet	<ul style="list-style-type: none"> Nom de la colonne: TRIP_SEQUENCE Ordre du sous-trajet dans une journée, commence à 1
4 Type de mobilité	<ul style="list-style-type: none"> Nom de la colonne: MOBILITY_TYPE Valeur: STAY ou MOVE
5 Mode de transport	<ul style="list-style-type: none"> Nom de la colonne: TRANSPORT_MODE Mode de transport du sous-trajet correspondant Valeur: STAY, WALK, CAR
6 Distance totale	<ul style="list-style-type: none"> Nom de la colonne: TOTAL_DISTANCE Distance totale de déplacement du sous-trajet, en mètres
7 Durée totale	<ul style="list-style-type: none"> Nom de la colonne: TOTAL_TIME Durée totale de déplacement du sous-trajet, en secondes
8 Date de début	<ul style="list-style-type: none"> Nom de la colonne: START_TIME Date de début du sous-trajet Format: HH24:mm:ss Exemple: 23:20:00
9 Date de fin	<ul style="list-style-type: none"> Nom de la colonne: END_TIME Date de fin du sous-trajet Format: HH24:mm:ss Exemple: 23:20:00
10 Total des points	<ul style="list-style-type: none"> Nom de la colonne: TOTAL_POINTS Nombre total de points dans le sous-trajet

Résultats	Contenu des colonnes au format CSV
11 Listes de points	<ul style="list-style-type: none"> Nom de la colonne: POINT_LISTS Liste de points dans le sous-trajet Format: N° heure latitude longitude N° est le numéro d'ordre, commence à 1 Heure: aaaa-MM-jj HH24:mm:ss Latitude et longitude au format décimal Chaque point est séparé par «;» Exemple: 1 2015-06-06 13:53:23 6.373743 -10.772951

Note: dans le présent projet, l'interpolation du segment MOVE est définie à intervalles d'1 minute.

Analyse des segments de trajet aux fins de vérification du résultat de l'interpolation

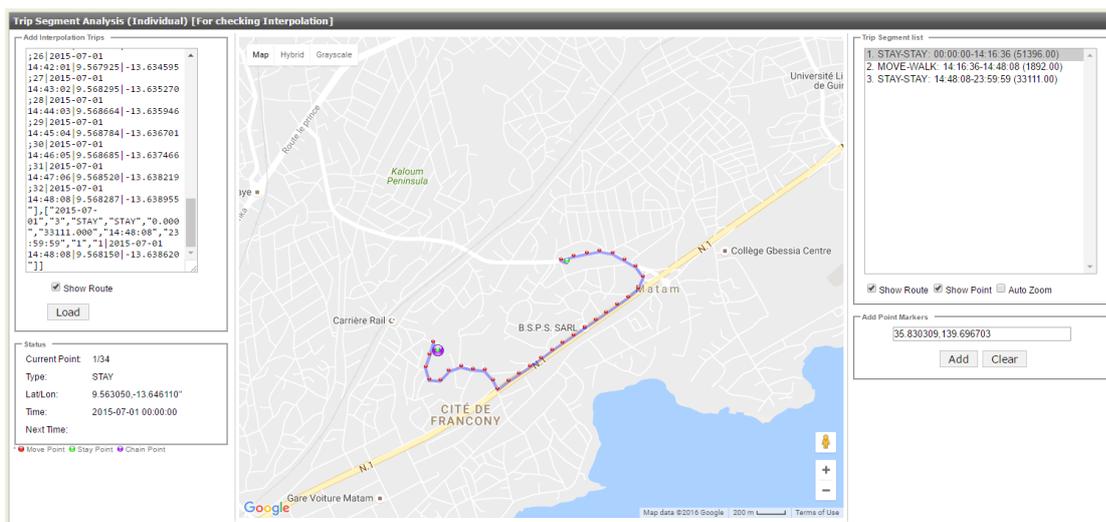
UID, DATE (yyyy-MM-dd), TRIP_SEQUENCE_ID, MOBILITY_TYPE, TRANSPORT_MODE, TOTAL_DISTANCE, TOTAL_TIME, START_TIME, END_TIME, TOTAL_POINTS, POINT_LIST

422a837717,2015-07-01,1,STAY,STAY,0.000,51396.000,00:00:00,14:16:36,1,1|2015-07-01 00:00:00|9.563050|-13.646110

4 2 2 a 8 3 7 7 1 7 , 2 0 1 5 - 0 7 - 0 1 , 2 , M O V E , W A L K , 2 5 5 8 . 7 1 9 , 1 8 9 2 . 0 0 0 , 1 4 : 1 6 : 3 6 , 1 4 : 4 8 : 0 8 , 3 2 , 1 | 2 0 1 5 - 0 7 - 0 1 1 4 : 1 6 : 3 6 | 9 . 5 6 3 5 6 2 | - 1 3 . 6 4 6 3 5 3 ; 2 | 2 0 1 5 - 0 7 - 0 1 1 4 : 1 7 : 3 7 | 9 . 5 6 2 8 2 8 | - 1 3 . 6 4 6 5 7 7 ; 3 | 2 0 1 5 - 0 7 - 0 1 1 4 : 1 8 : 3 8 | 9 . 5 6 2 0 9 5 | - 1 3 . 6 4 6 8 0 2 ; 4 | 2 0 1 5 - 0 7 - 0 1 1 4 : 1 9 : 3 9 | 9 . 5 6 1 4 1 5 | - 1 3 . 6 4 6 5 9 0 ; 5 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 0 : 4 0 | 9 . 5 6 1 3 5 4 | - 1 3 . 6 4 5 9 5 5 ; 6 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 1 : 4 1 | 9 . 5 6 1 8 9 1 | - 1 3 . 6 4 5 4 5 2 ; 7 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 2 : 4 2 | 9 . 5 6 2 1 7 4 | - 1 3 . 6 4 4 7 4 5 ; 8 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 3 : 4 3 | 9 . 5 6 1 9 8 4 | - 1 3 . 6 4 4 0 6 7 ; 9 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 4 : 4 4 | 9 . 5 6 1 9 4 7 | - 1 3 . 6 4 3 3 7 5 ; 1 0 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 5 : 4 5 | 9 . 5 6 1 3 6 3 | - 1 3 . 6 4 2 8 8 9 ; 1 1 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 6 : 4 6 | 9 . 5 6 0 8 3 8 | - 1 3 . 6 4 2 6 2 3 ; 1 2 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 7 : 4 7 | 9 . 5 6 1 3 0 0 | - 1 3 . 6 4 2 0 0 7 ; 1 3 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 8 : 4 8 | 9 . 5 6 1 7 4 5 | - 1 3 . 6 4 1 3 7 7 ; 1 4 | 2 0 1 5 - 0 7 - 0 1 1 4 : 2 9 : 4 9 | 9 . 5 6 2 1 7 8 | - 1 3 . 6 4 0 7 4 1 ; 1 5 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 0 : 5 0 | 9 . 5 6 2 6 2 3 | - 1 3 . 6 4 0 1 1 1 ; 1 6 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 1 : 5 1 | 9 . 5 6 3 0 6 7 | - 1 3 . 6 3 9 4 8 1 ; 1 7 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 2 : 5 2 | 9 . 5 6 3 5 0 6 | - 1 3 . 6 3 8 8 4 8 ; 1 8 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 3 : 5 3 | 9 . 5 6 3 9 4 4 | - 1 3 . 6 3 8 2 1 4 ; 1 9 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 4 : 5 4 | 9 . 5 6 4 3 8 2 | - 1 3 . 6 3 7 5 8 0 ; 2 0 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 5 : 5 5 | 9 . 5 6 4 8 2 3 | - 1 3 . 6 3 6 9 4 8 ; 2 1 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 6 : 5 6 | 9 . 5 6 5 2 6 6 | - 1 3 . 6 3 6 3 1 7 ; 2 2 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 7 : 5 7 | 9 . 5 6 5 7 0 5 | - 1 3 . 6 3 5 6 8 4 ; 2 3 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 8 : 5 8 | 9 . 5 6 6 1 4 4 | - 1 3 . 6 3 5 0 5 1 ; 2 4 | 2 0 1 5 - 0 7 - 0 1 1 4 : 3 9 : 5 9 | 9 . 5 6 6 6 2 7 | - 1 3 . 6 3 4 4 5 6 ; 2 5 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 1 : 0 0 | 9 . 5 6 7 3 0 8 | - 1 3 . 6 3 4 1 7 4 ; 2 6 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 2 : 0 1 | 9 . 5 6 7 9 2 5 | - 1 3 . 6 3 4 5 9 5 ; 2 7 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 3 : 0 2 | 9 . 5 6 8 2 9 5 | - 1 3 . 6 3 5 2 7 0 ; 2 8 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 4 : 0 3 | 9 . 5 6 8 6 6 4 | - 1 3 . 6 3 5 9 4 6 ; 2 9 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 5 : 0 4 | 9 . 5 6 8 7 8 4 | - 1 3 . 6 3 6 7 0 1 ; 3 0 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 6 : 0 5 | 9 . 5 6 8 6 8 5 | - 1 3 . 6 3 7 4 6 6 ; 3 1 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 7 : 0 6 | 9 . 5 6 8 5 2 0 | - 1 3 . 6 3 8 2 1 9 ; 3 2 | 2 0 1 5 - 0 7 - 0 1 1 4 : 4 8 : 0 8 | 9 . 5 6 8 2 8 7 | - 1 3 . 6 3 8 9 5 5

422a837717,2015-07-01,3,STAY,STAY,0.000,33111.000,14:48:08,23:59:59,1,1|2015-07-01 14:48:08|9.568150|-13.638620

Figure 7.11: Vérification de la trajectoire (usage interne)



Source: UIT

7.6 Système Hadoop pour le traitement des données

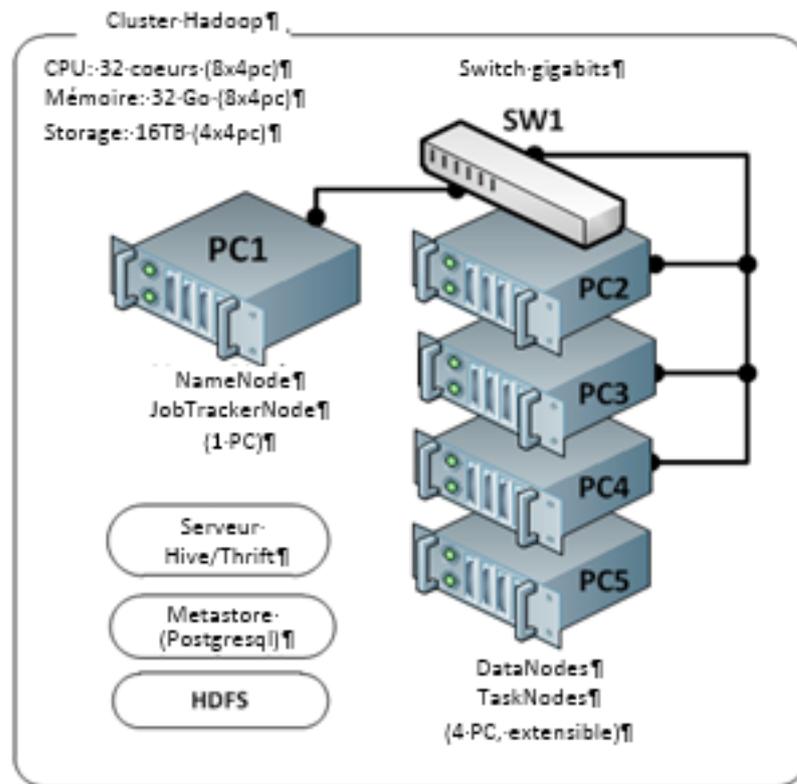
Les données CDR forment un ensemble de données très volumineux que les ordinateurs ordinaires ou systèmes de bases de données ne peuvent traiter dans un délai acceptable. C'est pourquoi le système Hadoop a été introduit comme principal outil de traitement des données pour ce projet. Hadoop est un cadre logiciel open source déployé dans le cloud pour les applications distribuées et la gestion intensive des données. Hadoop propose plusieurs services et fonctionnalités, mais seuls Hadoop Distributed File System (HDFS) et Hive ont été utilisés dans le cadre du présent projet. Pour que le système soit pleinement installé et opérationnel, cinq composants ont été nécessaires: NameNode, DataNodes, Secondary NameNode (SNN), JobTracker et TaskTrackers.

NameNode, c'est l'administrateur du système de fichiers HDFS; il conserve trace de la façon dont les fichiers sont fractionnés en blocs de fichiers ainsi que des nœuds qui stockent ces blocs et de la santé globale du système de fichiers distribués. Les DataNodes sont les travailleurs du système. Ils stockent et restituent les blocs de données sur demande (du client ou du NameNode) et présentent périodiquement au NameNode les listes des blocs stockés. Le NameNode secondaire (SNN) est un daemon d'arrière-plan qui contrôle l'état du cluster HDFS et prend des snapshots du NameNode pour aider à limiter les temps d'arrêt et les pertes de données. JobTracker assure la liaison entre l'application et Hadoop. Après le lancement du code sur le cluster, le JobTracker définit le plan d'exécution en déterminant les fichiers à traiter, affecte les noeuds à différentes tâches et contrôle toutes les tâches en cours d'exécution. Les TaskTrackers sont responsables de l'exécution des tâches individuelles que le JobTracker affecte et gèrent l'exécution des tâches individuelles sur chaque nœud esclave.

L'expérimentation Hadoop a porté sur cinq ordinateurs de mêmes caractéristiques: Xeon 2,6 GHz, mémoire de 8 Go, disque de 2x2 To avec CentOS 6.0 64 bits pour le système de bases de données et l'application de gestion de bibliothèque. PostgreSQL 9.0.6 avec PostGIS 1.5.3 a été installé sur le système. L'un des ordinateurs fonctionne sur le NameNode; les autres sont sur les DataNodes et TaskNodes.

La Figure 7.12 détaille le système utilisé pour ce projet: un système Hadoop avec 32 coeurs, 32 Go de mémoire et 16 To de stockage, capable de gérer jusqu'à 28 tâches en même temps. Les versions utilisées sont les suivantes: Hadoop 0.20.2, Hive 0.8.0 et JTS 1.12.

Figure 7.12: Exemple de cluster Hadoop



Source: UIT

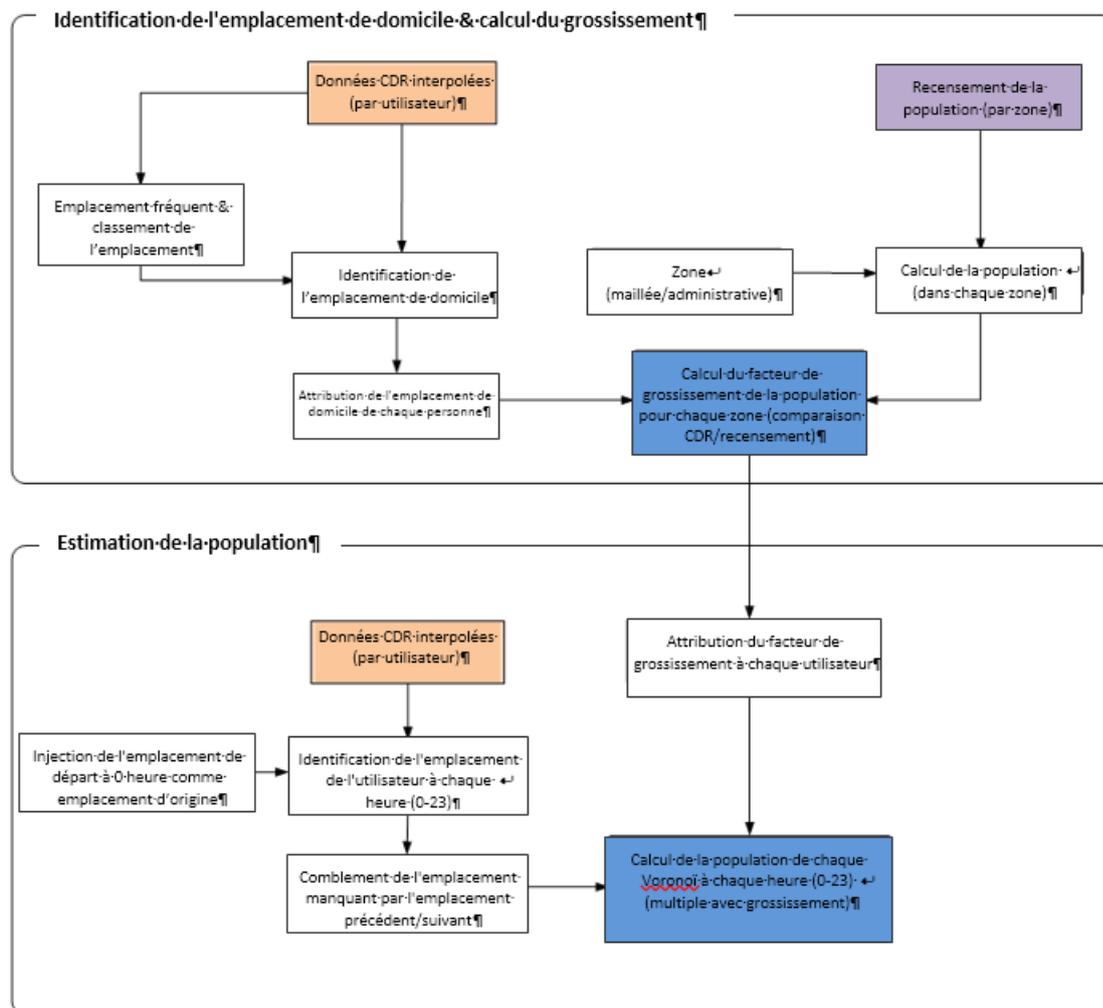
Hive est un ensemble d'entrepôts de données qui s'exécute au-dessus d'Hadoop. Les utilisateurs doivent être familiers et à l'aise avec l'utilisation de SQL pour effectuer des requêtes ad hoc, réaliser des compressions et procéder à des analyses de données. Des interfaces Web GUI et Java Database Connectivity (JDBC) sont fournies pour interagir avec Hive en émettant des requêtes dans un langage de type SQL appelé HiveQL.

