UNIVERSITÉ DE GENÈVE

Faculté des Sciences de la société (SDS)

Master en Socioéconomie

Année académique 2020-2021

Cours : Sources de Données en Sciences Sociales

Professeur : Pr. Philippe Wanner

Twitter : collecte de données géolocalisées pour des prédictions migratoires

Vestin Cyuzuzo HATEGEKIMANA

TABLE DES MATIERES

Introduction 3
Twitter 3
Prédiction migratoire 4
Type de données 7
Geotagging 7
Qui utilise le *Geotagging* ? 7
Collecte de données 8
API 8
Limites 10
Population 10
Technique 11
Discussion 14
Bibliographie 16

INTRODUCTION

Les prédictions migratoires sont des éléments importants en démographie qui permettent d'adapter les politiques publiques. Les méthodes classiques de prédiction prennent énormément de temps et d'effort pour obtenir des résultats. Plusieurs méthodes altératives ont été mises en place pour pallier ces soucis. Notamment l'utilisation des réseau sociaux tel que Facebook, Twitter ou Linkedin. Naturellement ce type de méthode apporte aussi son lot de défauts.

D'après le rapport de la commission européenne (European Commission. Directorate General for Employment, Social Affairs and Inclusion. et al., 2016) les sources classiques de données pour la migration sont les recensements, les registres de populations, les sources administratives et les statistiques de frontière. Le rapport présente aussi 2 nouveaux moyens originaux de collecter des données sur la migration : les données des téléphones mobiles et les réseaux sociaux. En ce qui concerne les réseau sociaux, l'article se concentre sur la présentation de 6 plateformes : Foursquare, Ficklr, Facebook, Linkedin, Google Latitude et Twitter. C'est sur ce dernier que nous allons consacrer cet article.

Twitter

Le réseau social Twitter a été créé en 2006 et sert de source d'information pour ses consommateurs. Chacun est tenu de se créer un compte avec lequel il va pouvoir exploiter la plateforme. Chaque utilisateur peut en suivre d'autre, les suiveurs d'un compte sont appelés *followers* (abonnés en français) et les personnes suivies d'un utilisateur sont appelés *friends* (abonnement en français). L'affichage du réseau social fonctionne comme un fil d'actualité qui dépend des préférences de l'utilisateur et des personnes qu'il suit. Sur ce fil d'actualité apparaît ce qu'on appelle des *tweets* qui sont des messages limités à 140 caractères écrits par d'autres utilisateurs de Twitter. À ces *tweets* peuvent être ajoutés des liens, des images ou des vidéos (il existe d'autres éléments plus ou moins intéressants), le but étant de transmettre de l'information. Il est également possible de *retweeter* un *tweet*, c'est-à-dire le repartager à ses *followers* avec ou sans commentaires. Entre 2009 et 2019, il était possible pour les utilisateurs d'ajouter un *geotag*[1] à leurs *tweets* (Cebeillac & Rault, 2016; Porter, 2020). Cela signifie

[1] À la suite de ce travail, les terme localisation, géolocalisation, géotagger et d'autres formes d'anglicisme ou de variantes seront employés indifféremment. Les seules précisions qui seront données concernent les localisations s'appliquant à des *tweets*.

qu'ils ont la possibilité d'ajouter une localisation très précise (latitude et longitude) de leur position au moment où ils postent leur *tweet*. Ainsi un utilisateur peut communiquer à ses *followers* où il se trouve au moment où il *tweet*. C'est grâce à ces données de localisation que les prédictions migratoires peuvent être faites.

Les réseaux sociaux tel que Twitter offrent l'avantage de pouvoir croiser beaucoup d'informations, dont des données sociodémographiques qui seront utiles pour étudier les flux migratoires. Il existe également des moyens de pouvoir inférer des informations tel que l'âge et le sexe avec plus ou moins de précision (Sloan et al., 2013; Yildiz et al., 2017). Certaines études ont montré une cohérence entre les données collecté par ce biais et des données officielles d'après le rapport de l'union Européenne (European Commission. Directorate General for Employment, Social Affairs and Inclusion. et al., 2016) mais aussi d'après les quelques études choisis pour cet articles (Hawelka et al., 2014; Jurdak et al., 2015; Zagheni et al., 2014). Naturellement, ces données possèdent des limites que nous présenterons plus tard dans ce document. Mais il est important de noter l'intérêt de ces données : elles peuvent servir de validation de données officielles existantes et elles peuvent servir d'estimation biaisée, mais utilisable pour les prédictions de flux migratoires en attendant des données de meilleure qualité.

