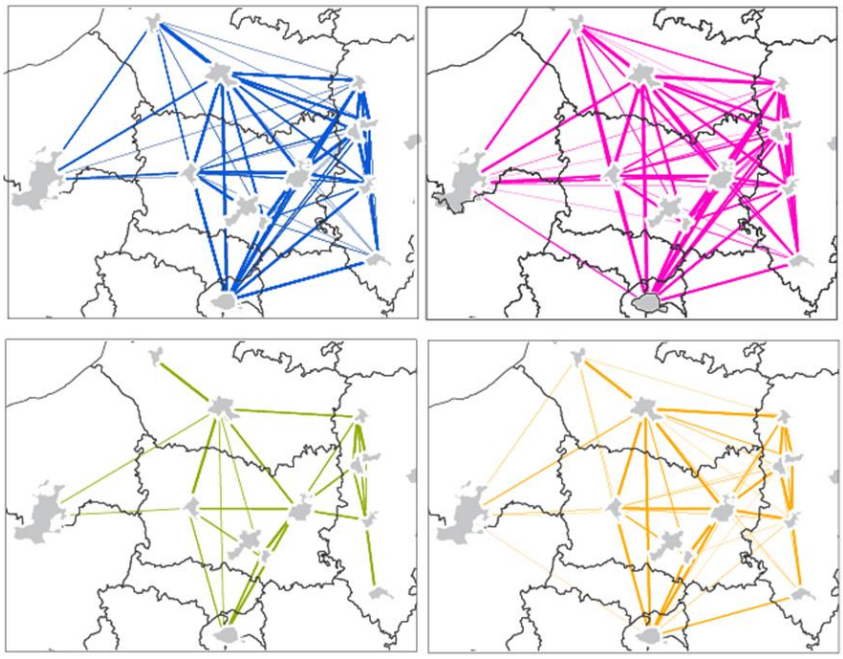
Potentiel des données géolocalisées issues de la foule pour les questions de mobilité et tourisme
- quelques exemples issus de la littérature -

Laurence Jolivet, Arnaud Le Guilcher, Sébastien Mustière, Ana-Maria Raimond, Guillaume Touya
Univ. Paris Est, IGN ENSG, LaSTIG, COGIT



Flux origine/destination issus de données de téléphonie, extrait de [Olteanu-Raimond et al. 2013]

13/01/2017

Préparé dans le cadre du projet PEPS MOBITOURGEO

Objet du document

Le projet Mobitourgéo pose la question du potentiel de diverses données issues de la foule pour aider à analyser et quantifier les pratiques de mobilité des touristes et résidents. Les données issues des opérateurs téléphoniques sont particulièrement ciblées par le projet. Mais, plus largement, d'autres sources d'information peuvent être utilisées : photos partagées sur le Web, tweets, traces de randonnées partagées sur des sites spécialisés, etc. L'objet de ce document est d'illustrer, à travers des exemples de la littérature, quelles études peuvent être faites à partir de ces données en relation avec la problématique de Mobitourgéo, et quelles questions ces données soulèvent.

TABLE DES MATIERES

1 Introduction 4
2 Notions relatives aux données géographiques produites par la foule 5
2.1 VGI, UGC, Big Data… terminologie autour de ces données 5
2.2 Présentation des données spatialisées issues de la foule 8
1.2.1 Les données de téléphonie mobile 8
1.2.2 autres types de données géolocalisées 9
2.3 Quelques défis liés aux données spatialisées issues de la foule 9
2.3.1 Précision de localisation 10
2.3.2 Pérennité 11
2.3.3 Hétérogénéité, complétude, représentativité 12
2.3.4 Manipulation d'un volume important de données 16
2.3.5 Respect de la vie privée 17
2.3.1 Aspects légaux 17
3 Exemples d'analyses des données géolocalisées issues de la foule 18
3.1 Etude des comportements et de l'espace pratiqué 18
3.2 détection et analyse d'événements 24
3.3 estimation des flux origines/destinations 29
3.4 Identification de repères individu-centrés / espace perçu 31
4 Quelques questions méthodologiques ou techniques pour l'analyse des données 32
4.1 Analyse de densité 32
4.2 Recalage et segmentation de trajectoires 34
4.3 Analyse de texte 36
4.4 Analyse visuelle 36
4.5 Techniques d'anonymisation 37
5 Références 38

1 INTRODUCTION

Des données géolocalisées sont utilisées dans l'étude de suivi de personnes, en général dans le contexte de l'utilisation d'un espace particulier (espace urbain, gare, etc.). L'espace peut être lui-même le sujet d'étude ou un contexte des comportements spatiaux individuels, de groupes de personnes et des relations entre ces personnes.