7.7 Estimation de la population dynamique

Il est essentiel de connaître la répartition de la population plutôt que la répartition des abonnés au téléphone mobile dans une zone donnée pendant les heures de la journée ou dans n'importe quel intervalle de temps pour calculer les trajectoires et les tendances des mouvements, et le défi consiste à savoir comment: 1) grossir les données CDR ou le nombre d'abonnés pour refléter l'état de la population réelle à partir du recensement de la population et 2) représenter la trajectoire des personnes ou savoir où et quand les personnes sont localisées. La procédure de grossissement est décrite à la Figure 7.13.

Selon cette procédure, les données CDR sont utilisées pour extraire les emplacements de domicile des utilisateurs de téléphones mobiles en fonction des emplacements des stations de base. Les emplacements de domicile des utilisateurs de téléphones mobiles sont réaffectés à l'aide des POI et des données complémentaires sur l'utilisation des terres et la répartition des habitations autour de chaque station de base. Les données des utilisateurs de téléphone mobile ainsi réaffectées sont ensuite agrégées pour chaque zone de recensement (secteur) de la population. Le facteur de grossissement, qui décrit le nombre de personnes représentées par un utilisateur de téléphone mobile, peut être calculé en comparant la population réelle d'une zone de recensement avec le nombre agrégé d'utilisateurs de téléphone mobile. Par exemple, un utilisateur de téléphone mobile peut représenter 10 personnes dans une zone de recensement donnée. Le facteur de grossissement est également utilisé pour calculer la population dynamique sur chaque période.

Figure 7.13: Identification de l'emplacement de domicile avec grossissement de la population

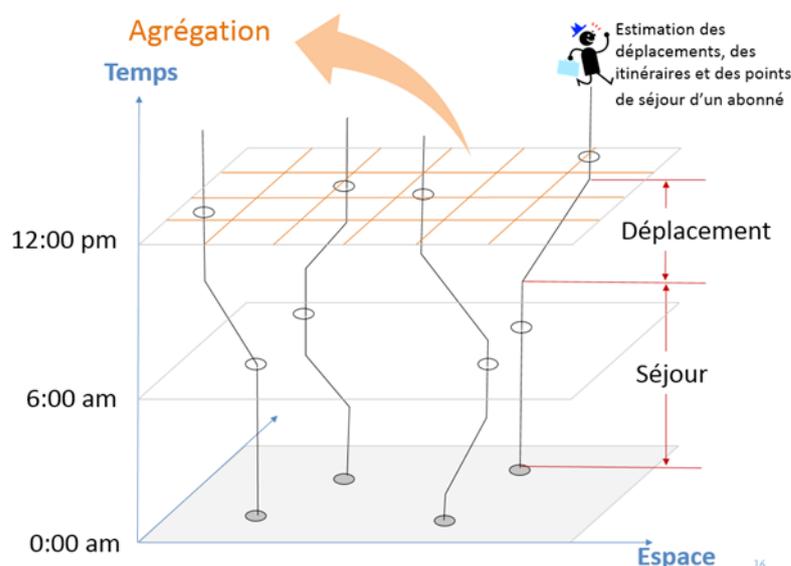


Source: UIT

La population journalière dynamique (population ambiante), contrairement à la population résidentielle, prend en compte les mouvements des individus dans une zone donnée, comme le montre la Figure 7.14. Par exemple, aucune personne ne vit sur une route principale, mais elle les emprunte de temps à autre. La cartographie d’une population résidente n’affiche normalement personne sur la route, tandis que la population ambiante révèle la présence d’individus en fonction de facteurs spécifiques à la route.

L’analyse des données CDR permet d’estimer les emplacements de domicile et les autres points de séjour ainsi que les itinéraires de déplacement et les données d’horodatage des utilisateurs de téléphone mobile de manière à refléter pleinement la trajectoire, à l’aide des facteurs de grossissement. La population dynamique peut ainsi être calculée à un emplacement donné et dans un intervalle de temps donné.

Figure 7.14: Agrégation des données journalières aux fins d'une estimation de la population dynamique



Source: UIT

Les données CDR de chaque téléphone mobile sont traitées pour identifier l'emplacement sur chaque période de temps, aux fins de l'estimation de la population dynamique. Les valeurs Voronoï sont ainsi rattachées à chaque tranche horaire de chaque téléphone mobile. Dans le cas où il n'y a pas d'enregistrement d'appel sur une certaine période, l'emplacement précédent ou suivant sera utilisé à la place pour compléter toutes les données de localisation manquantes. Enfin, le nombre de personnes est calculé pour chaque Voronoï et valeur temporelle, en combinaison avec le facteur de grossissement de chaque utilisateur. Dans le processus de calcul, chaque téléphone mobile aura son propre facteur de grossissement de manière à refléter de façon plus détaillée les emplacements et les mouvements de la population dans le temps.

7.8 Téléphone mobile et clonage de carte SIM

Le clonage de téléphone mobile⁷ est une technique utilisée pour copier des données privées (vol d'identité) d'un téléphone mobile vers un autre téléphone mobile, qui devient alors la réplique exacte du téléphone d'origine. Par conséquent, les appels peuvent être effectués et reçus par les deux téléphones, mais seul l'abonné légitime est facturé car l'opérateur de réseau mobile n'a pas les moyens de distinguer le téléphone légitime du téléphone «cloné». Si un abonné est surpris de recevoir une énorme facture, il est fort probable que son téléphone ait été cloné. De nombreux téléphones mobiles, GSM ou CDMA, courent le risque d'être clonés.

Il existe plusieurs types d'informations d'identité spécifiques à l'utilisateur en fonction du système de réseau mobile. Sur le réseau GSM, chaque combiné possède un numéro IMEI unique et nécessite l'insertion d'une carte SIM pour pouvoir fonctionner. Les cartes SIM peuvent être clonées et placées dans d'autres combinés, et parce que les combinés peu coûteux peuvent souvent avoir le même IMEI (pour réduire les coûts de production), le clonage des cartes SIM génère un sérieux problème de fraude. Sur les réseaux CDMA, chaque combiné possède un numéro ESN (numéro de série électronique) et MIN (numéro d'identification mobile) unique qui est gravé sur sa puce et le clonage de ce type de combiné nécessite de remplacer ces numéros d'identité.

Le clonage peut être détecté au moment de la duplication d'identité lorsque l'utilisation du téléphone mobile est transférée d'un endroit à un autre endroit à des vitesses improbables (piège de la vitesse).

⁷ <https://www.movzio.com/howto/cell-phone-cloning/>

Par exemple, si un appel est émis à Conakry et qu'un autre appel est passé depuis le Mali 5 minutes plus tard à 300 kilomètres de là, on peut en déduire qu'il y a deux appareils avec la même identité sur le réseau. Lorsqu'il détecte un clonage de téléphone mobile, l'opérateur concerné bloque souvent les deux téléphones de sorte que l'abonné légitime contacte le service clientèle de son opérateur pour se plaindre. Les enregistrements détaillés des appels tels que numéro d'identité, heure de début d'appel et emplacement de station de base sont des éléments clés essentiels pour l'identification des cas de clonage. Toutefois, une telle analyse requiert le traitement d'un grand nombre de données, et, partant, la présence de systèmes hautes performances.

Dans le cadre de ce projet, des tentatives ont été faites pour détecter le clonage de téléphone mobile sur les réseaux GSM à partir des données CDR en mettant à profit la technologie des mégadonnées et le traitement spatial. Les résultats sont représentés à la section 8.5 de ce rapport.

8 Résultats des analyses

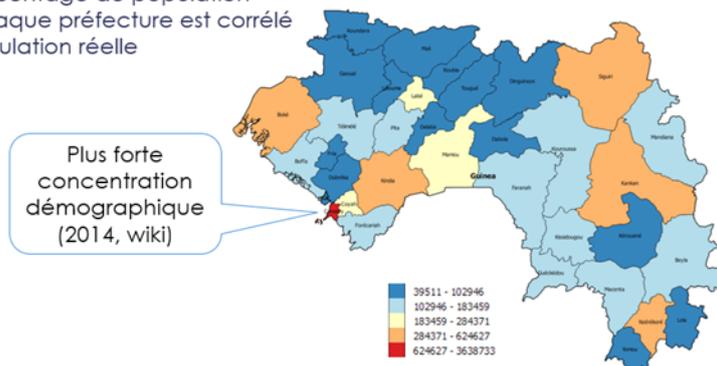
8.1 Population d'utilisateurs de téléphone mobile par préfecture

Cette section est consacrée à la répartition de la population à partir des données CDR. La Figure 8.1 affiche le nombre d'utilisateurs de téléphones mobiles estimé à partir des données CDR et à partir des données démographiques issues du recensement national de la population en République de Guinée, et confirme que la proportion d'utilisateurs de téléphone mobile dans les préfectures est en forte corrélation avec la proportion des populations réelles obtenue à partir des données du recensement. La Figure 8.2 montre quels sont les emplacements les plus visités par les personnes qui vivent à Conakry (Coyah, Kindia et Forécariah arrivent en tête). La Figure 8.3 révèle que Coyah est le deuxième emplacement le plus visité par les mêmes personnes.

Figure 8.1: Population d'utilisateurs de téléphone mobile par préfecture

Population d'utilisateurs de téléphone mobile par préfecture

- Le pourcentage de population dans chaque préfecture est corrélé à la population réelle



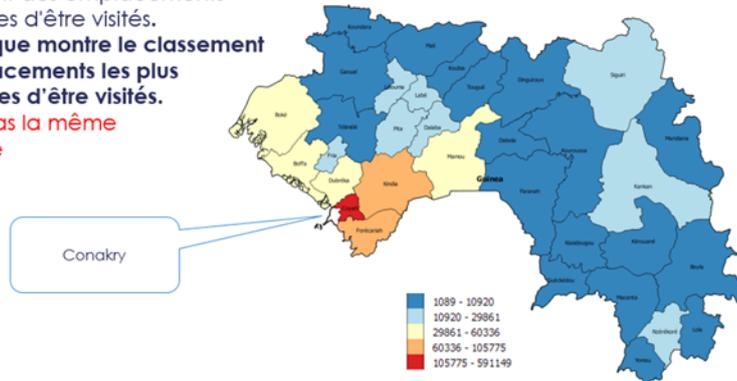
Population par préfecture (sur la base des emplacements fréquents)
Le petit cercle correspond à l'emplacement de la station de base (pas de grossissement)

Source: UIT

Figure 8.2: Emplacements les plus visités par les personnes vivant à Conakry

Conakry - Emplacements les plus visités

- Focale sur les habitants de Conakry.
- Classement des emplacements susceptibles d'être visités.
- **Le graphique montre le classement des emplacements les plus susceptibles d'être visités.**
- **N'inclut pas la même préfecture**



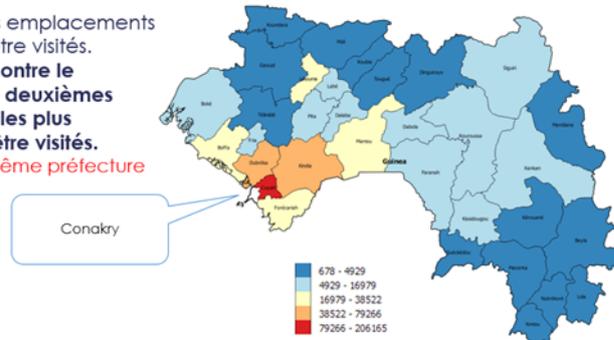
- La majorité des habitants de Conakry visitent **Coyah, Kindia et Forécariah**

Source: UIT

Figure 8.3: Emplacements les plus visités par les personnes vivant à Conakry

Conakry - Deuxièmes emplacements les plus visités

- Focale sur les habitants de Conakry.
- Classement des emplacements susceptibles d'être visités.
- **Le graphique montre le classement des deuxièmes emplacements les plus susceptibles d'être visités.**
- **N'inclut pas la même préfecture**



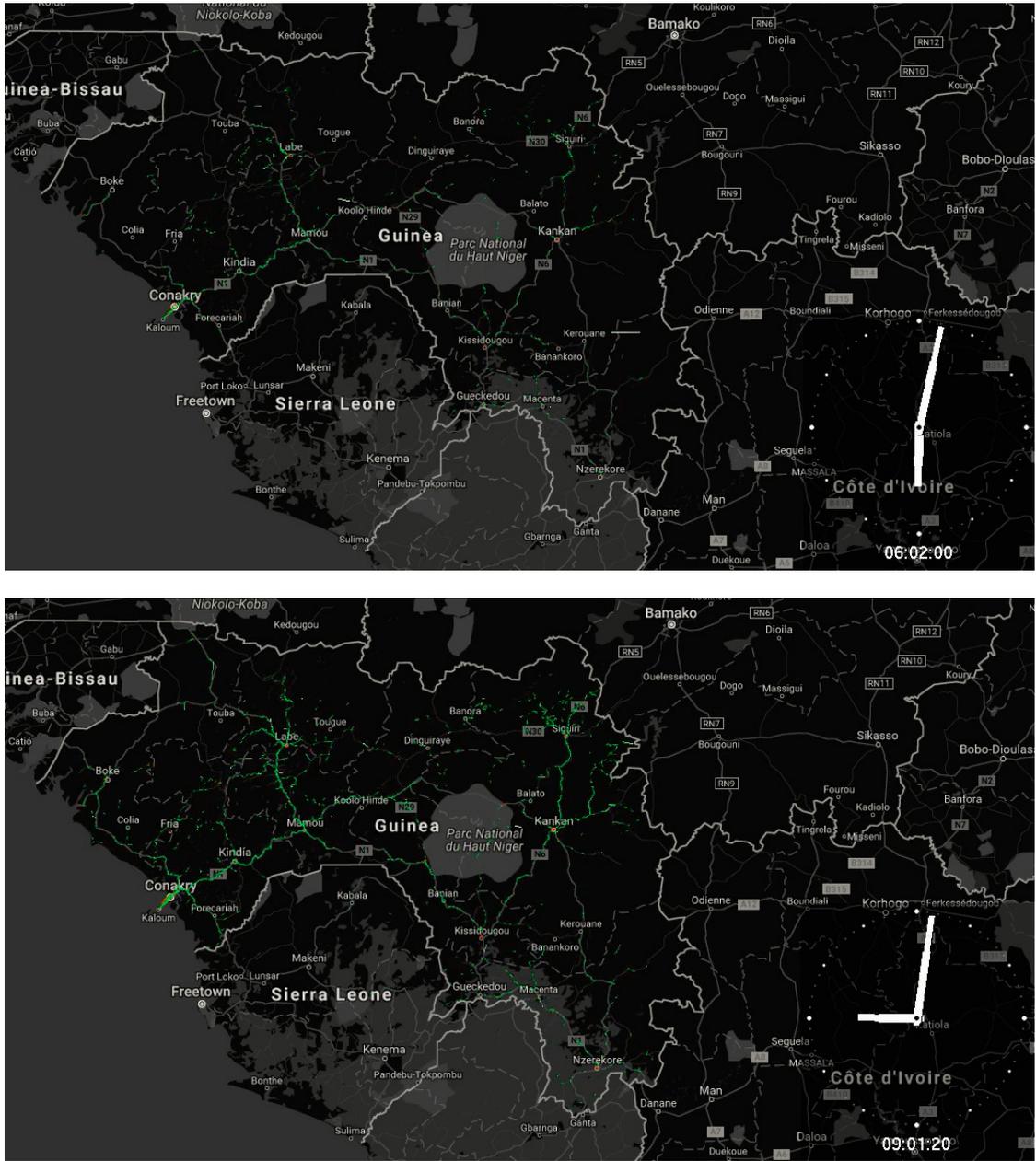
- De nombreux habitants de **Conakry** visitent **Coyah (deuxième emplacement le plus visité)**