Prédiction migratoire

Comme nous l'avons abordé en introduction, l'utilisation des données de géolocalisation de Twitter permet de faire des prédictions migratoires. Puisque les utilisateurs peuvent indiquer leur localisation sur chacun de leurs *tweets*, ils laissent en quelque sorte des traces de leur déplacement au cours du temps. Tout ce qu'il suffit de faire au chercheur, c'est de collecter les *tweets* sur une période défini et d'observer le mouvement des utilisateurs. Dans cette partie nous allons brièvement présenter trois travaux sur la migration réalisé avec les données de Twitter. Ces trois travaux ne sont pas représentatifs de la diversité des recherches dans le domaine, mais ils ont au moins le mérite de présenter les méthodologies utiliser pour valider ou estimer des données sur la migration. Ainsi nous pouvons comprendre comment ce réseau social peut être utilisé pour les prédictions migratoires, comment se passe la collecte puis

l'échantillonnage des *tweets* et nous pouvons avoir un premier aperçu de l'une des API de Twitter (interface de programmation d'application), la Stream API[2].

La première étude *Understanding Human Mobility from Twitter* (Jurdak et al., 2015) utilise un ensemble de données Twitter avec plus de six millions de *tweets* géolocalisés postés en Australie, soit 7 811 004 *tweets* provenant de 156 607 utilisateurs du réseau social de septembre 2013 à avril 2014 afin de déterminer dans quelle mesure les modèles de mobilité basés sur Twitter sont représentatifs des mouvements de la population et des individus. Les auteurs de l'étude avancent que la plateforme peut servir d'excellent proxy pour analyser la migration humaine. Les auteurs n'ignorent tout de même pas le biais d'échantillonnage que représente les données issues du réseau social, la modalité de communication et les possibles de biais de localisation concernant l'envoi des *tweets*. Nous reviendrons sur ces biais dans la partie limites du document. Afin d'appuyer leurs propos les auteurs identifient les schémas de mobilité observés par le biais de Twitter avec les schémas observés par le biais d'autres technologies, telles que les enregistrements de données d'appels. De manière générale, les modèles de mobilité au niveau de la population sont bien représentés par les *tweets* géolocalisés, tandis que les modèles au niveau individuel sont plus sensibles aux facteurs contextuels, tels que le degré de préférence de l'individu pour *tweeter* à partir d'un ou de plusieurs endroits. Les auteurs affirment ainsi que leurs résultats peuvent aider à améliorer la modélisation des mouvements humains, puisque leurs modèles de mobilité humaine extraits de *tweets* géolocalisés présentent des caractéristiques globales similaires à celles observées dans les enregistrements de téléphones portables, ce qui démontre que Twitter est un bon proxy pour étudier la mobilité humaine. Néanmoins, il existe des différences par rapport à d'autres modalités de mesure de mobilité.

La deuxième étude *Inferring International and Internal Migration Patterns from Twitter Data* (Zagheni et al., 2014). Ce document utilise des données géolocalisées pour environ 500 000 utilisateurs de Twitter. Les données proviennent d'utilisateurs des pays de l'OCDE pour la période allant de mai 2011 à avril 2013. Les auteurs évaluent alors pour le sous-échantillon d'utilisateurs qui ont publié régulièrement des *tweets* géolocalisés, les mouvements géographiques à l'intérieur et entre les pays pour des périodes indépendantes de quatre mois. Sachant que les utilisateurs de Twitter ne sont

[2] Plus d'information sur les API de Twitter dans la partie API.

pas représentatifs de la population de l'OCDE, les auteurs ont utilisé la méthode des doubles différences afin de réduire le biais de sélection dans les tendances des taux d'émigration pour un seul pays. Les auteurs affirment que leurs méthodes peuvent être utilisées pour prévoir les points de retournement des tendances migratoires et que les données géolocalisées de Twitter peuvent améliorer considérablement notre compréhension des relations entre les migrations internes et internationales. Le but de l'étude était de montrer la faisabilité de ces méthodes en l'absence de données officielles, contrairement aux 2 autres études présenté dans ce travail qui comparent les données collectées aux données officielles. Toutefois, les auteurs indiquent que les résultats doivent être pris avec précaution.