De nombreux travaux de recherche étudient l'intérêt d'utiliser les données spatialisées issues de la foule pour répondre à des questionnements variés, comme par exemple l'évaluation de l'activité des villes (Couronné et al. 2011, Reades et al. 2008; Ahas et al. 2015), l'analyse de la mobilité spatiale (Gonzalez *et al*. 2008, Whang et al. 2012, Schneider et al. 2013, Jiang et al. 2013), l'épidémiologie (Tatem et al. 2009, Wesolowski et al., 2014), la gestion de crise naturelle (Lu et al, 2012), l'aménagement du territoire (Louail et al., 2014). Pour plus de détails sur des méthodes proposés dans le cadre des thématiques liées aux déplacements nous invitons le lecteur à lire notamment Williams et al. (2015).

Dans le domaine du tourisme, certains travaux s'intéressent au tourisme en ville, aux lieux d'intérêt culturels ou à des événements particuliers rassemblant de nombreuses personnes. D'autres travaux s'intéressent au tourisme de nature, de l'observation de paysage, faune ou flore, à la pratique de la randonnée ou d'un autre sport. Des enjeux parmi d'autres en matière de tourisme sont de connaître les fréquentations et les différentes pratiques touristiques. Plusieurs méthodes de suivis existent pour évaluer cela. Le Corre et al. (2012) classent les méthodes en trois familles : quantitatives (ex. comptages), qualitatives (ex. enquêtes) et comportementales (ex. traceur GPS). La téléphonie mobile est mentionnée dans ce dernier article comme encore peu testée. Cependant, l'apport récent des données géolocalisées issues de techniques GPS ou de téléphonie est souligné dans Shoval et al. (2012). Des observatoires, notamment en France, permettent de réunir et de diffuser ces données. Les données peuvent être agrégées dans le calcul d'indicateurs permettant de suivre l'évolution d'un site touristique dans le temps. Certaines données, en particulier les données d'enquête et les données géolocalisées, peuvent aussi être utilisées au niveau de l'individu ou du groupe d'individus afin de caractériser à une grande échelle spatiale et temporelle les comportements.

Dans ce document nous présentons les caractéristiques de données spatialisées issues de la foule et qui peuvent être d'intérêt dans le cadre de l'analyse des mobilités et du tourisme. Nous présentons ensuite des analyses effectuées sur ces données en ne focalisant que sur des recherches aux objectifs proches des objectifs et des données envisagés dans le projet Mobitourgéo (les données agrégés du flux Vision d'Orange voire les données issues de réseaux sociaux).

2 NOTIONS RELATIVES AUX DONNEES GEOGRAPHIQUES PRODUITES PAR LA FOULE

2.1 VGI, UGC, BIG DATA… TERMINOLOGIE AUTOUR DE CES DONNEES

Nous décrivons ici quelques termes utilisés en sciences de l'information pour qualifier les données, en particulier celles produites « par la foule ». Classiquement, les termes utilisés ont des acceptions variables selon les personnes ou communautés, l'ambition de cette partie n'est donc pas de faire une analyse terminologique précise mais juste de fixer quelques éléments de vocabulaire (souvent à travers les acronymes anglais qui sont de fait beaucoup plus répandus que ceux en français). Pour une analyse détaillée des différents termes utilisés dans la littérature nous invitons le lecteur à lire See et al. (2016).

La notion de Web 2.0 désigne plus un paradigme que le Web lui-même. Il désigne la révolution des pratiques de production des informations, dans laquelle tout un chacun participe à la création, la collecte et l'analyse de ces informations. Ce concept marque surtout que la frontière n'est plus franche entre producteur et usager des informations. On parle parfois également de « produser », contraction de « producer » et « user », pour désigner ces personnes (Bruns 2008). De manière emblématique, le Time Magazine a mis ce concept en avant en 2006 en désignant tout un chacun, soit « You », la personnalité de l'année 2016, pour son pouvoir dans la société de l'information (Figure 1).

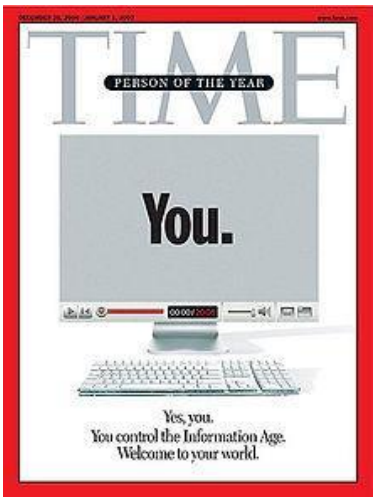


Figure 1. La personnalité de l'année 2006 pour le Time Magazine (la surface grise est un miroir)

Crowdsourcing (ou cloud sourcing, ou action participative/citoyenne) désigne plus particulièrement l'action de réaliser une tâche à laquelle beaucoup de contributeurs participent, en retour d'une satisfaction économique, sociale ou liée à un besoin personnel (Estellés-Arolas et González-Ladrón-de-Guevara, 2012). On désigne par là l'action d'une communauté, souvent connectée, souvent ouverte, souvent volontaire (ou au moins non opposée) et non l'action d'employés ou de fournisseurs d'une entreprise par exemple. On oppose d'ailleurs parfois crowdsourcing et community sourcing. Dans le deuxième cas,

seules des personnes identifiées et autorisées participent à la tâche, alors que dans le premier cas tout un chacun peut le faire.