Source: UIT

8.2 Visualisation des flux de population grâce au résultat d'interpolation

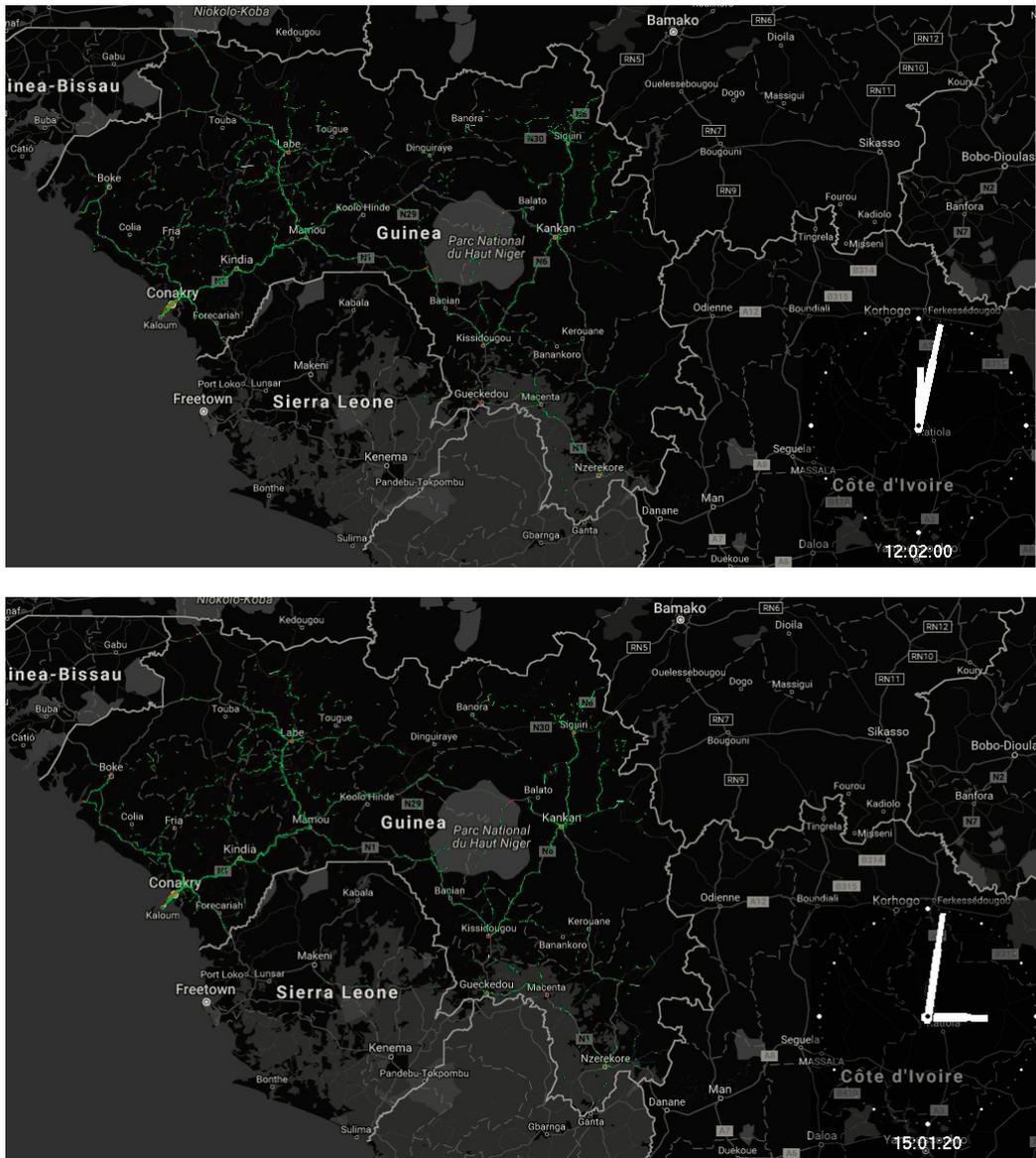
Cette section présente la visualisation des données CDR interpolées sur la base du processus présenté à la section 7. Les Figures 8.4a et 8.4b montrent les flux de population reconstitués à partir des résultats d'interpolation à 6 heures, 9 heures, 12 heures et 15 heures. A Monrovia, on observe moins de personnes le matin et plus de personnes l'après-midi. Les résultats d'interpolation comportent 11 colonnes: Identité d'utilisateur, Date, Séquence du trajet, Type de mobilité, Mode de transport, Distance totale, Durée totale, Date de début, Date de fin, Total des points et Listes de points (voir la section 7.5). Les données CDR sont interpolées à intervalles d'1 minute.

Figure 8.4a: Flux de population à 6 heures et 9 heures



Source: UIT

Figure 8.4b: Flux de population à 12 heures et 15 heures

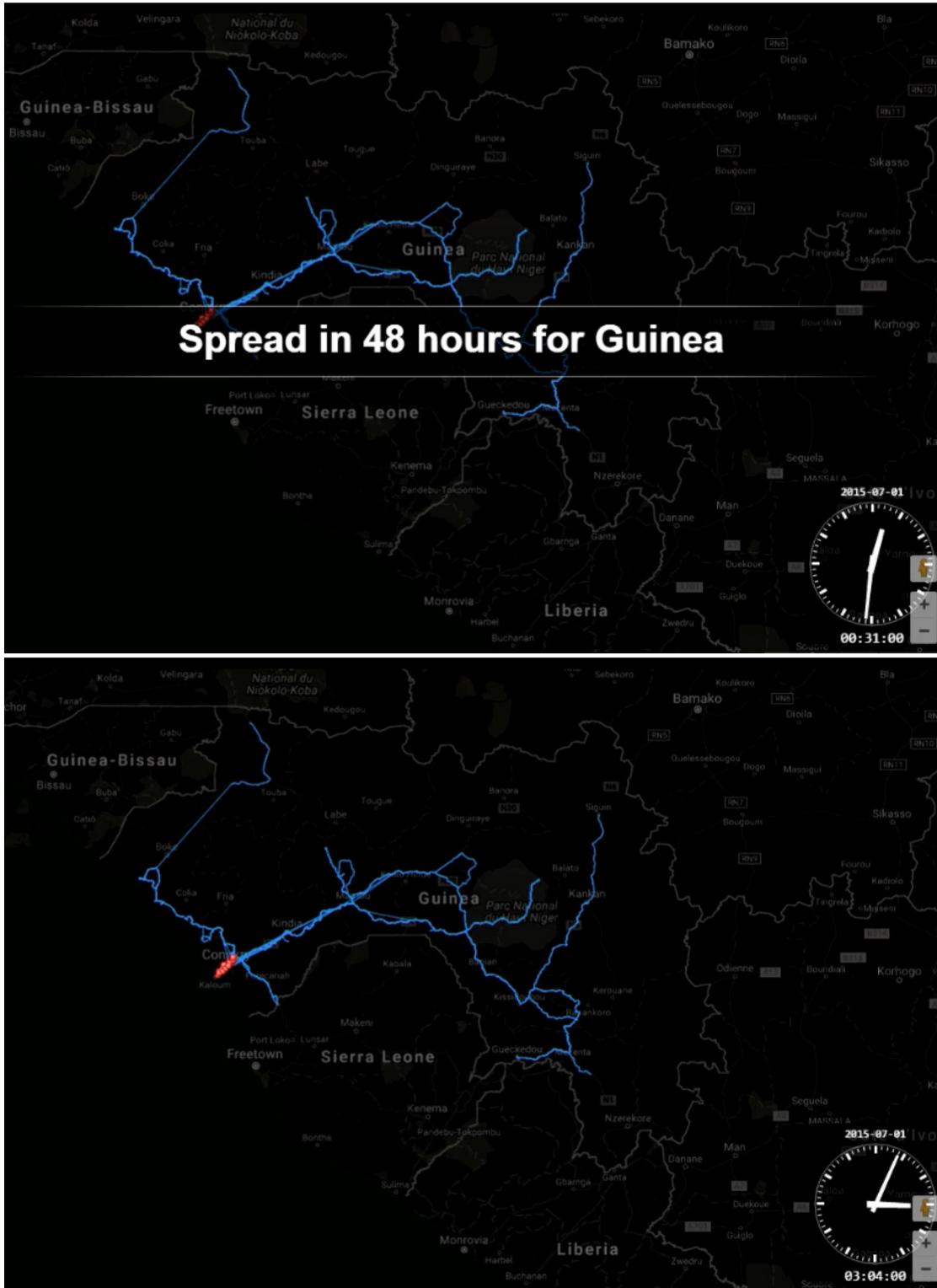


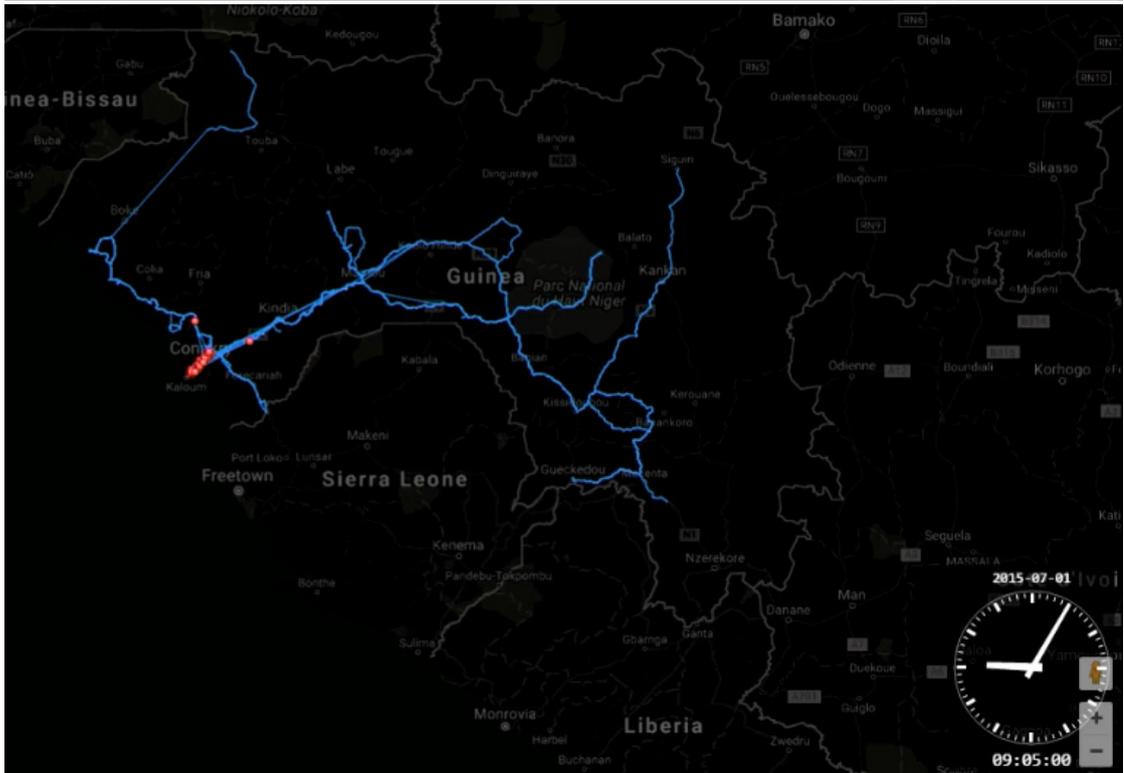
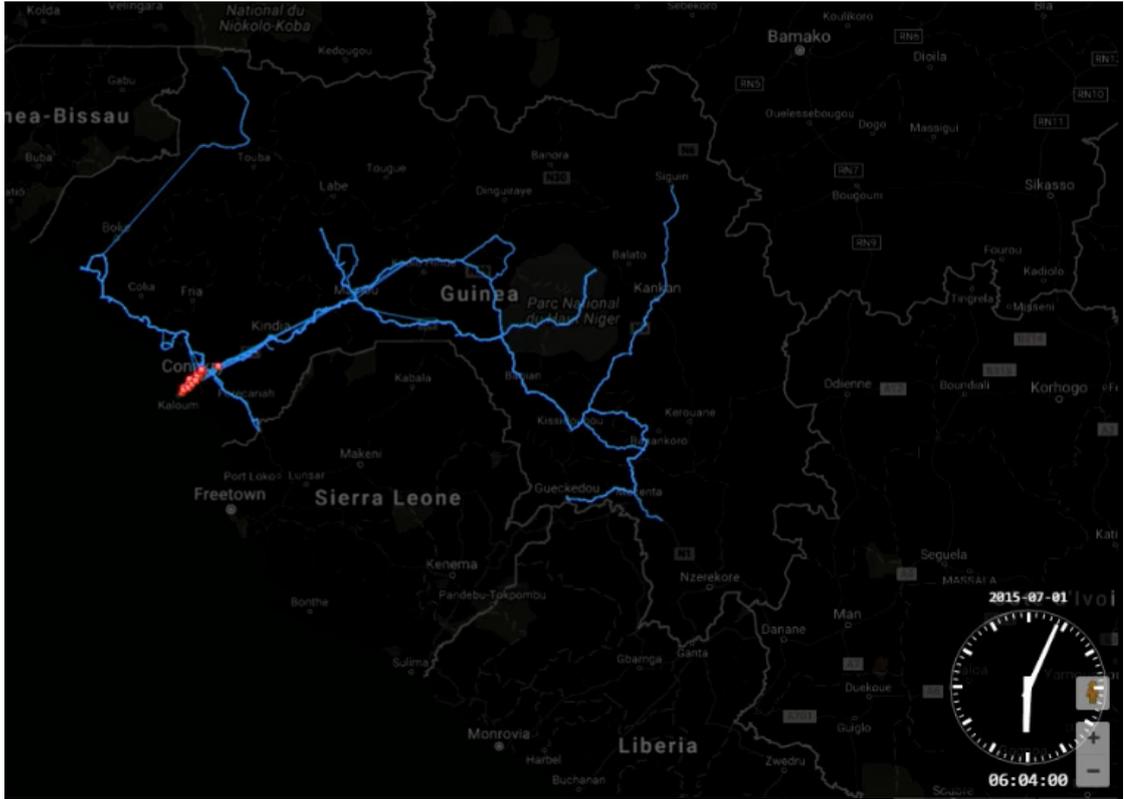
Source: UIT