La dernière étude *Geo-located Twitter as proxy for global mobility patterns* (Hawelka et al., 2014) se base sur près d'un milliard de *tweets* enregistrés en 2012 pour estimer le volume de voyageurs internationaux par pays de résidence. Les auteurs ont collecté une année complète de *tweets* géolocalisés qui ont été postés par des utilisateurs du monde entier entre le 1er janvier 2012 et le 31 décembre 2012. La base de données comprend 944 millions d'enregistrements générés par un total de 13 millions d'utilisateurs. Le flux a été recueilli par l'API de diffusion en continu de Twitter (API de streaming). Les auteurs valident leurs résultats en utilisant les statistiques du tourisme mondial et les modèles de mobilité fournis par d'autres auteurs. Ainsi, ils soutiennent que le réseau social est exceptionnellement utile pour comprendre et quantifier les modèles de mobilité mondiale. Le but de l'étude était de valider la représentativité des *tweets* géolocalisé en tant que source mondiale de données sur la mobilité. Malgré la répartition inégale entre les différentes parties du monde et le biais possible en faveur d'une certaine partie de la population, dans de nombreux cas, les auteurs soutiennent que les *tweets* géo-localisés peuvent et doivent être considérés comme de précieux indicateurs de la mobilité humaine, en particulier au niveau des flux de pays à pays. Dans de nombreux cas, les résultats ont bien confirmé les attentes des auteurs.

Dans les trois articles, les auteurs sont enthousiastes quant à l'utilisation des données géolocalisées de Twitter pour les prédictions migratoires. La première et la troisième étude valident leurs résultats en les comparant à des sources officielles disponibles, tandis que la deuxième étude présente une méthodologie pour obtenir les prédictions les plus fiables possibles. Alors que la première étude présente le cas de l'Australie, les

deux autres ont eu le projet ambitieux d'aborder la migration internationale, entrainant par la même occasion bien plus de biais. Dans tous les cas, les auteurs ne semblent pas ignorer les biais que la collecte de données géolocalisées sur Twitter peuvent apporter, mais ils pensent que les avantages (vitesse, disponibilité et cohérence) peuvent contrebalancer ces défauts et présenter des résultats assez valables de la migration. C'est la vision qui se dégage de la littérature citée par chacun des trois articles.

TYPE DE DONNÉES

Geotagging

Les données de géolocalisation de Twitter sont des données très pratiques puisqu'elles indiquent la longitude et la latitude d'une personne au moment où elle *tweet*. Le *tweet* conserve ces informations, même si la personne change d'emplacement et fait un nouveau *tweet* géolocalisé. L'utilisateur laisse donc une trace de son déplacement derrière lui. Le nombre de *tweets* géolocalisés disponible pour le chercheur est tout de même limité, puisque l'utilisateur doit d'abord activer sa localisation puis indiquer volontairement sa position sur un *tweet*, ces données demandent donc le consentement total des utilisateurs. Nous parlerons des problèmes que cela peut générer plus tard dans la partie « limites ». Ce qu'il faut retenir c'est que bien que l'information soit précise, elle n'est disponible qu'à la volonté de l'utilisateur. Le *geotagging* est l'action spécifique d'indiquer la localisation sur un *tweet* alors que l'activation de la localisation permet simplement en temps réel de voir où la personne se trouve. La dernière solution n'est pas exploitable puisqu'elle ne laisse pas de trace au cours du temps.

Qui utilise le *Geotagging* ?

Puisque l'utilisation du *geotagging* n'est pas automatique et dépend entièrement de la volonté de l'utilisateur, il ne serait pas étonnant alors de constater des différences d'utilisation d'un groupe d'individu à un autre. Ce sont ces différences que l'étude *Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter* (Sloan & Morgan, 2015) met en évidence. Les auteurs expliquent qu'il y a des différences entre ceux qui activent leur localisation ou non et ceux qui utilisent le *geotagging* ou non (une fois que la localisation a été activée). Le but de l'étude est de voir comment le choix de l'un ou l'autre de ces comportements est associé au sexe, à l'âge, à la classe sociale, à la