UGC, pour « User Generated Content », est le terme le plus général désignant des données (spatiales ou non spatiales) produites de manière volontaire ou involontaire par crowdsourcing, en insistant sur le fait que les utilisateurs des données en sont aussi les producteurs (éventuellement partiellement) (Krumm, 2008).

VGI, pour « Volunteered Geographic Information », est un terme introduit par Goodchild (2007) et très répandu en information géographique pour désigner le crowdsourcing volontaire d'informations géolocalisées. Cela englobe clairement les initiatives visant à collecter des données spatialisées pour décrire le territoire de manière général (comme OpenStreetMap -Figure 2-, Google Map Maker ou Wikimapia), ou des initiatives plus ciblées sur des thématiques particulières (comme la diffusion d'informations sur la criminalité dans SpotCrime), ou lors d'évènements particuliers (comme la gestion de catastrophes, la saisie de données suite au tremblement de terre d'Haïti étant un exemple emblématique [Zook et al. 2010]).

Figure 2. Extrait d'OpenStreetMap, le projet emblématique de VGI. Ici des données cartographiques générales, mais aussi des données plus ciblées sur une application (la Loire à Vélo), source [Brando 2013]

Dans l'acronyme VGI, il y a clairement l'idée de collecte par la foule. On oppose donc parfois VGI et données d'autorité (ou données de référence), produite par un producteur institutionnel de données comme l'IGN (Goodchild et Glennon, 2010). Mais la frontière entre les deux s'étiole doucement, une autorité pouvant avoir pour rôle de certifier des données produites par le grand public par exemple, ou des données VGI pouvant servir de référence de fait voire être intégrées dans des données d'autorité.

L'acronyme VGI met bien en avant l'aspect géolocalisé des informations, et est très utilisé. Mais la notion de "volontaire" porte évidemment à débat pour délimiter ce que recoupe le terme VGI. Le "volontariat" pouvant être plus ou moins conscient et ses conséquences plus ou moins maîtrisées. En dehors de la communauté scientifique spécialisée, le terme VGI peut être utilisé par abus de langage pour désigner toutes les données localisées issues de la

foule. Des catégorisations plus fines existent, mais les terminologies associées ne sont pas encore complètement partagées et répandues.

On oppose ainsi parfois VGI et UVGI (pour UnVolonteered Geographic Information). On peut signaler aussi le terme iVGI, introduit par Fisher (2012), pour désigner les données ayant une composante spatiale importante, mais que le contributeur ne partage pas volontairement : c'est clairement le cas des données issues des téléphones portables qui sont nécessairement localisées par l'opérateur mais sans que l'utilisateur en ait nécessairement conscience ni envie de partager cette information de localisation. Les termes Ambiant Geographic Information (Stefanidis et al., 2013) et Social Media Geographic Information (Campagna et al. 2015) sont quant à eux utilisés pour désigner les informations issues de réseaux sociaux, qui sont spatialisées sans que la localisation géographique ne soit l'information essentielle : c'est le cas, par exemple, de certains tweets ou certaines photographies partagées sur le Web si le contributeur a fourni sa localisation, sur la base du volontariat. McKenzie et Janowicz (2014) proposent aussi une classification des données de la foule plus ou moins volontaires, en fonction des critères suivants : accès bidirectionnel ou non aux données, limitation de l'utilisation des données, conscience de la contribution par l'utilisateur, conscience de l'utilisation future des données et implication active de l'utilisateur. A partir de ces critères, on va trouver OpenStreetMap à un bout de l'échelle (très volontaire) et Facebook ou FourSquare de l'autre côté (plutôt coercitif).

Pour toutes ces raisons, pour désigner précisément les données qui nous intéressent dans le cadre du projet Mobitourgéo, à savoir générées par une foule plus ou moins importante, de manière volontaire ou non, avec une forte composante géolocalisée, nous conseillons plutôt l'acronyme UGSC pour « User Generated Spatial Content » (Antoniou 2009, Brando et Bucher 2010), même si il n'est pas très répandu.

Par ailleurs, la notion de crowdsourcing est souvent associée aux notions d'Open Data et Big Data, que nous précisons ci-dessous.