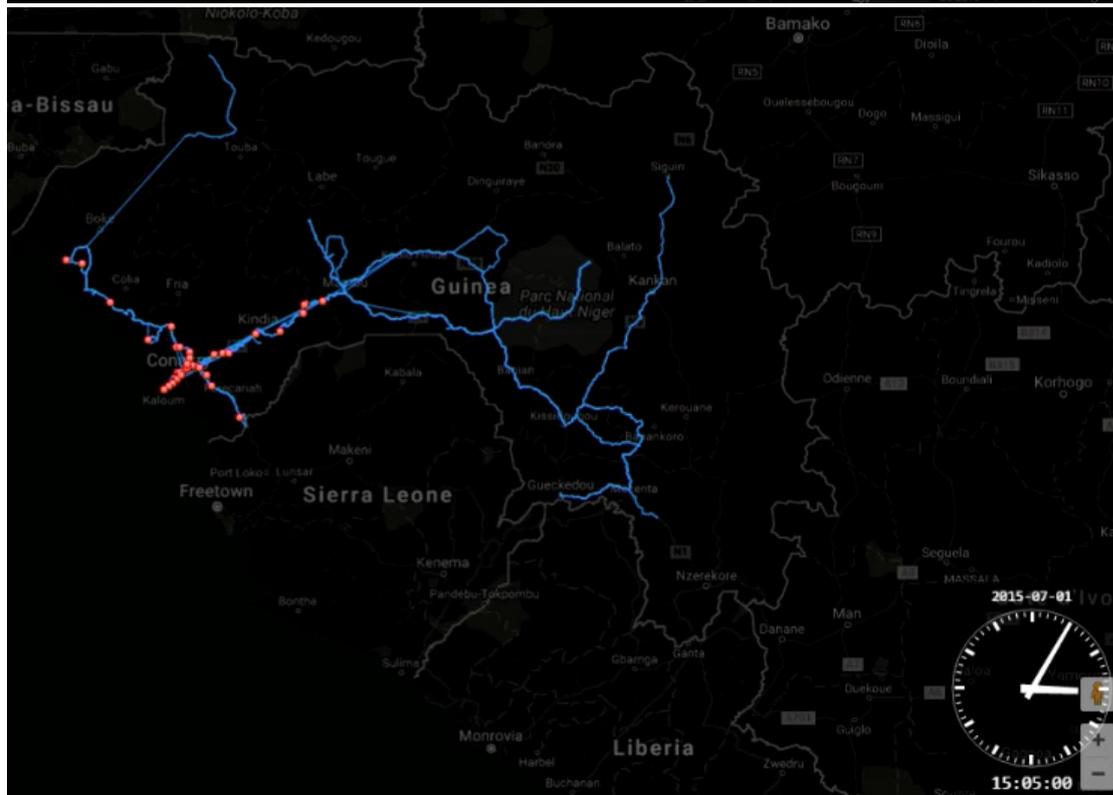
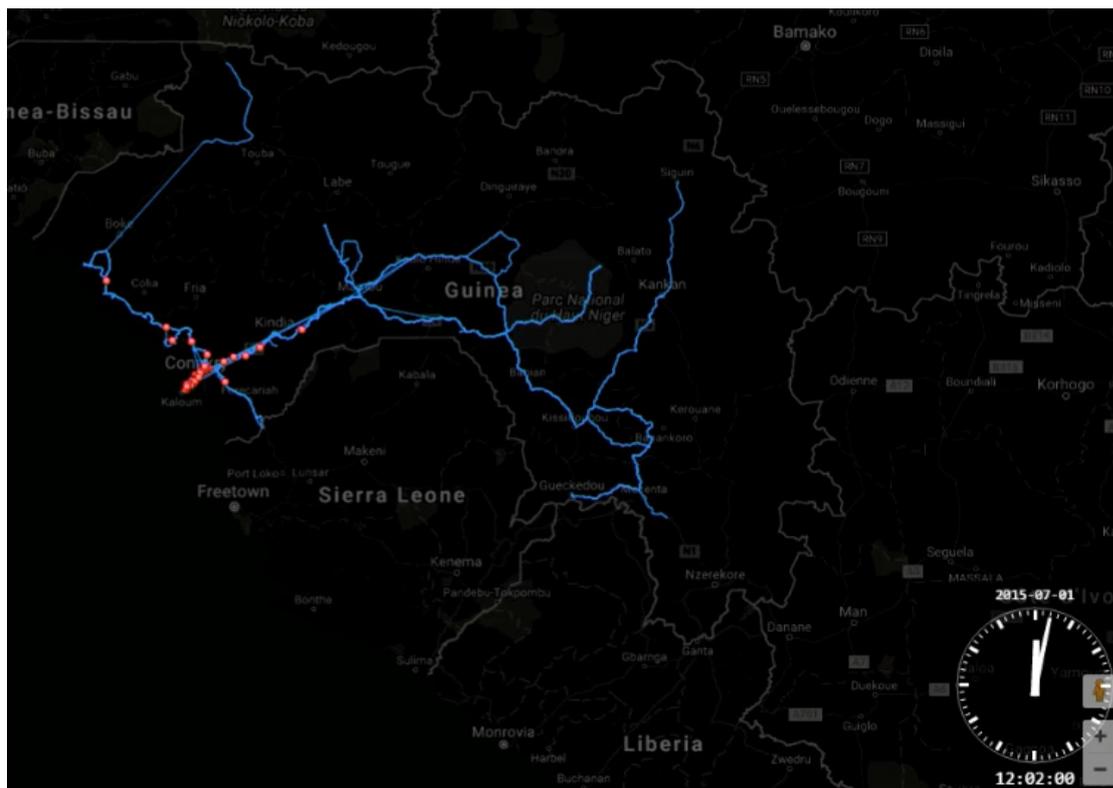
8.3 Mouvements de population dans une ville (sur 48 heures)

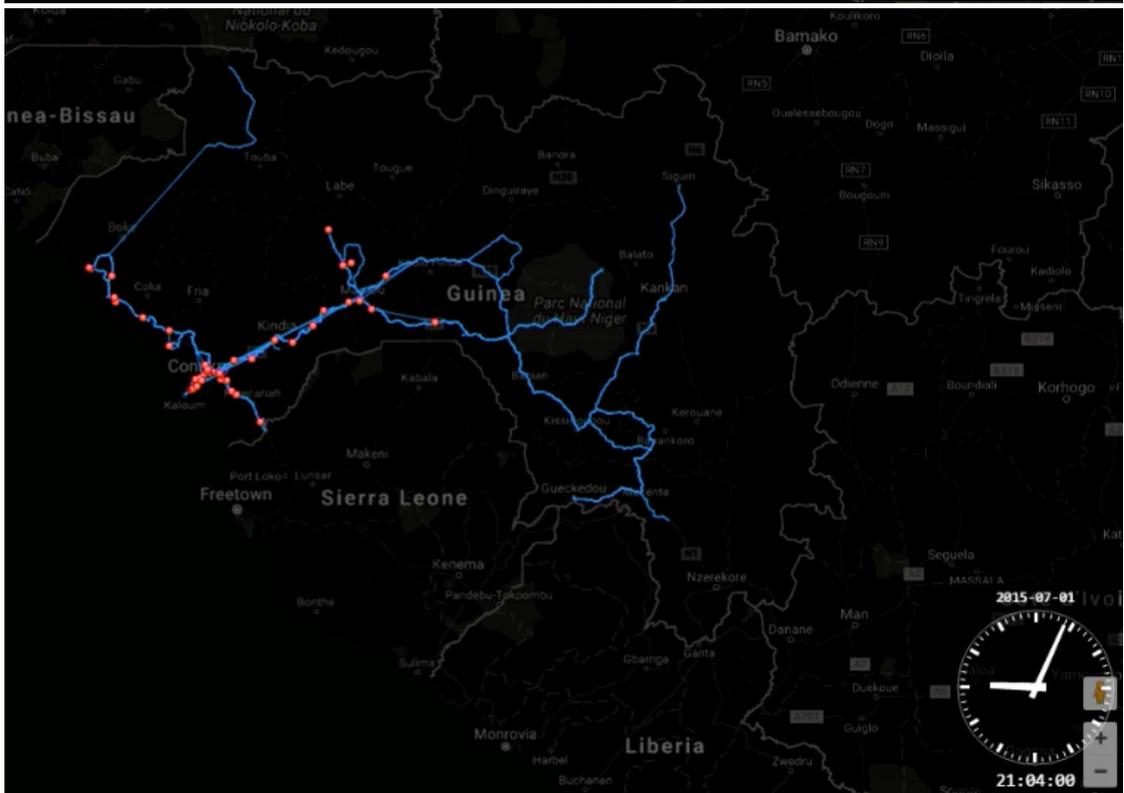
La Figure 8.5 montre les mouvements de population dans une ville par tranche de 3 heures sur une période de 48 heures. Force est de constater, en cas d'incidents dans la ville, qu'une maladie transmissible pourrait se propager rapidement compte tenu de la mobilité de ses habitants.

Figure 8.5: Flux de population depuis une ville donnée (sur 48 heures)



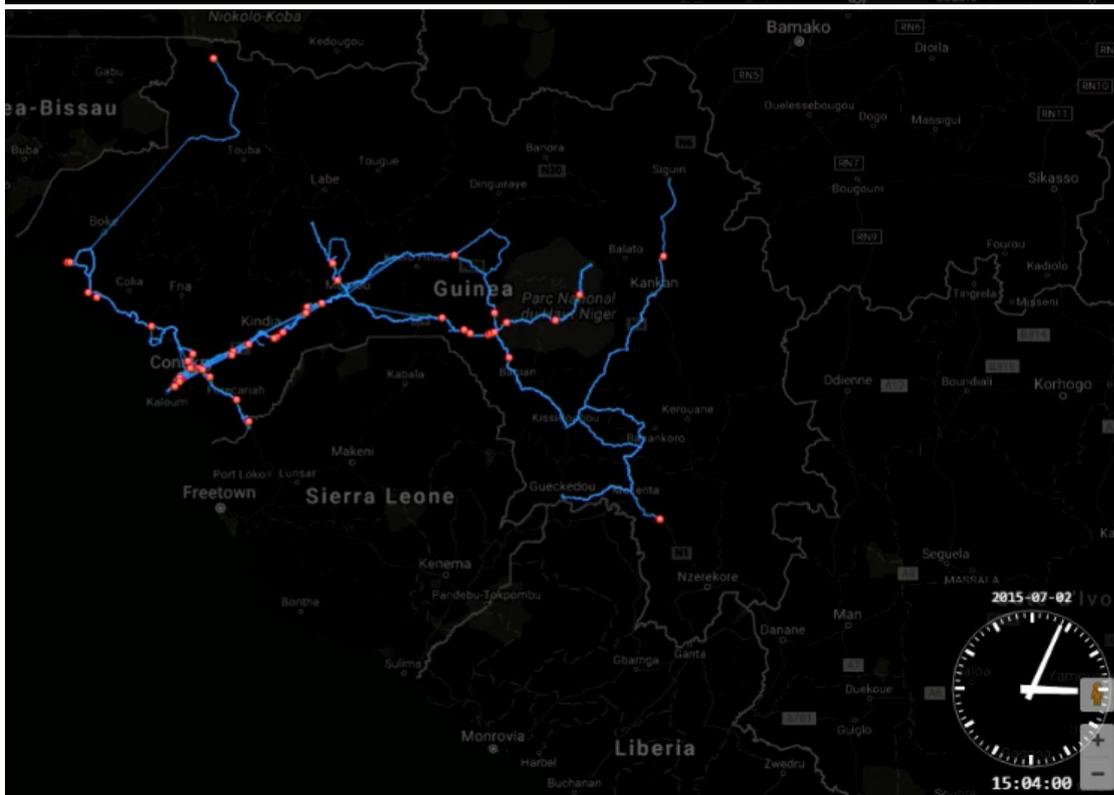
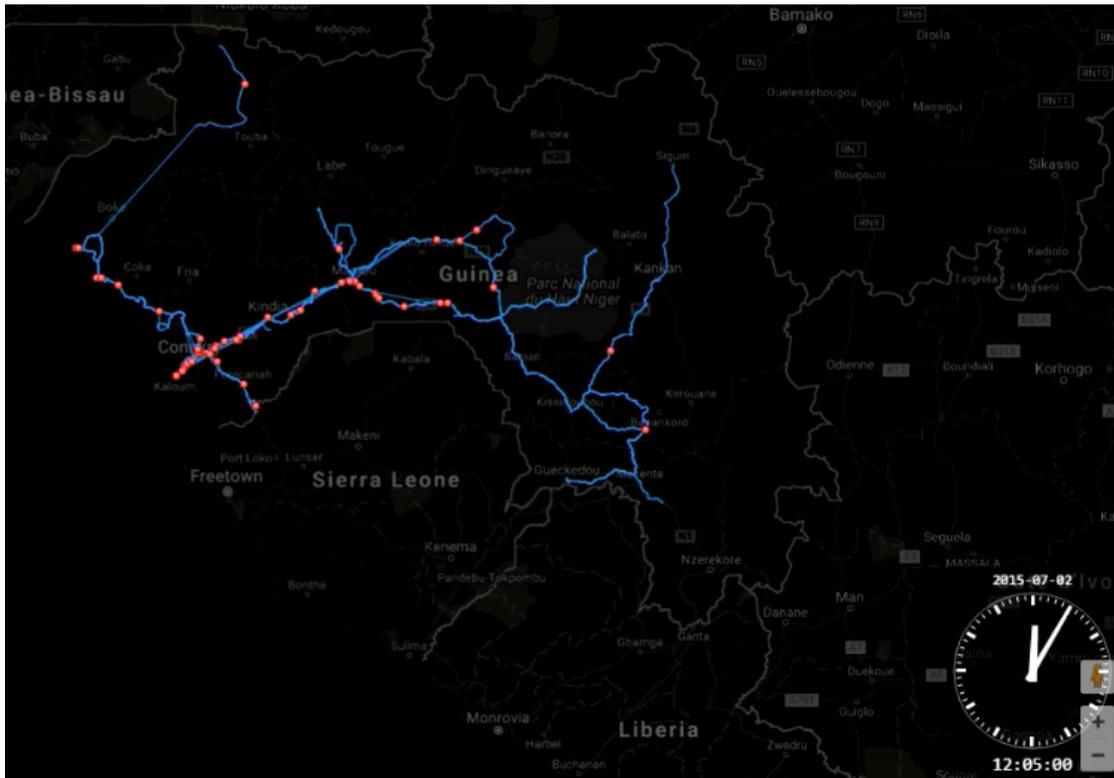


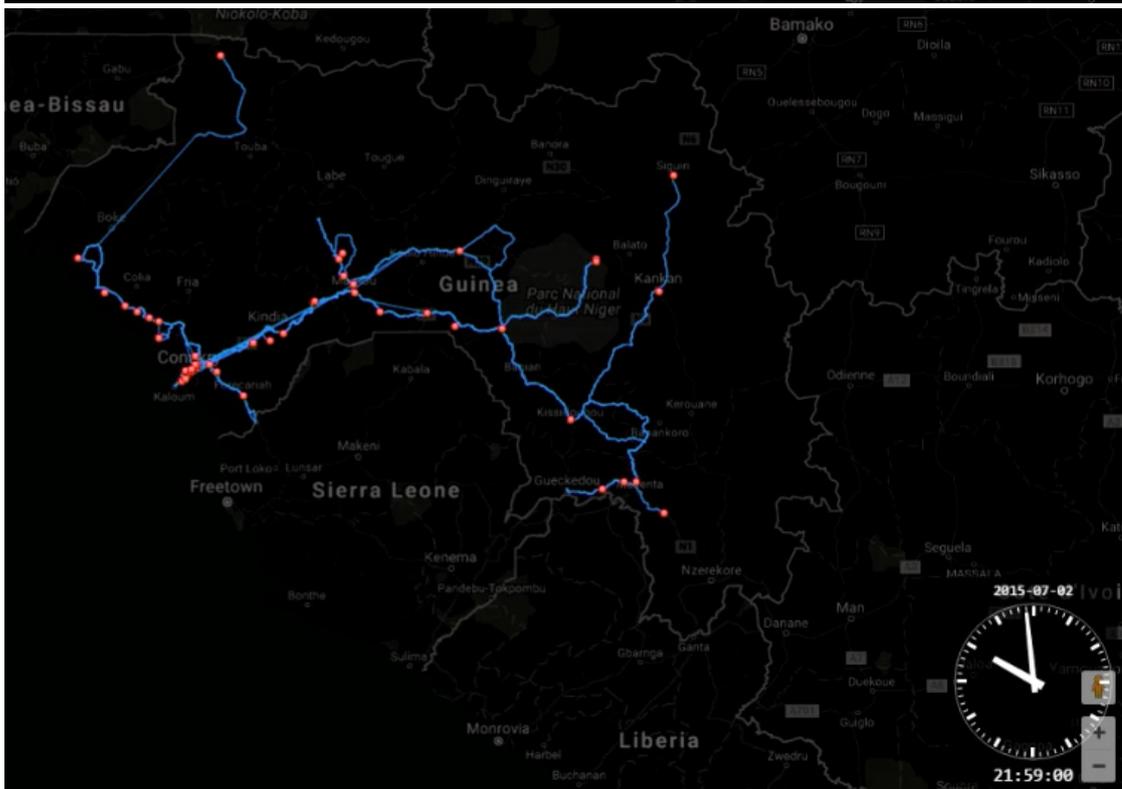












Source: UIT

8.4 Surveillance des personnes ayant traversé une zone à risque

Les Figures 8.6a, 8.6b, 8.7a et 8.7b montrent la répartition des personnes qui ont traversé ou non une zone à risque. La zone à risque est représentée en bleu. Sur la Figure 8.6A, les gros points rouges à

l'intérieur du polygone bleu désignent les personnes qui se trouvent dans la zone à risque; les petits points jaunes à l'extérieur du polygone représentent les personnes qui se trouvent à l'extérieur de la zone à risque à une heure donnée. Les Figures 8.6B, 8.7A et 8.7B montrent comment les personnes situées dans la zone à risque (à l'heure de la Figure 8.6A) se répartissent sur les heures qui suivent.