langue dans laquelle les *tweets* sont écrits et à la langue dans laquelle les utilisateurs interagissent avec l'interface utilisateur de Twitter. Afin de d'obtenir les informations démographiques des personnes, ils ont utilisé des algorithmes de classification. Les auteurs constatent qu'il existe des différences statistiquement significatives entre les deux comportements pour toutes les caractéristiques démographiques citées, bien que l'ampleur de l'association diffère considérablement selon les facteurs. Malgré les limites de leurs données, ils suggèrent que les utilisateurs de Twitter qui publient des informations géographiques ne sont pas représentatifs de la population du réseau social au sens large. Nous reviendrons sur ce constat dans la partie limite. Néanmoins, ils obtiennent des différences (taille d'effets) faibles pour le sexe et l'âge ou peu exploitable pour les classes sociales. Pour toutes ces raisons, bien que les résultats soient significatifs, ils n'ont pas l'aide d'indiquer une réelle différence dans l'activation de la localisation et dans l'utilisation de tweets géolocalisé. Il est donc apriori pas probable que ces différences produisent un biais supplémentaire dans l'échantillonnage. Par contre la langue des utilisateurs et la langue d'utilisation des plateformes indique qu'il y a bien des différences importantes.

COLLECTE DE DONNÉES

API

API signifie *Application programming interface*, soit interface de programmation d'application (Bucher, 2013; Wu et al., 2020). En tant qu'objets protocolaires, les API permettent aux parties intéressées d'accéder aux données et aux fonctionnalités des services en ligne populaires, le tout de manière très contrôlée. C'est la partie visible d'un programme qui peut être utilisée pour avoir accès aux fonctionnalités et données du programme de base sans avoir à connaître son fonctionnement pour l'utilisateur. Il existe plusieurs types d'API, mais nous nous intéressons ici aux API web. Ce type est basé sur le web, donc sur l'accès par programmation aux données et aux fonctionnalités via HTTP. Les API sont donc principalement des interfaces conçues par un développeur pour faciliter l'accès contrôlé à des fonctionnalités et aux données contenues par un service ou un programme. Elles sont surtout proposées par des médias sociaux pour permettre à des développeurs tiers de créer des logiciels ou applications annexes fonctionnant avec le média sans avoir à connaître le code derrière. Ainsi un il serait possible à un développeur indépendant d'utiliser des

données de Twitter par exemple pour créer une application se basant dessus (ex. bots, alarmes, analyseur, etc.).

L'accès à l'API n'est pas libre et demande aux développeurs de respecter un certain nombre de critères. Les chercheurs sont soumis aux mêmes règles pour avoir accès aux données du réseau. Ils doivent premièrement posséder un compte Twitter. Le compte ne doit pas nécessairement être actif mais il constitue la base à l'utilisation des API (les actions passent par ce compte. L'utilisateur souhaitant exploiter l'interface devra alors activer son compte développeur[3]. Une fois que c'est fait, il devra demander la permission à la plateforme pour créer une application. Dans cette demande, il devra indiquer en anglais l'utilisation qu'il compte faire de l'API et répondre à diverses questions. Une fois la requête faite s'il n'y a aucuns problèmes, la personne reçoit dans les 3 jours une autorisation avec différents codes d'authentifications pour son application et c'est à partir de là que les API peuvent être utilisées.

La plateforme possède deux types d'API qui permettent aux utilisateurs de gérer leurs comptes, mais aussi de collecter des données dans les archives et en temps réel : la REST API et l'API de streaming dont la documentation est en libre accès en ligne (Bucher, 2013; Twitter, s. d.). La première permet d'accéder aux utilisations de base de Twitter : poster des *tweets*, analyser les *followers* et les *friends*, rechercher des *tweets*, envoyer et récupérer des messages. La seconde permet de collecter des *tweets* en temps réel sans limitation. Il existe deux versions d'utilisation. La version développeur, qui représente la version de base et la version premium, qui représente la version payante qui a beaucoup moins de limite que la version développeur.