Figure 8.6A: Répartition des personnes à l'intérieur/extérieur de la zone à risque



Source: UIT

Figure 8.6B: Surveillance des personnes ayant traversé une zone à risque (10 heures plus tard)



Source: UIT

Figure 8.7A: Où sont-elles allées dans les 48 heures? (12 heures après)



Source: UIT

Figure 8.7B: Où sont-elles allées dans les 48 heures? (47,5 heures après)



Source: UIT

Comme le montrent les Figures 8.7A et 8.7B, les personnes peuvent, en 48 heures, se retrouver loin de l'emplacement d'origine, ce qui met en péril une très grande zone dans le cas d'une épidémie d'Ebola.

8.5 Tentatives de détection des cas de clonage de téléphones et de cartes SIM

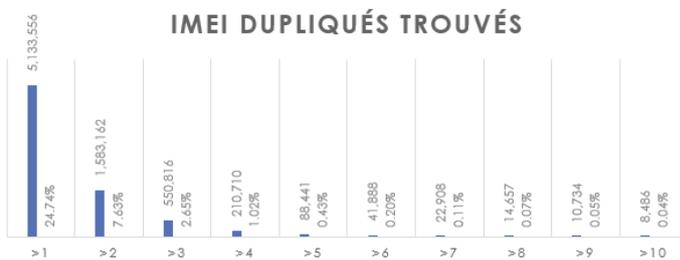
Comme mentionné à la section 7.8 ci-avant, les données CDR peuvent être utilisées pour déceler les cas de clonage. La Figure 8.8 montre la proportion d'IMEI (identifiants uniques de téléphonie mobile dans les données CDR) provenant d'enregistrements d'appels depuis plusieurs SIM ou identités internationales d'abonné mobile (IMSI). Le graphique du haut révèle que le nombre d'identifiants uniques IMEI dans les données CDR du projet est près de deux fois plus important que le nombre d'identifiants uniques IMSI et que près de 22 980 numéros de téléphones mobiles uniques ont été utilisés avec plus de sept cartes SIM (IMSI) indiquant des cas probables de clonage mobile. Le présent projet a mis la focale sur les identifiants IMEI utilisés avec plus de sept IMSI.

Le graphique du bas affiche deux enregistrements d'appel provenant d'un IMEI identique avec différents identifiants IMSI. Leurs localisations sont trop éloignées l'une de l'autre pour qu'un appel ait pu être passé à un endroit et le deuxième à un autre en moins d'une minute. Le traitement des données CDR au moyen des identifiants IMSI et IMEI peut donc aider à déceler les cas de clonage. C'est cette méthode qui a été utilisée pour le projet.

Figure 8.8: Détection des cas de clonage (clonage IMEI)

Tentative de détection de clonage de téléphone mobile (GSM) au moyen de la technologie des mégadonnées et du traitement spatial

- **Clone IMEI (même IMEI, IMSI différents)**
 - % d'IMEI utilisés avec au moins 2 IMSI (SIM)



Enregistrements détaillés des appels	
Total des enregistrements	5,207,803,431
Taille du fichier	1,458 GB
Nombre de jours	61
Id unique (IMSI)	9,827,544
Id unique (IMEI)	20,746,397
Emplacement unique	1954

- Focale sur "IMSI trouvés > 7", liste des heures appel par appel
- Contrôle des id. IMSI différents, de la distance spatiale et des heures différentes
- Procédure rapide sur la plate-forme cloud Hadoop avec support spatial (<1h)

Tentative de détection de clonage de téléphone mobile (GSM) au moyen de la technologie des mégadonnées et du traitement spatial

- **Clone IMEI (même IMEI, IMSI différents)**
 - Focale sur "IMSI trouvés > 7", liste des heures appel par appel
 - Contrôle des id. IMSI différents, de la distance spatiale et des heures différentes
- **Highlight**
 - Localisations différentes avec distance importante
 - 4 292 IMEI ont été identifiés comme clones IMEI possibles

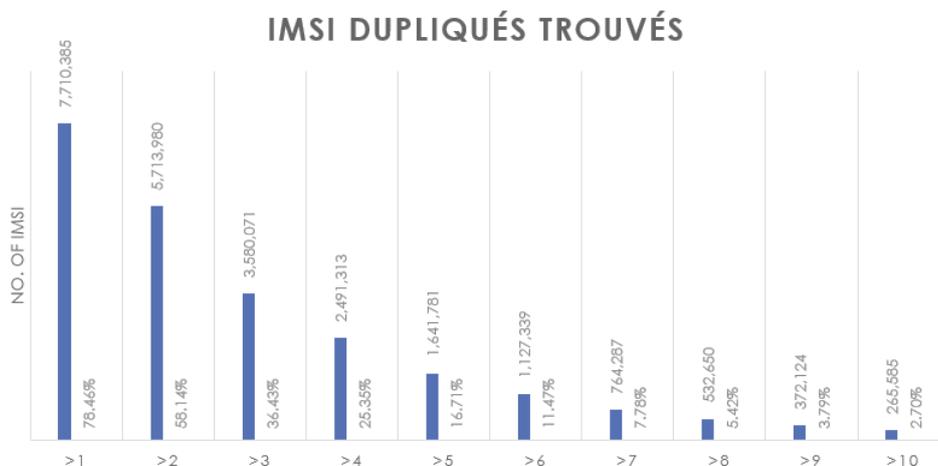
IMEI	TIME	PREV_TIME	SITE	PREV_SITE	IMSI	PREV_IMSI	DISTANCE (km)	TIME_DIFF (min)
3e6a5	2015-06-01 10:08:21	2015-06-01 10:04:10	31086	20252	157e949cf	3a7a44dc72cb	227.97	4.18
3e6a5	2015-06-01 10:46:57	2015-06-01 10:30:37	31086	20201	157e949cf	bb6182663b23	169.29	16.33
3e6a5	2015-06-01 10:49:11	2015-06-01 10:46:57	20201	31086	bb618266	157e949cf6895	169.29	2.23
3e6a5	2015-06-01 13:28:12	2015-06-01 13:14:21	20252	31086	3a7a44dc7	157e949cf6895	227.97	13.85
3e6a5	2015-06-01 13:44:18	2015-06-01 13:28:12	30092	20252	157e949cf	3a7a44dc72cb	227.04	16.1
3e6a5	2015-06-01 13:52:51	2015-06-01 13:44:18	10742	30092	7bd5ef153	157e949cf6895	196.39	8.55
3e6a5	2015-06-01 13:56:58	2015-06-01 13:52:51	20252	10742	3a7a44dc7	7bd5ef153589	164.68	4.12
3e6a5	2015-06-01 14:12:55	2015-06-01 14:02:45	31085	20251	157e949cf	3a7a44dc72cb	227.97	10.17
3e6a5	2015-06-01 14:27:19	2015-06-01 14:12:55	10452	31085	c797e88e3	157e949cf6895	195.58	14.4
3e6a5	2015-06-01 14:46:16	2015-06-01 14:30:52	20252	10452	3a7a44dc7	c797e88e3a96	162.35	15.4
3e6a5	2015-06-01 14:54:10	2015-06-01 14:48:07	10452	20252	c797e88e3	3a7a44dc72cb	162.35	6.05

Source: UIT

La Figure 8.9, dans sa partie supérieure, montre une tentative de détection de clonage de carte SIM: plus de 7,7 millions de cartes SIM uniques ont été utilisées sur au moins deux combinés pendant la même période et plus de 764 287 cartes SIM ont été utilisées sur sept combinés, ce qui peut probablement être attribué à un clonage de carte SIM. Dans le cadre du projet, la focale et les filtres ont été placés sur les identifiants IMSI trouvés sur plus de sept IMEI.

Figure 8.9: Détection des cas de clonage (clonage IMSI)

- **Clone IMSI (SIM)**
 - Les factures sont adressées à l'abonné mobile légitime (fraude)
 - % des id. IMSI trouvés sur 2 IMEI ou plus (combinés)



- **Clone IMSI (SIM)**
 - Focale sur "IMSI trouvés > 7", liste des heures appel par appel
 - Contrôle des id. IMSI différents, de la distance spatiale et des heures différentes
- **Highlight**
 - Localisations différentes avec distance importante
 - 73 586 IMSI ont été identifiés comme clones IMSI possibles (clone sim)

IMSI	TIME	PREV_TIME	SITE	PREV_SITE	IMEI	PREV_IMEI	DISTANCE (km)	TIME_DIFF (min)
d3afbe658515ec99b4	2015-07-20 12:32:07	2015-07-20 12:30:10	52854	22413	c5bff5e63	48e5c2d72628	685.39	1.95
801ca9ffe40a78d717	2015-07-15 23:33:21	2015-07-15 23:33:05	51301	13304	206842fa5	ce06f0aea734	657.67	0.27
6890438b81b32631a	2015-06-05 17:58:52	2015-06-05 17:55:01	35842	495	d646044df	e28a6166c1d3	632.24	3.85
6bb32097834efda08	2015-07-05 10:03:50	2015-07-05 10:03:02	35842	492	2c3092703	66b5575d71a3	632.24	0.8
6bb32097834efda08	2015-07-05 10:41:08	2015-07-05 10:39:59	492	35842	66b5575d	2c3092703f1c6	632.24	1.15
6bb32097834efda08	2015-07-05 10:42:25	2015-07-05 10:41:08	35842	492	2c3092703	66b5575d71a3	632.24	1.28
6bb32097834efda08	2015-07-05 11:05:08	2015-07-05 11:01:01	35842	492	2c3092703	66b5575d71a3	632.24	4.12
6bb32097834efda08	2015-07-05 11:25:00	2015-07-05 11:23:00	492	35842	66b5575d	2c3092703f1c6	632.24	2

Source: UIT

La Figure 8.9, dans sa partie inférieure, montre (en jaune) un enregistrement effectué avec un même IMSI, mais avec différents identifiants IMEI. Leurs localisations sont trop éloignées l'une de l'autre pour que les deux appels aient pu être émis avec la même carte SIM. Sur cette série de données, 73 586 IMSI ont été identifiés comme clones possibles (clonage de carte SIM).

8.6 Analyse transfrontière

Cette section présente les résultats de l'analyse des mouvements transfrontières. Elle montre comment les données CDR peuvent être utilisées pour comprendre les mouvements de population dans les régions frontalières des pays voisins.

Il existe deux types de mouvements de population transfrontières:

- le mouvement transfrontière **entrant**, lorsqu'une personne retourne dans son pays d'origine depuis un pays voisin; et
- le mouvement transfrontière **sortant**, lorsqu'une personne quitte son pays d'origine pour visiter un pays voisin.

L'analyse transfrontière utilise l'IMEI comme indicateur clé. Les données CDR du Libéria et de la Sierra Leone ont été fusionnées avec les données de la Guinée pour suivre les mouvements transfrontières. A l'issue de cette fusion, le pays d'origine de l'utilisateur est identifié en fonction de la fréquence des emplacements appelés.

8.6.1 Ensemble de données CDR avec IMEI transfrontière

La Figure 8.10 présente les informations transfrontières qui sont obtenues à partir des données CDR, y compris l'IMEI qui indique la source des téléphones mobiles utilisés dans les pays voisins. Trois opérateurs de réseau mobile (MNO)- MTN, Orange et Cellcom- ont fourni des données CDR (avec un code IMEI crypté). Malheureusement, seuls les codes IMEI MNT et Cellcom ont pu être comparés. Autrement dit, les résultats d'analyse des sections suivantes ne tiennent pas compte des mouvements transfrontières des abonnés Orange. Si l'on se base sur la taille des données, le nombre d'abonnés Orange est supposé être le plus important de tous. Il n'est donc pas exclu, dans ces conditions, que certaines tendances de mouvement, représentées par les abonnés d'Orange, aient été omises.

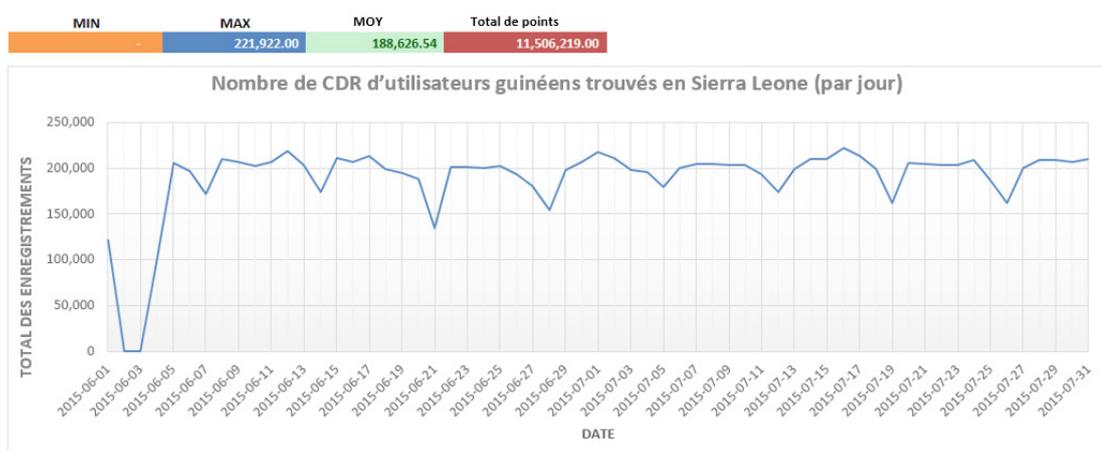
Figure 8.10: Ensemble de données CDR avec IMEI transfrontière en Guinée

NO	Pays	MNO	Voix SMS									
			Taille des données	Période	Total des enregistrements	IMEI (appelant)	IMSI (appelant)	Nb d'appellants	Horo datage	LAC/cell ID	Durée	Appel entrant/sortant
1	Guinée	MTN	303 G	Juin-juillet 2015 (sauf 19-25 juin)	1,003,898,498	O (6,979,708)	O (2,699,233)	O (2,660,893)	O	O	O	-
2		Orange	1,100 G	Juin-juillet 2015	3,915,651,949	O (10,386,715)	O (5,265,388)	O (10,285,506)	O	O	O	-
3		Cellcom	53 G	Juin-juillet 2015	279,441,946	O (3,379,974)	O (1,862,923)	-	O	O	-	-
4		Intercell	2.1 G	Juin-juillet 2015	8,811,038	-	-	O (99,531)	O	O	O	-

Source: UIT

Les Figures 8.11 et 8.12 présentent le registre des appels journaliers émis par les abonnés au téléphone mobile de la Guinée et enregistrés sur les réseaux des pays voisins (Sierra Leone et Libéria). L'activité quotidienne représente près de 188 626 appels vers la Sierra Leone contre 68 379 appels vers le Libéria, ce qui signifie que les abonnés guinéens sont deux fois plus susceptibles de se rendre en Sierra Leone qu'au Libéria.