Pour pouvoir exploiter une API, il faut utiliser l'interface mise à disposition par les développeurs de l'application. Mais la tâches peut s'avérer compliquée dans le cas de de notre réseau social à cause de la quantité de travail à faire pour obtenir des données propres et utilisables. Par exemple, pour interagir avec les API de Twitter, les utilisateurs doivent, en plus d'identifier et de digérer les informations pertinentes de la documentation des développeurs de la plateforme, construire/envoyer/recevoir des requêtes, gérer les limites de taux et se battre pour que les objets de réponse imbriqués et en temps réel soient intégrés dans des structures de données faciles à analyser. Les données n'étant pas directement formatées et la quantité de travail pouvant être

[3] La procédure se fait sur le lien suivant : https://developer.twitter.com/en (consulté le 08.11.2020)

importante, il n'est pas rare que les chercheurs utilisent des intermédiaires d'autre langage de programmation (Java, Python, R, etc.) tel que le package *retweet* du langage de programmation R pour pouvoir collecter les données et leur éviter du travail supplémentaire (Kearney, 2019).

LIMITES

Bien que les données soient accessibles au public et qu'elles concernent une large population d'utilisateurs, elles ont plusieurs limites importantes qui d'une part les rendent moins fiables et d'autre part demandent plus d'effort pour les rendre utilisables. Une partie de ces limites a déjà été mentionnée par les auteurs évoqués plus tôt (Hawelka et al., 2014; Jurdak et al., 2015; Zagheni et al., 2014) et sera approfondie dans cette partie. Nous avons regroupé les limites sous deux catégories : les limites liées à la population qui provoquent des biais d'échantillonnages et les limites liées à l'utilisation de l'API menant à des restrictions techniques.

Population

Limite 1 : Représentativité de la population

Sans grandes surprises, la population de Twitter n'est pas représentative de la population générale. Une étude se basant sur le cas britannique par exemple a montré que la population anglaise du réseau social était en moyenne plus jeune et plus instruite que la population réelle (Mellon & Prosser, 2017). Il existe bien des méthodes pour corriger la représentativité de la population (Hino & Fahey, 2019), mais cela ne suffis pas à éliminer le biais de sélection de base. Il faut avoir accès à internet, avoir une certaine connaissance des réseaux sociaux et savoir lire et écrire ce qui limite déjà l'utilisation de Twitter a beaucoup de monde. D'un pays à un autre les biais peuvent être approfondis. C'est déjà un problème important lorsque le sujet d'étude est la migration puisque les facteurs démographiques affectés jouent un rôle important sur la mobilité des personnes.

Limite 2 : Utilisation de la géolocalisation

Au-delà de la représentativité de la population de Twitter avec la population réelle, il existe aussi le même problème entre les utilisateurs du réseau social classique et ceux qui postent leur localisation dans leurs *tweets*. Sachant que suivant certaines information démographique l'utilisation de la géolocalisation change et sachant que

sur le réseau seulement 41.6% activent leur localisation et que dans cet échantillon seulement 3.1% utilisent la localisation dans leur *tweet*, nous pouvons estimer que cela apporte de nouveau biais dans l'échantillon qui peuvent également poser problème dans le recueil de données (Sloan & Morgan, 2015)[4]. Par exemple, dans les études citées plus tôt, avant de pouvoir utiliser les données, il était nécessaire de refaire plusieurs échantillonnages et/ou de collecter de nouvelles données pour obtenir un échantillon intéressant pour une analyse (Hawelka et al., 2014; Jurdak et al., 2015; Zagheni et al., 2014).

Limites 3 : Biais de communication

Le biais de communication ne doit pas être écarté. Twitter est un réseau social qui sert avant tout à communiquer. La localisation est un outil comme un autre qui accompagne le message (ex. image, vidéo, lien etc.). Lorsque les utilisateurs n'ont rien à dire, ils deviennent en quelque sorte invisibles. De plus il n'y a pas réellement de constance dans le flux de *tweets*, ce qui pourrait aussi amené son lot de problème. Rappelons également que Twitter limite le nombre de caractères par *tweets* à 140. Il est important de comprendre si cette limite stricte du contenu des *tweet* peut avoir un impact sur le partage de la géolocalisation. Dans quel type de contenu une personne est plus susceptible d'indiquer sa position ? Par exemple, on aurait plus de facilité à partagé sa position lorsqu'elle est accompagnée d'une photo d'un lieu touristique que lorsque l'on parle d'autres sujets. Enfin, il n'est pas clair si les utilisateurs de Twitter envoient des messages depuis des endroits spécifiques (comme le domicile ou le lieu de travail), s'il y a des préférences de zone pour les envois et comment ces préférences peuvent avoir un impact sur la méthode de collecte de donnée. Ici, ce sont simplement des questions ouvertes qui sont posées, car nous n'avons pas trouvé de travaux mesurant l'effet du problème de communication sur Twitter.

Technique

Limite 4 : Utilisation des APIs

Comme nous l'avons présenté plus tôt, il existe deux types d'API sur Twitter. Une permettant de collecter des *tweets* dans les archives du réseau social (REST API) et l'autre permettant de collecter les *tweets* en continue (API de streaming). C'est la

[4] D'autre travaux cités dans l'étude ont estimé l'utilisation du *geotagging* correspond à 0.85% des *tweets* postés en général.

dernière méthode qui est privilégié par les chercheurs pour des raisons de gratuité des données. Effectivement, il existe des limites concernant la collecte de *tweets* via la REST API qui n'existent pas dans la collecte en temps réelle. La plus importante étant la limitation temporelle : Il n'est pas possible de collecter des *tweets* remontant à plus d'une semaine dans la version gratuite de l'API. Il faut avoir la version premium pour pouvoir avoir accès à toutes les archives de Twitter et encore, la quantité de *tweets* collectables reste limitée. Ainsi les chercheurs sont poussés à utiliser la deuxième méthode gratuite et sans limite. Bien que le service du streaming fixe une limite sur la quantité de données accessibles à moins de 1 % du flux total de Twitter au moment de la collecte, cela reste tout de même un nombre conséquent au vu du trafic journalier du réseau. Le problème étant alors que l'accès à d'ancien *tweets* est limité réduisant ainsi la faisabilité des étude rétrospectives.

Notons tout de même qu'avec les progrès du *web scraping*[5] et la motivation des développeurs Python, un *package*[6] est récemment sorti (octobre 2020) pour ce langage de programmation : TWINT (*How to Scrape Tweets Without Twitter's API Using TWINT*, 2020). Ce package permet de collecter des données de Twitter sans avoir à utiliser l'API de la plateforme. Ainsi, il permet non-seulement d'outrepasser les limites de collecte de *tweets* imposées par la plateforme, mais aussi de faire le travail gratuitement et librement. En effet, il n'est donc plus nécessaire de posséder un compte ou de faire la démarche pour obtenir une autorisation d'utilisation de l'API. Et il n'y a apriori pas de surveillance possible de la part du réseau social. Cet outil est donc prometteur pour les études rétrospectives sur Twitter de manière générale. N'ayant trouvé à l'heure actuelle aucune étude utilisant ce *package* et n'ayant pas encore eu l'opportunité de le tester, il est difficile de pouvoir évaluer ses qualités et ses désavantages.

Limites 5 : Restriction sur les *tweets* géolocalisés

La dernière limite provient de la fin de l'utilisation libre des *tweets* géolocalisés (Hu & Wang, 2020). Dans un *tweet* du 18 juin 2019, Twitter a annoncé le retrait de la géolocalisation des *tweets* :

[5] Le *web scraping* est une méthode de collecte de données en ligne.
[6] Le *package* est une extension d'un logiciel ou d'un langage de programmation ayant sa propre documentation permettant d'exécuter des tâches spécifiques.

Screenshot pris sur Twitter 1[7]

Le retrait de la fonctionnalité a été fait, bien que l'utilisation puisse se faire avec certaines fonctionnalités comme l'indique The verge se référant au premier *tweet*:

> « However, you'll still be able to tag the precise location of your photographs, and TechCrunch notes that you'll also still be able to add your location to *tweets* via the service's integration with mapping services like FourSquare and Yelp. » (Porter, 2020)

Les raisons évoquées à ce retrait sont la faible utilisation de cette option et la volonté de simplification de l'expérience des utilisateurs. Ce changement a pour effet de limiter partiellement la diffusion de position précise des *tweets* puisqu'il n'est plus possible de le faire volontairement à chacun d'entre eux. Seules les personnes ayant l'habitude de publier leur position avec des photos prises par l'application Twitter ou d'autres services extérieurs à la plateforme pourront communiquer leur position. Puisque les débats ont été très animées dans sur le réseau social suite à cette modification Twitter a ajouté des précisions dans un second *tweet* :

Screenshot pris sur Twitter 2[8]

C'est la localisation précise donc le *geotagging* indiquant la longitude et la latitude qui est retirée, mais il est toujours possible d'indiquer sont emplacement dans tous les *tweets*. Cette dernière option permet d'indiquer son emplacement parmi des éléments déjà répertoriés en ligne tel que les restaurant, des lieux publiques ou des monuments.