Figure 8.11: Registre des appels journaliers des utilisateurs mobiles guinéens en visite en Sierra Leone



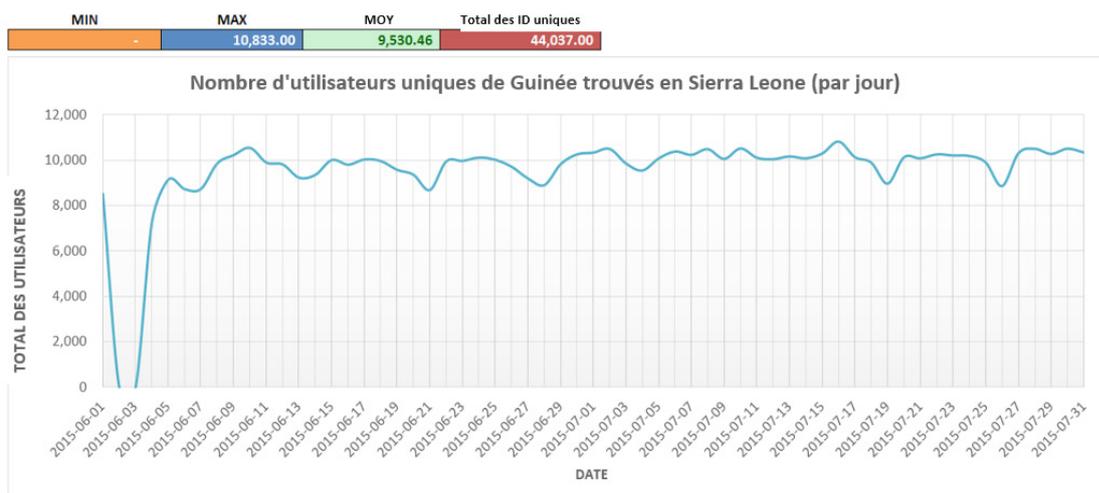
Source: UIT

Figure 8.12: Registre des appels journaliers des utilisateurs mobiles guinéens en visite au Libéria



Source: UIT

Figure 8.13: Utilisateurs mobiles guinéens uniques en visite en Sierra Leone (par jour)



Source: UIT

Les Figures 8.13 et 8.14 présentent le nombre d'utilisateurs mobiles guinéens uniques qui visitent la Sierra Leone et le Libéria (près de 9 530 utilisateurs pour la Sierra Leone et 3 872 utilisateurs pour le Libéria).

Figure 8.14: Utilisateurs mobiles guinéens uniques en visite au Libéria (par jour)

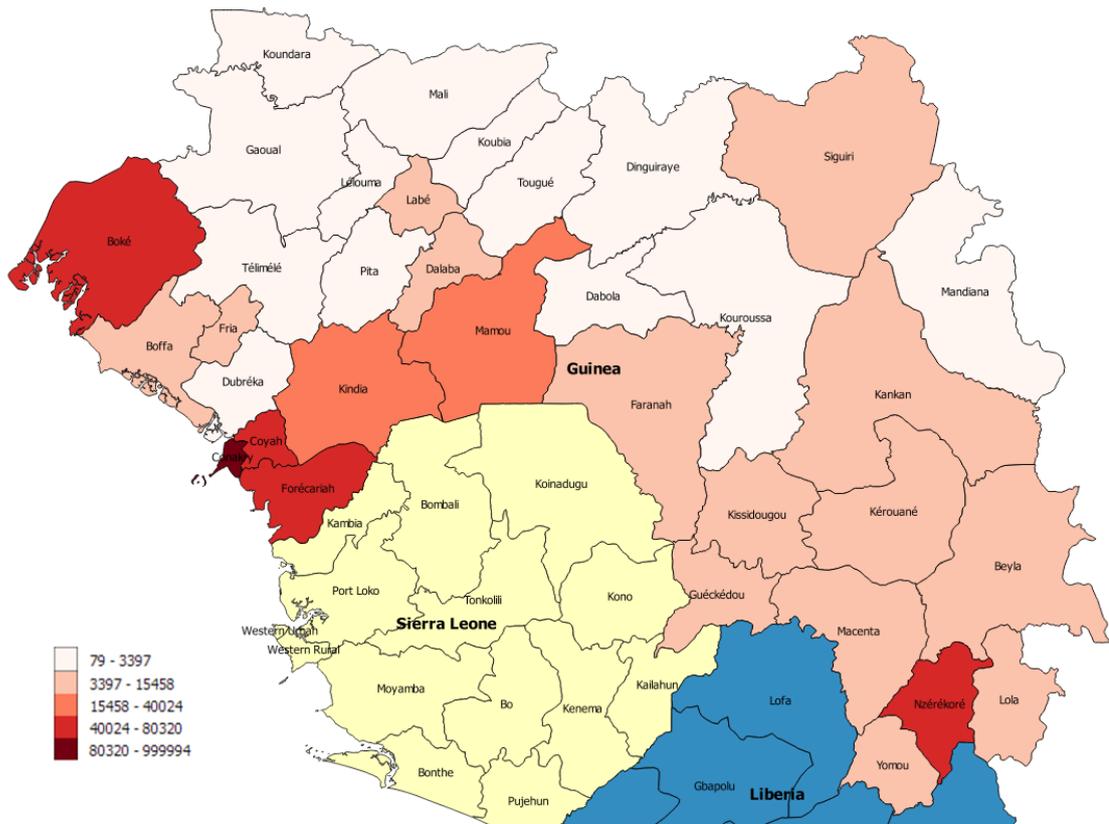


Source: UIT

8.6.3 Mouvements de population transfrontières entrants

La Figure 8.15 représente l'empreinte des mouvements transfrontières entrants, de la Sierra Leone et du Libéria vers la Guinée. Les empreintes ont été calculées à partir de l'activité d'utilisation mobile des utilisateurs, indiquant les endroits qui ont été traversés et où des activités d'utilisation mobiles élevées ont été enregistrées. La préfecture la plus visitée est Conakry, qui abrite la capitale de la Guinée. Coyah et Boké sont les deux autres préfectures les plus visitées.

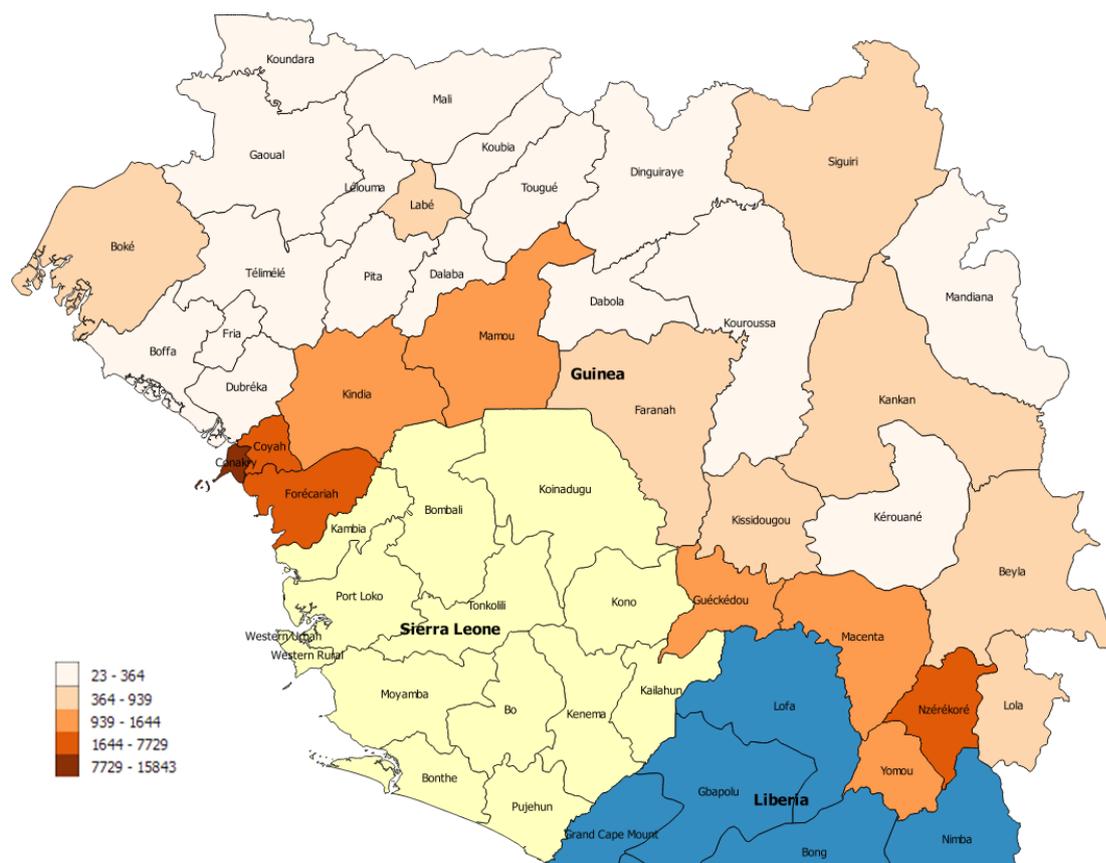
Figure 8.15: Empreinte des utilisateurs de téléphone mobile vers la Guinée



Source: UIT

La Figure 8.16 représente les mouvements transfrontières des utilisateurs de téléphone mobile de la Sierra Leone et du Libéria vers la Guinée. Cette population a été calculée à partir du nombre total d'utilisateurs mobiles uniques qui viennent en Guinée. Elle montre que les préfectures les plus visitées par la population transfrontière entrante en Guinée sont Conakry, Forécariah et Coyah. Forécariah semble par ailleurs être la principale préfecture frontalière permettant de passer de la Sierra Leone en Guinée. Nzérékoré serait le principal district permettant de passer du Libéria en Guinée.

Figure 8.16: Nombre d'utilisateurs mobiles entrant en Guinée

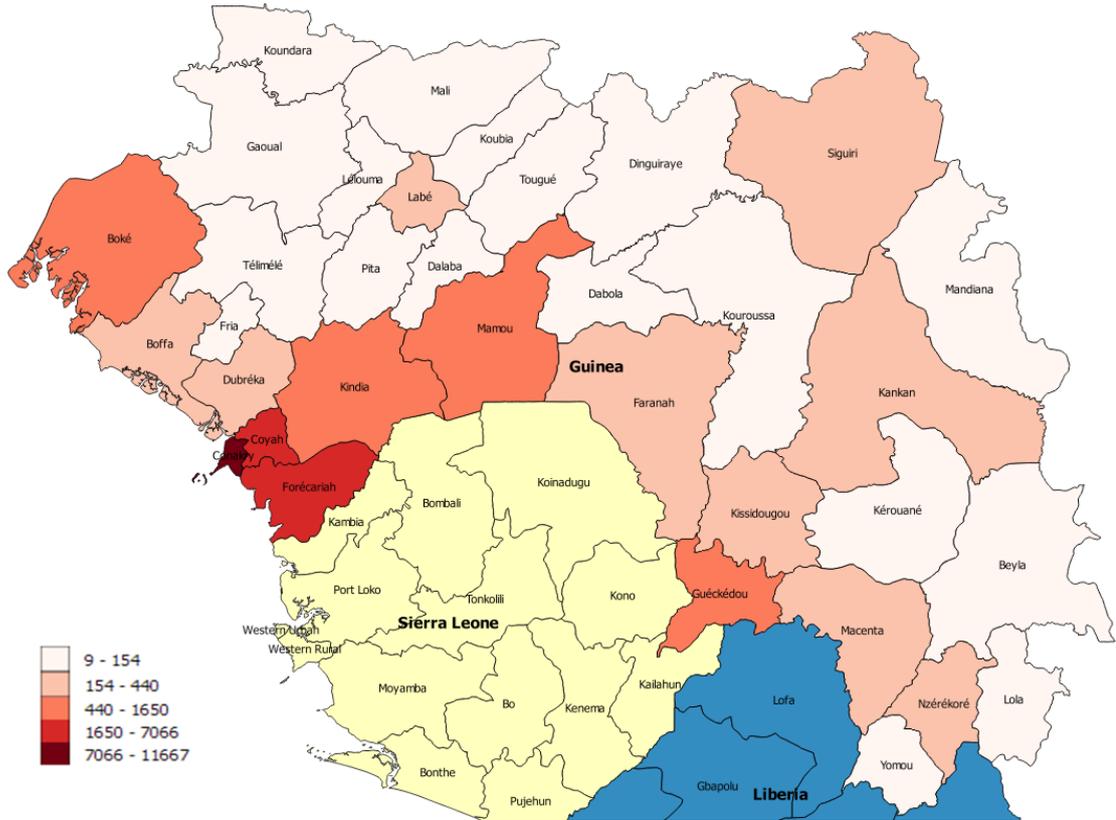


Source: UIT

8.6.4 Mouvements de population transfrontières sortants

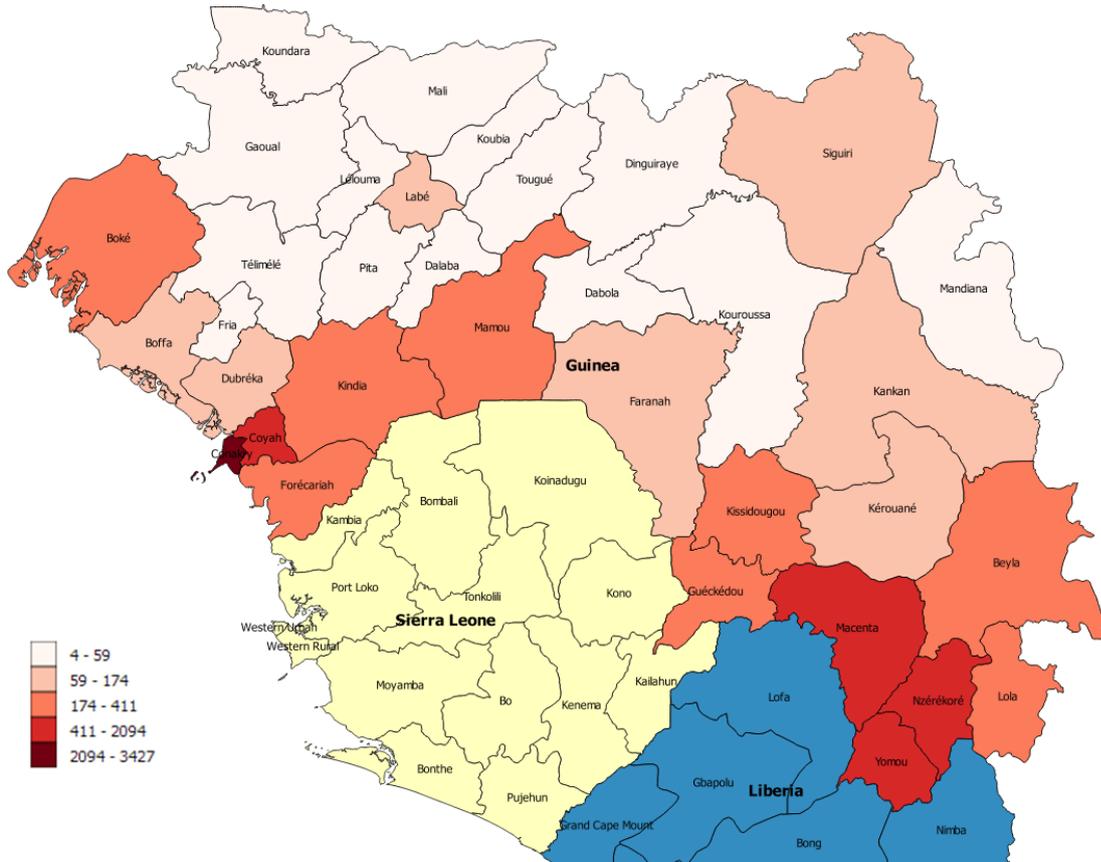
La Figure 8.17 représente les mouvements transfrontières des utilisateurs de téléphone mobile visitant la Sierra Leone depuis la Guinée. Elle montre que Conakry, Forécariah et Coyah sont les trois préfectures les plus visitées.

Figure 8.17: Empreinte des visiteurs vers la Sierra Leone



Source: UIT

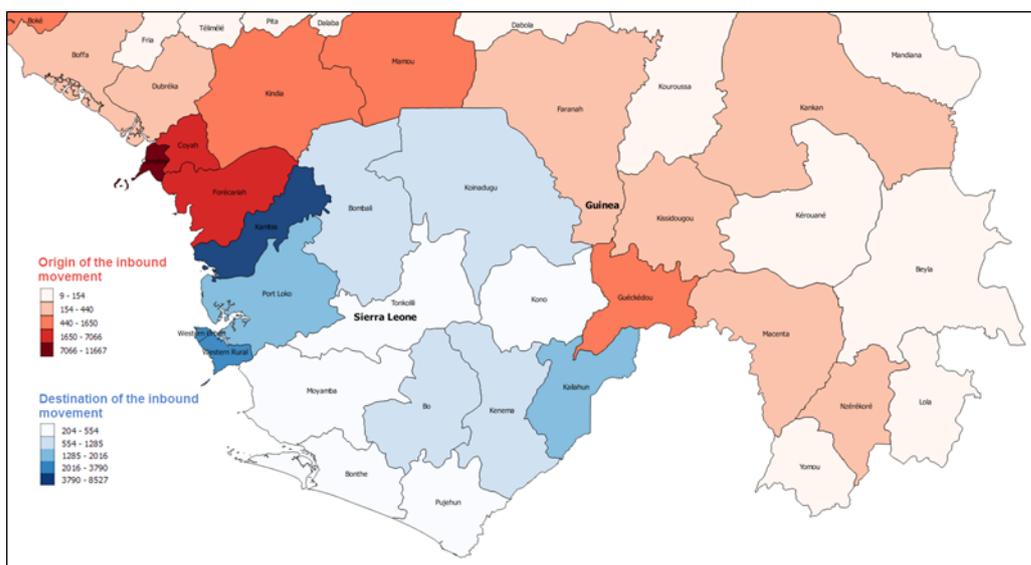
Figure 8.18: Empreinte des visiteurs vers le Libéria



Source: UIT

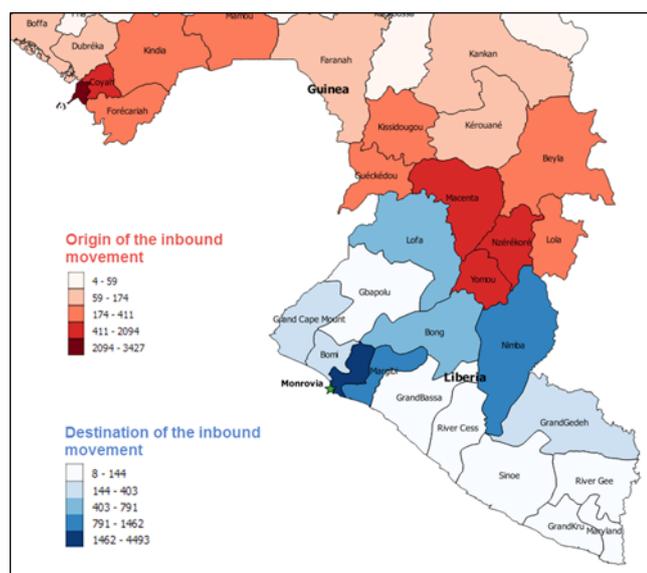
La Figure 8.18 représente les mouvements transfrontières des utilisateurs de téléphone mobile visitant le Libéria depuis la Guinée. Elle montre que Conakry, Forécariah et Coyah sont les trois préfectures les plus visitées avec une population transfrontière importante. Les Figures 8.19 et 8.20 montrent les villes visitées par la population transfrontière en Sierra Leone et au Libéria.

Figure 8.19: Origine et destination de la population transfrontière vers la Sierra Leone



Source: UIT

Figure 8.20: Origine et destination de la population transfrontière vers le Libéria



Source: UIT

9 Discussions

Le manque d'informations précises et horodatées sur les mouvements de personnes lors d'une catastrophe naturelle, d'une situation d'urgence ou en cas d'épidémie peut sérieusement peser sur les performances ou nuire à l'efficacité de la préparation, des interventions et des activités de rétablissement humanitaires. Il est désormais possible cependant d'accéder à ces informations sachant que les téléphones mobiles sont omniprésents, et les données CDR fournissent de précieuses informations sur le comportement des personnes en cas de catastrophe ou d'urgence. Par exemple, les variations du nombre de téléphones actifs connectés à chaque station de base révèlent des modèles d'activité dans des endroits spécifiques pendant et après un événement et peuvent être utilisées pour déterminer les déplacements de population et la connectivité sociale entre les villes et au sein de celles-ci. L'analyse des données CDR constitue un outil puissant pour améliorer les mécanismes d'alerte précoce et de gestion des urgences et révéler d'éventuels schémas de propagation d'épidémie.

Ce rapport décrit comment les données CDR pourraient contribuer spécifiquement à surveiller et à contrôler les épidémies, comme l'Ebola, en estimant les trajectoires humaines et la répartition spatio-temporelle des populations à partir des données CDR au niveau local (rural et urbain) ainsi que dans les zones frontalières.

Les résultats de l'analyse ont révélé la présence d'une forte corrélation entre la répartition des utilisateurs de téléphone mobile extraite des données CDR et les populations réelles extraites des données nationales du recensement; une importante mobilité transfrontière depuis les zones rurales et urbaines dans les pays voisins de la Guinée a également été observée.

L'utilisation des données CDR a montré un énorme potentiel en tant qu'outil pour aider les autorités en cas d'urgence, de catastrophe et d'épidémies. Parce que l'Ebola est une maladie épidémique, il est essentiel de comprendre les mouvements des porteurs potentiels de maladie, non seulement de ville à ville mais également au niveau national, pour que les interventions politiques soient efficaces.

Dans le cadre du présent projet, les données CDR ont été analysées puis compilées pour produire des données statistiques dans une région donnée. Les emplacements géographiques des appareils de téléphonie mobile sont déterminés par l'emplacement des stations d'émission-réception de base du

réseau mobile auxquelles les personnes se connectent lorsqu'elles émettent ou reçoivent un appel. Ceci est très important pour le calcul et l'estimation de la population réelle.

L'analyse des positions quotidiennes de millions d'appareils mobiles collectées sur deux mois a été utilisée pour estimer l'emplacement d'origine de la population d'utilisateurs mobiles en Guinée, ce qui explique également les phénomènes de hausse et de baisse du nombre de personnes au cours d'une même journée à des emplacements spécifiques grâce à une méthode unique de corrélation de l'activité du téléphone mobile avec la population.

La question se pose cependant de savoir comment préserver la vie privée des utilisateurs. Le fait de devoir assurer la confidentialité limite considérablement l'utilisation des données de téléphonie mobile, mais cela est important pour maintenir l'acceptation par le public de l'utilisation des données et aussi parce que ces données risquent d'être utilisées pour compromettre les droits de l'homme, en particulier ceux des personnes vulnérables. Dans la plupart des cas, l'accès aux données impose l'anonymisation des données pour préserver la confidentialité des utilisateurs. C'est pour cette raison qu'un logiciel a été développé. Il peut être fourni aux opérateurs ou aux autorités locales qui, avec une formation minimale, peuvent effectuer la tâche d'anonymisation des données. Malgré cela, l'utilisation de données anonymisées montre des preuves très prometteuses de son efficacité dans les situations d'urgence.

Dans le cadre de ce projet pilote, les populations étudiées ont été limitées aux utilisateurs de téléphone mobile. Pour présenter la répartition des populations réelles, y compris celles qui ne sont pas présentes dans les données CDR, un examen plus approfondi est nécessaire, par exemple pour définir des facteurs d'échelle. Par ailleurs, le projet a porté sur les aspects quantitatifs des utilisateurs de téléphone mobile, tels que le nombre de personnes ou de groupes de personnes et leurs trajectoires. Cependant, sans aspects qualitatifs, comme le genre et l'âge des utilisateurs, il est difficile d'analyser comment les mouvements de population extraits peuvent être liés à d'autres facteurs socio-économiques. L'inclusion d'attributs personnels pourrait grandement accroître la capacité des données CDR à régler certaines questions de société dont le contrôle des maladies épidémiques.

Les données CDR anonymisées ne comprennent pas d'informations personnelles, mais il est possible d'estimer les attributs démographiques de base. Une clé pour estimer les attributs démographiques des données CDR est l'utilisation de données supplémentaires concernant le comportement d'appel et les attributs démographiques. Les données peuvent être collectées auprès des utilisateurs de téléphonie mobile, par exemple au moyen d'une enquête de terrain, et être utilisées comme données de formation et données de validation pour l'estimation. Compte tenu de l'importance que revêtent les informations sur les attributs pour l'utilisation des données CDR à des fins sociétales, il est essentiel de recueillir des données supplémentaires auprès des utilisateurs de téléphonie mobile.

Les trajectoires humaines et la répartition spatio-temporelle de la population extraite des données CDR peuvent servir à analyser l'impact de la mobilité humaine sur la propagation des maladies transmissibles et à estimer le nombre d'évacués en cas de catastrophe ou d'urgence à un moment et à un lieu donnés.

Les cartes de répartition de la population par heures sont utiles lors de l'extraction d'informations statistiques (par exemple, le nombre de personnes au niveau non agrégé plutôt qu'au niveau des unités administratives). Elles peuvent aussi couvrir les populations difficiles à atteindre, sans passer par l'utilisation d'un système d'enregistrement national avancé, compte tenu du fort taux de pénétration du téléphone mobile.

Du point de vue de l'opérateur, toute publication de données comporte à la fois des efforts et des risques ainsi que des contraintes contractuelles lorsqu'il s'agit de publier des données sur les clients ou, tout au plus, un sentiment de perte de confiance des consommateurs si les enregistrements sont transférés en dehors du contrôle de l'entreprise. Les opérateurs peuvent néanmoins publier des données avant ou pendant une urgence, en raison de considérations relatives à la responsabilité

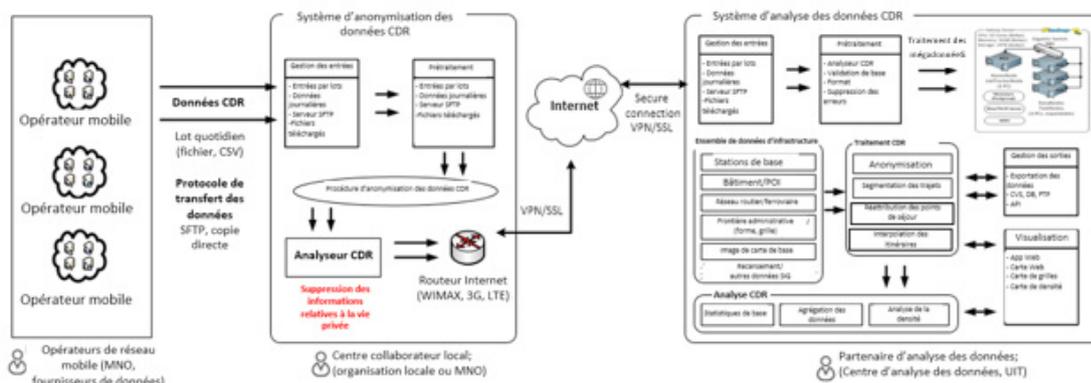
sociale des entreprises. De plus, et parce que les questions internationales ne connaissent pas de frontières, il est important que les gouvernements locaux et les opérateurs des différents pays soient bien organisés pour assurer une utilisation efficace des données.

Le rôle toujours plus important de la technologie mobile dans la communication d'informations pendant les situations d'urgence et les catastrophes (collecte de données et diffusion de données) doit être reconnu pour favoriser l'utilisation des données CDR aux fins de la gestion des catastrophes et des risques d'épidémie.

Enfin, il est recommandé d'élaborer de nouvelles politiques permettant l'utilisation de données CDR pour la sécurité et l'intervention en cas d'urgence de sorte que les données CDR puissent être enregistrées, récupérées, analysées et utilisées pour anticiper et atténuer les effets des futurs foyers épidémiques.

Annexe 1: Concept global de l'analyse automatisée des données CDR

Figure A.1.1: Concept global de l'analyse des données CDR



Source: UIT

Dans le cadre de ce projet, la procédure globale nécessite toujours certaines opérations manuelles par exemple pour la gestion des entrées, le prétraitement et l'analyse des données. Le système automatisé d'analyse de données CDR permet à l'ensemble du système de fonctionner automatiquement avec une moindre participation de l'opérateur. Ce système est recommandé en cas d'extension vers un déploiement à grande échelle.

- Fournisseur de données (opérateur mobile):** Les données sont collectées auprès de l'opérateur de réseau mobile (fournisseur de données). L'opérateur recueille les données de localisation des dispositifs mobiles et les enregistre avant de les exporter ou de les transférer vers un centre de collaborateur local (LCO) pour traitement ultérieur, généralement sous la forme d'un fichier CSV au format compressé (zip). Tout transfert de données entre le fournisseur de données et le LCO doit être effectué via un canal sécurisé (SFTP, VPN) ou par copie directe.
- Centre collaborateur local (LCO):** Le LCO stocke les données CDR anonymisées (fonction généralement assurée par le régulateur ou le fournisseur de licence de téléphonie mobile), sachant que le fournisseur de données peut également fonctionner comme un LCO. Dans ce cas de figure, les données sont anonymisées par le LCO (par exemple l'opérateur) parallèlement au transfert de données vers un système de stockage en nuage (partenaire de stockage de données).
- Partenaire d'analyse des données (DAP):** Le DAP conserve toutes les données CDR assainies (éventuellement plusieurs opérateurs) et veille à la sécurisation de ces données et informations. Cette unité fournit un système de stockage en nuage qui permet à d'autres fournisseurs de données de télécharger leurs données sur le système de stockage en nuage pour une analyse plus approfondie. Elle contrôle la gestion des utilisateurs de même que l'autorisation de télécharger en flux ascendant ou en flux descendant.
- Système d'anonymisation des données CDR:** Ce système gère la procédure d'anonymisation des données CDR. Il reçoit les données CDR brutes au format csv et procède à leur traitement pour supprimer toutes les informations liées à la confidentialité qui seront ensuite téléchargées sur le système de stockage en nuage. Lors de cette phase, l'anonymisation est une application en ligne de commande utilisée pour fournir certains paramètres tels que le chemin d'entrée, le chemin de sortie, les données de départ et l'analyseur de format CDR. Des analyseurs de format CDR spécifiques seront parfois développés sachant que le format de données CDR peut être différent parmi les fournisseurs de données. Le programme comprend également des fonctions de validation de base, de contrôle des erreurs et de suppression.
- Système d'analyse des données CDR:** Ce système effectue une analyse approfondie et conserve toutes les données CDR de tous les fournisseurs de données. Il intègre plusieurs modules dont

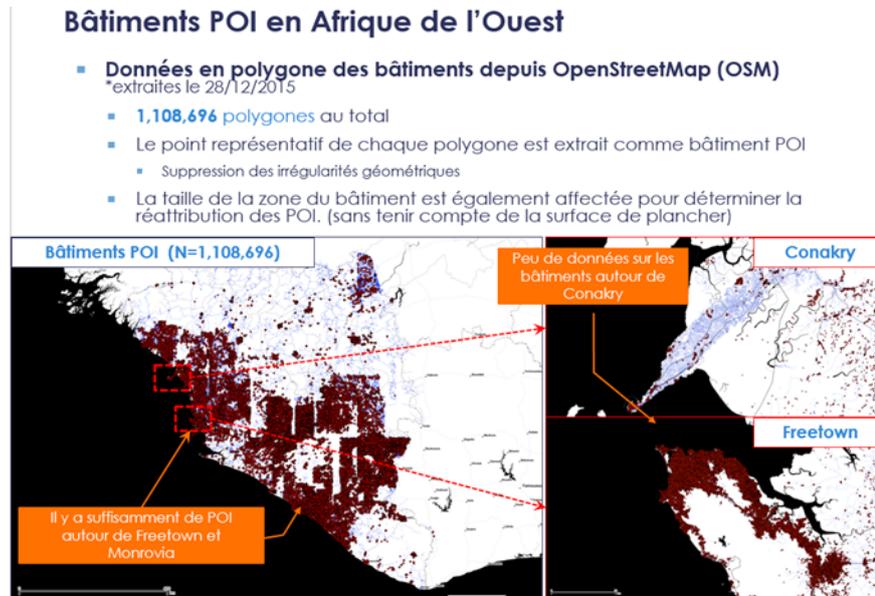
la gestion des entrées, le prétraitement, l'unité de traitement des mégadonnées et le traitement des données CDR. L'estimation de la population est un produit de ce système, mais la livraison du système ne fait pas partie de cette phase.