[7] https://twitter.com/TwitterSupport/status/1141039841993355264?ref_src=twsrc%5Etfw
[8] https://twitter.com/twittersupport/status/1142130343715078144

Elle est donc beaucoup moins précise et bien plus aléatoire. Cette décision a pour effet d'augmenté la protection de la vie privée des personnes, mais a des impactes sur la recherche se basant sur la géolocalisation des *tweets*. Ce sont ces impactes qui nous sont présentés dans l'article *Understanding the removal of precise geotagging in tweets* (Hu & Wang, 2020). L'article explique qu'il reste encore 3 méthodes pour géolocaliser un *tweet* : l'option de géolocalisation générale en attribuant un lieu à un *tweet*, l'option de géolocalisation précise, mais uniquement pour les photos capturées par l'application mobile de Twitter et l'utilisation d'une application tierce connectée à la plateforme (par exemple, Instagram) pour attribuer des coordonnées précises à un *tweet*. Sur les 3 méthodes, la collecte de la longitude et de la latitude des *tweet*s ne peut pas être faites avec la première méthode conformément à la communication de Twitter. À la place, il y juste le nom de la position d'où la personne a *tweeté*. Sur les 2 autres options par contre, il est toujours possible de collecter la localisation précise. Les auteurs de l'étude ont montré en se basant sur deux jeux de données collectés dans une autre étude qu'en réalité la grande majorité des *tweets* géotagués proviennent de sources tierces (de 72% à 88%), alors que seulement une petite partie provient de Twitter (de 8% à 25%). Leur raisonnée est réalisé naturellement en prenant en compte le fait que les jeux de données ne soient pas forcément représentatifs. Donc conformément à l'annonce du réseau social dans son *tweet* du 18 juin 2019, il semblerait que le *geotagging* soit réellement très peu utilisé. Ce qui fait que l'impacte ne sera pas d'une très grande ampleur sur les données collectées. Une question encore ouverte serait de savoir si la population utilisant les *tweets* géolocalisés depuis la plateforme Twitter est différente de celle utilisant des application tierce ; donc est-ce que cette légère perte de données entraine un biais supplémentaire dans la représentativité de la population. Néanmoins les auteurs invitent les chercheurs à prendre en compte dans leurs recherche la vie privée des utilisateurs, mais aussi de diversifier leurs sources de données pour ne pas être trop dépendants de Twitter.

DISCUSSION

Dans cet article, nous avons présenté l'apport des collectes de données géolocalisées sur Twitter dans les prédictions migratoire et nous avons également présenter plusieurs limites de cette approche. Premièrement, cette technique permet d'obtenir rapidement des données à moindre effort et permettrait de confirmer les valeurs officielles. Deuxièmement, elle permettrait aussi d'obtenir un proxy pour plusieurs

valeurs observables et migratoires en l'attente de données plus fiables. Mais plusieurs limites sont imposées à cette méthodologie. Les limites issues de la population, telle que la non-représentativité, l'utilisation ou non de la localisation et les problèmes de communication créent un biais de sélection important. Les limites techniques telles que les restrictions sur l'utilisation de l'API et les restrictions sur l'utilisation des *tweets* géolocalisés limites l'efficacité de la collecte de données. La limite qui peut paraître majeure vient de la mise à jour de Twitter en 2019 entrainant la diminution des *tweets* géolocalisés. Elle n'entraine au final que peu de changement, mais ouvre la voie à plusieurs questions. Notamment des interrogations sur l'accessibilité à des données aussi privées que la localisation, pouvant révéler par exemple des trajets quotidiens. Ces données croisées à d'autres données sensible disponible sur le réseau comme le nom, l'âge ou autre informations privées partagées par les utilisateurs compromettent sérieusement leur sécurité. La restriction sur l'utilisation du *geotagging* semble avoir été une réponse à ce problème, mais en engendre un autre pour les chercheurs. Dans le cas où des restrictions futures amèneraient à la suppression totale de *tweets* géolocalisés, quelles alternatives resterait-il aux chercheurs ?