- **Transfert des données (LCO->DAP):** Des connexions sécurisées (VPN/SSL) devraient être utilisées pour crypter toutes les données transférées. Ces connexions seront établies et entièrement contrôlées par le collaborateur local. Les connexions Internet peuvent se faire via un réseau mobile 3G, un réseau WiMAX ou une connexion par ligne fixe. De telles connexions peuvent aussi être utilisées pour la gestion à distance et le provisionnement du système. Le LCO peut également télécharger des données via une connexion FTP sécurisée (SFTP).

Annexe 2: Bâtiments POI (points d'intérêt) et réseaux routiers

La Figure A.2.1 présente les données POI mises sur pied dans le cadre de ce projet. Celles-ci comprennent 1 108 696 polygones, représentant chacun un bâtiment. La taille de la zone des bâtiments est utilisée pour calculer la zone des bâtiments POI auxquels les emplacements des points de séjour sont affectés.

Figure A.2.1: Bâtiments POI en Afrique de l'Ouest

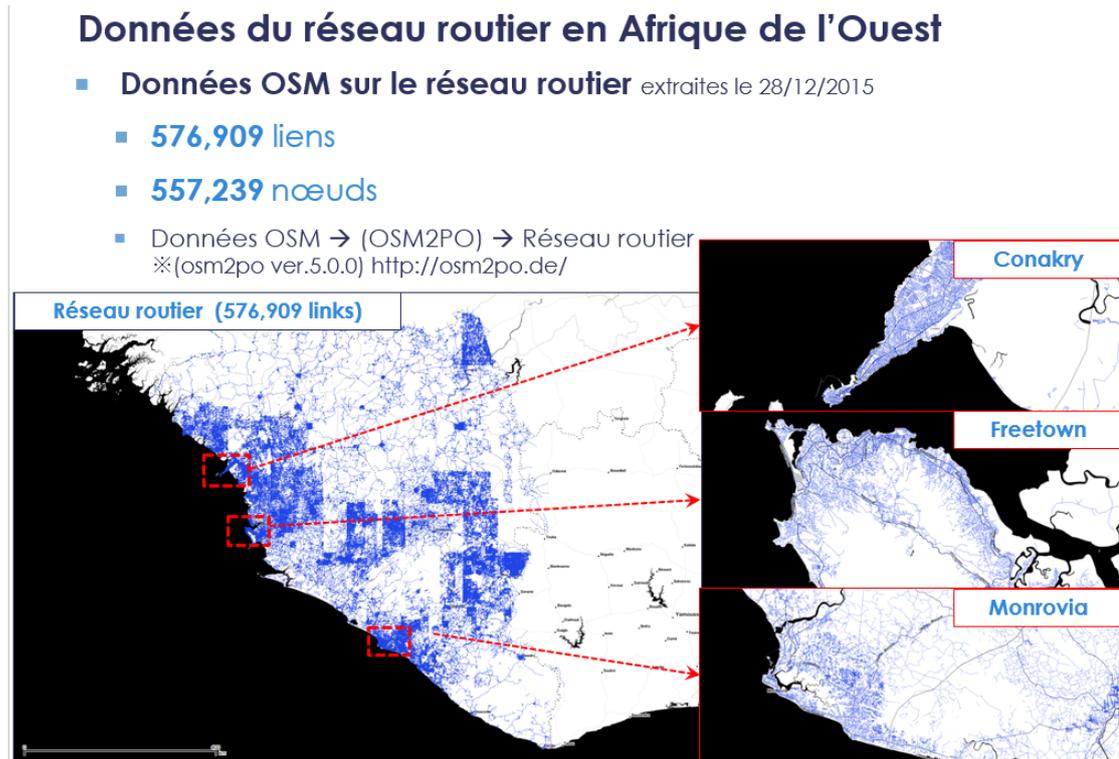


Source: UIT

Les données du réseau routier mises à disposition pour ce projet se fondent sur les données OSM (OpenStreetMap)⁸ du 28 décembre 2015. Elles comprennent 576 909 liens et 557 239 nœuds.

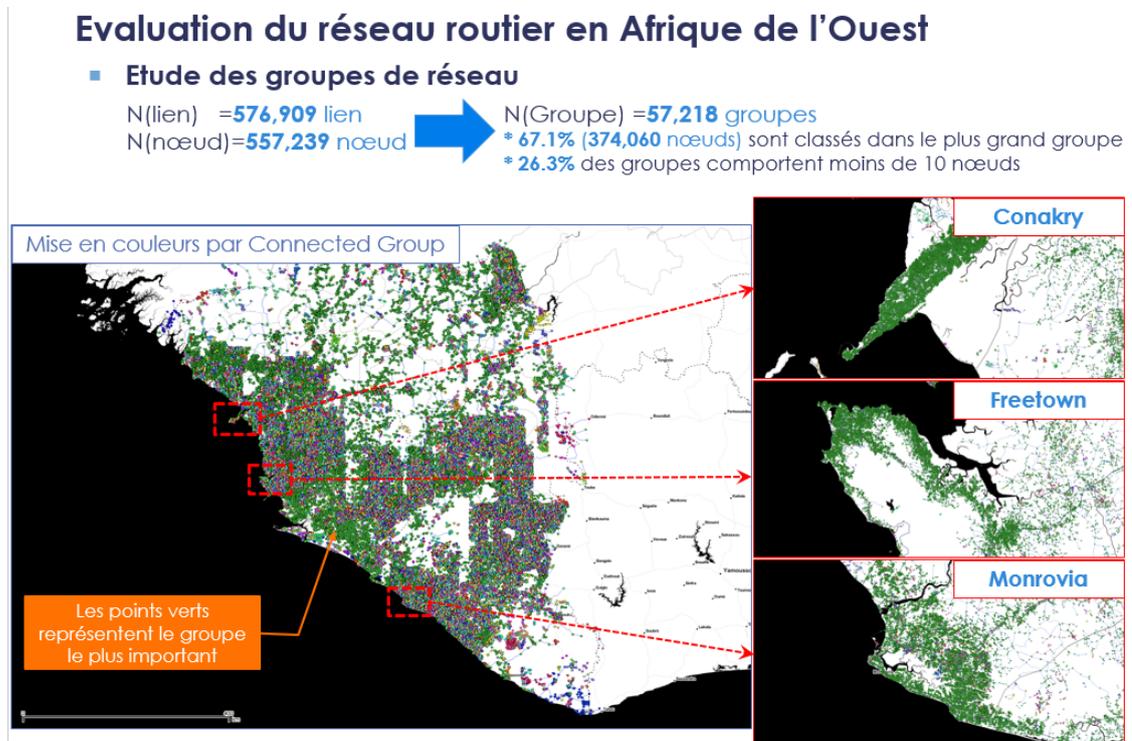
⁸ <https://www.openstreetmap.org/#map=5/51.500/-0.100>

Figure A.2.2: Données du réseau routier en Afrique de l'Ouest



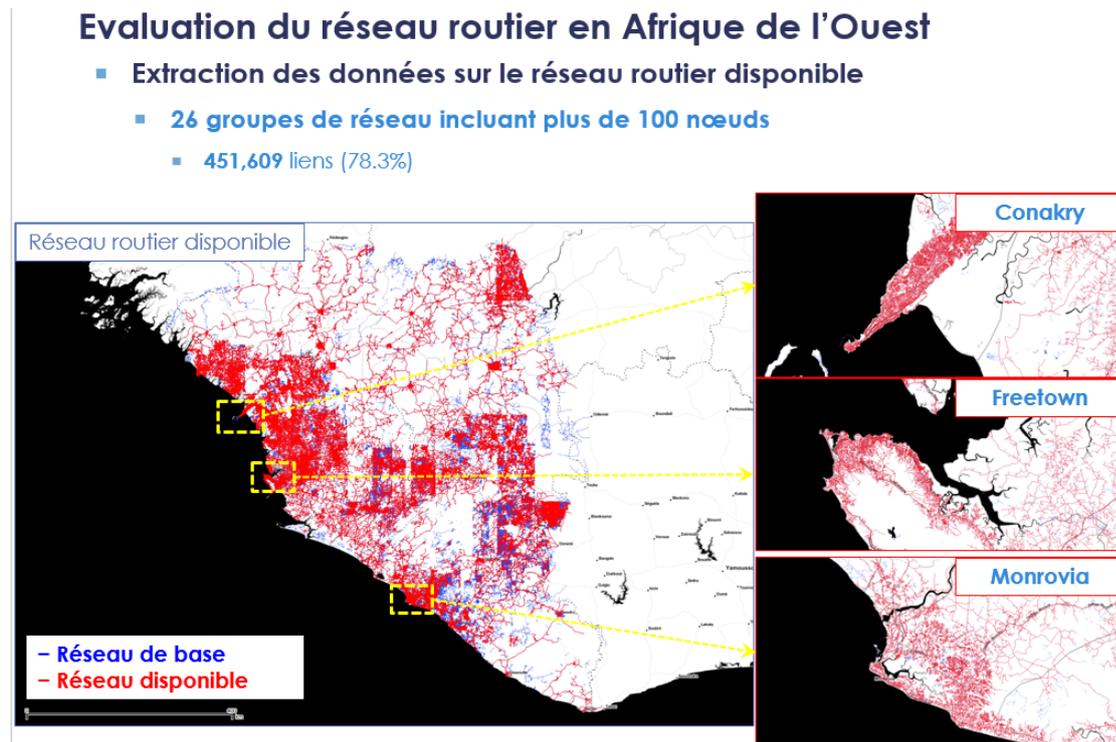
Source: UIT

Figure A.2.3: Evaluation du réseau routier en Afrique de l'Ouest



Source: UIT

Figure A.2.4: Extraction des données du réseau routier en Afrique de l'Ouest



Source: UIT

Union internationale des télécommunications (UIT)
Bureau de développement des télécommunications (BDT)
Bureau du Directeur
Place des Nations
CH-1211 Genève 20 – Suisse
Courriel: bdtdirector@itu.int
Tél.: +41 22 730 5035/5435
Fax: +41 22 730 5484

Adjoint au directeur et
Chef du Département de
l'administration et de la
coordination des opérations (DDR)
Courriel: bdtdputydir@itu.int
Tél.: +41 22 730 5784
Fax: +41 22 730 5484

Département de l'environnement
propice aux infrastructures et
aux cyberapplications (IEE)

Courriel: bdtee@itu.int
Tél.: +41 22 730 5421
Fax: +41 22 730 5484

Département de l'innovation et des
partenariats (IP)

Courriel: bdtip@itu.int
Tél.: +41 22 730 5900
Fax: +41 22 730 5484

Département de l'appui aux projets et
de la gestion des connaissances (PKM)

Courriel: bdtpkm@itu.int
Tél.: +41 22 730 5447
Fax: +41 22 730 5484

Afrique

Ethiopie
International Telecommunication
Union (ITU)
Bureau régional
P.O. Box 60 005
Gambia Rd., Leghar ETC Building
3rd floor
Addis Ababa – Ethiopie

Courriel: itu-addis@itu.int
Tél.: +251 11 551 4977
Tél.: +251 11 551 4855
Tél.: +251 11 551 8328
Fax: +251 11 551 7299

Cameroun
Union internationale des
télécommunications (UIT)
Bureau de zone de l'UIT
Immeuble CAMPOST, 3^e étage
Boulevard du 20 mai
Boîte postale 11017
Yaoundé – Cameroun

Courriel: itu-yaounde@itu.int
Tél.: +237 22 22 9292
Tél.: +237 22 22 9291
Fax: +237 22 22 9297

Sénégal
Union internationale des
télécommunications (UIT)
Bureau de zone de l'UIT
19, Rue Parchappe x Amadou
Assane Ndoye
Immeuble Fayçal, 4^e étage
B.P. 50202 Dakar RP
Dakar – Sénégal

Courriel: itu-dakar@itu.int
Tél.: +221 33 849 7720
Fax: +221 33 822 8013

Zimbabwe
International Telecommunication
Union (ITU)
Bureau de zone
TelOne Centre for Learning
Corner Samora Machel and
Hampton Road
P.O. Box BE 792 Belvedere
Harare – Zimbabwe

Courriel: itu-harare@itu.int
Tél.: +263 4 77 5939
Tél.: +263 4 77 5941
Fax: +263 4 77 1257

Amériques

Brésil
União Internacional de
Telecomunicações (UIT)
Bureau régional
SAUS Quadra 06, Bloco "E"
11^o andar, Ala Sul
Ed. Luis Eduardo Magalhães (Anatel)
70070-940 Brasília, DF – Brazil

Courriel: itubrasilia@itu.int
Tél.: +55 61 2312 2730-1
Tél.: +55 61 2312 2733-5
Fax: +55 61 2312 2738

La Barbade
International Telecommunication
Union (ITU)
Bureau de zone
United Nations House
Marine Gardens
Hastings, Christ Church
P.O. Box 1047
Bridgetown – Barbados

Courriel: itubridgetown@itu.int
Tél.: +1 246 431 0343/4
Fax: +1 246 437 7403

Chili
Unión Internacional de
Telecomunicaciones (UIT)
Oficina de Representación de Área
Merced 753, Piso 4
Casilla 50484 – Plaza de Armas
Santiago de Chile – Chili

Courriel: itusantiago@itu.int
Tél.: +56 2 632 6134/6147
Fax: +56 2 632 6154

Honduras
Unión Internacional de
Telecomunicaciones (UIT)
Oficina de Representación de Área
Colonia Palmira, Avenida Brasil
Ed. COMTELCA/UIT, 4.º piso
P.O. Box 976
Tegucigalpa – Honduras

Courriel: itutegucigalpa@itu.int
Tél.: +504 22 201 074
Fax: +504 22 201 075

Etats arabes

Egypte
International Telecommunication
Union (ITU)
Bureau régional
Smart Village, Building B 147, 3rd floor
Km 28 Cairo – Alexandria Desert Road
Giza Governorate
Cairo – Egypte

Courriel: itucairo@itu.int
Tél.: +202 3537 1777
Fax: +202 3537 1888

Asie-Pacifique

Thaïlande
International Telecommunication
Union (ITU)
Bureau régional
Thailand Post Training
Center, 5th floor,
111 Chaengwattana Road, Laksi
Bangkok 10210 – Thaïlande

Adresse postale:
P.O. Box 178, Laksi Post Office
Laksi, Bangkok 10210 – Thaïlande

Courriel: itubangkok@itu.int
Tél.: +66 2 575 0055
Fax: +66 2 575 3507

Indonésie
International Telecommunication
Union (ITU)
Bureau de zone
Sapta Pesona Building, 13th floor
Jl. Merdan Merdeka Barat No. 17
Jakarta 10001 – Indonésie

Adresse postale:
c/o UNDP – P.O. Box 2338
Jakarta 10001 – Indonésie

Courriel: itujakarta@itu.int
Tél.: +62 21 381 3572
Tél.: +62 21 380 2322
Tél.: +62 21 380 2324
Fax: +62 21 389 05521

Pays de la CEI

Fédération de Russie
International Telecommunication
Union (ITU)
Bureau de zone
4, Building 1
Sergiy Radonezhsky Str.
Moscow 105120
Fédération de Russie

Adresse postale:
P.O. Box 25 – Moscow 105120
Fédération de Russie

Courriel: itumoscow@itu.int
Tél.: +7 495 926 6070
Fax: +7 495 926 6073

Europe

Suisse
Union internationale des
télécommunications (UIT)
Bureau de développement des
télécommunications (BDT)
Unité Europe (EUR)
Place des Nations
CH-1211 Genève 20 – Suisse
Courriel: eurregion@itu.int
Tél.: +41 22 730 5111



Union internationale des télécommunications
Bureau de Développement des Télécommunications
Place des Nations
CH-1211 Genève 20
Suisse
www.itu.int

ISBN: 978-92-61-22292-5



Imprimé en Suisse
Genève, 2017