Après les mesures de Twitter, l'utilisation des *tweets* géolocalisé à des fins de prédiction migratoire n'a pas cessé. Encore aujourd'hui, les méthodes d'analyse de la mobilité basées sur les *tweets* géolocalisés continuent de faire l'objet de recherche dans plusieurs domaines. À titre d'exemple, une récente étude parue en décembre 2020 a utilisé cette méthode pour observer la mobilité des individus durant la pandémie de covid-19 aux États-Unis (Xu et al., 2020). Malgré les défauts évoqués tout au long de ce document nous pouvons constaté que l'engouement pour la collecte de données géolocalisées pour les prédictions migratoires sur Twitter est toujours aussi importante qu'à l'époque des premières études et continuera de croitre à l'avenir.

BIBLIOGRAPHIE

Bucher, T. (2013). Objects of Intense Feeling : The Case of the Twitter API. *Computational Culture*, *3*. http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/

Cebeillac, A., & Rault, Y.-M. (2016). Contribution of geotagged Twitter data in the study of a social group's activity space : The case of the upper middle class in Delhi, India. *Netcom*, *30-3/4*, 231-248. https://doi.org/10.4000/netcom.2529

European Commission. Directorate General for Employment, Social Affairs and Inclusion., University of Washington, Seattle., Wittgenstein Centre, Vienna Institute of Demography., University of Manchester., University of Southampton., Qatar Computing Research Institute, Doha., & Flowminder Foundation, Stockholm. (2016). *Inferring migrations, traditional methods and new approaches based on mobile phone, social media, and other big data : Feasibility study on inferring (labour) mobility and migration in the European Union from big data and social media data.* Publications Office. https://data.europa.eu/doi/10.2767/61617

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260-271. https://doi.org/10.1080/15230406.2014.890072

Hino, A., & Fahey, R. A. (2019). Representing the Twittersphere : Archiving a representative sample of Twitter data under resource constraints. *International Journal of Information Management*, *48*, 175-184. https://doi.org/10.1016/j.ijinfomgt.2019.01.019

How to Scrape Tweets Without Twitter's API Using TWINT. (2020, octobre 15). [Blog]. Medium.Com. https://medium.com/towards-artificial-intelligence/how-to-scrape-tweets-without-twitters-api-using-twint-797b196b951c

Hu, Y., & Wang, R.-Q. (2020). Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour*, *4*(12), 1219-1221. https://doi.org/10.1038/s41562-020-00949-x

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding Human Mobility from Twitter. *PLOS ONE*, *10*(7), e0131469. https://doi.org/10.1371/journal.pone.0131469

Kearney, M. (2019). rtweet : Collecting and analyzing Twitter data. *Journal of Open Source Software*, *4*(42), 1829. https://doi.org/10.21105/joss.01829

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population : Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3), 205316801772000. https://doi.org/10.1177/2053168017720008

Porter, J. (2020, juin 19). *Twitter removes support for precise geotagging because no one uses it, But it will remain available for photographs* [Information]. Theverge.Com. https://www.theverge.com/2019/6/19/18691174/twitter-location-tagging-geotagging-discontinued-removal

Sloan, L., & Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLOS ONE*, *10*(11), e0142209. https://doi.org/10.1371/journal.pone.0142209

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters : Deriving Sociologically Relevant Demographics

from Twitter. *Sociological Research Online*, *18*(3), 74-84. https://doi.org/10.5153/sro.3001

Twitter. (s. d.). *API reference index* [API documentation]. Twitter.Com. Consulté 12 décembre 2020, à l'adresse https://developer.twitter.com/en/docs/api-reference-index

Wu, D., Jing, X., Zhang, H., Kong, X., Xie, Y., & Huang, Z. (2020). Data-driven approach to application programming interface documentation mining : A review. *WIREs Data Mining and Knowledge Discovery*, *10*(5). https://doi.org/10.1002/widm.1369

Xu, P., Dredze, M., & Broniatowski, D. A. (2020). The Twitter Social Mobility Index : Measuring Social Distancing Practices With Geolocated Tweets. *Journal of Medical Internet Research*, *22*(12), e21499. https://doi.org/10.2196/21499

Yildiz, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, *37*, 1477-1514. https://doi.org/10.4054/DemRes.2017.37.46

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from Twitter data. *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, 439-444. https://doi.org/10.1145/2567948.2576930