Open Data désigne le fait que des données sont librement réutilisables par tout le monde, sans trop de restriction sur l'usage qui en effet (ou plus précisément avec des licences relativement permissives). Des données sont ainsi mises à disposition par ce biais, pour en promouvoir leur usage. C'est le cas par exemple du taux de remplissage des stations Vélib à Paris. Il faut noter que le terme d'Open Data ne présume rien de la source des données (crowdsourcing ou non), et les données produites par crowdsourcing ne sont pas nécessairement en Open Data, même si cela est fréquent. La publication des données publiques ouvertes, en forte augmentation, est suivie depuis 2013, par l'indice "Global Open Data" (http://index.okfn.org), qui donne l'état des données publiques ouvertes en temps quasi-réel.

Big Data désigne des données tellement volumineuses ou complexes que leur usage est difficile. Au-delà de la notion de volume, clairement désigné par le terme « Big », on dit plus largement que les Big Data sont caractérisées par les « 5 V » (Volume, Variety, Velocity, Variability, Veracity), ce qui illustre que ces données sont caractérisées non pas seulement par leur volume, mais aussi leurs hétérogénéités, leurs confiances variables, leurs contradictions, leurs évolutions… Les données issues de la foule, nombreuses et

hétérogènes, peuvent donc souvent être vues comme du “big data”, mais ce n’est pas toujours le cas.

2.2 PRESENTATION DES DONNEES SPATIALISEES ISSUES DE LA FOULE

Dans cette partie nous présentons brièvement les caractéristiques de données qui peuvent nous intéresser dans le cadre du projet Mobtourgéo.

1.2.1 LES DONNEES DE TELEPHONIE MOBILE

Les données de téléphonie mobiles qui nous intéressent a priori sont les données localisées enregistrées de manière passive par le réseau de téléphonie mobile lorsqu’un événement se produit sur un téléphone mobile. Elles sont propres à chaque opérateur de téléphonie. Plusieurs types d'événements peuvent être distingués : a) l’allumage et l’extinction du téléphone; b) le début de communication (appel entrant ou appel sortant); c) l’envoi ou la réception d’un SMS; d) le *handover*[1]; e) la mise à jour de localisation effectuée par le réseau de téléphonie mobile lorsque le téléphone change de zone de référence LAC[2] ou s’il est inactif plus de trois heures; f) la connexion “Data” à chaque fois qu’une application est utilisée ou fait une mise à jour automatique. En plus des données individuelles, des données agrégées qui représentent le volume d’appel associé à l’antenne est également enregistrée par les opérateurs de téléphonie.

Ces données de téléphones mobiles contiennent d’autres informations que la localisation : l’estampille temporelle (année, mois, jour, minute, seconde), l’identifiant de la carte SIM (anonymisé lorsque les données sont stockées), l’identifiant du LAC et de l’antenne à laquelle le téléphone mobile est connecté et le type d’événement. Smoreda et al. (2014) font une description plus détaillée des données de téléphones mobiles, initialement sauvegardées pour des fins de facturation ou pour assurer la qualité du réseau téléphonique mais qui sont également sources d’informations pour analyser les mobilités.

En fonction de différentes politiques de gestion de données, de politiques commerciales, et de la législation en vigueur de chaque pays vis à vis de la protection de la vie privée, les données de téléphonie accessibles peuvent être de deux types :

- données à l’antenne : événements tous individus confondus liés à une antenne
- données à l’individu : évènements tous lieux confondus liés à un individu particulier.

Dans Gao et al. (2013), par exemple, ces deux types de données sont utilisés afin de déterminer les communautés d’utilisateurs. Ils étudient ainsi l’utilisation de l’espace physique par les données à l’individu (suivi des déplacements) et l’utilisation globale de l’espace. Le service Flux Vision de l’opérateur Orange par exemple s’appuie également sur ces deux types de données.

[1] Handover : un changement automatique de connexion d’antenne, d’une cellule à une autre pendant la communication si un changement de position du terminal le fait sortir de la zone de couverture de la première cellule

[2] Location Area Code (LAC) : des grandes zones regroupant des antennes voisines pour une meilleure gestion technique du réseau de téléphonie mobile.

Les données de téléphonie peuvent utiliser différentes unités de comptage (cf. Tiru et Ahas 2012 pour plus de détail): Erlang (mesure de charge de l'antenne spécifique au domaine de la téléphonie), CDR (call detailed record, comptage des événements de types b et c), A-BIS probe-based (comptage évènements de type a, b, c, d, e), MPS, Anonymous Bulk Location Data.... Notons que Tiru et Ahas (2012) considèrent que la mesure Erlang (charge de l'antenne) est la moins adaptée pour les études statistiques sur le tourisme.

Ahas et al. (2008) précisent que les défauts principaux des données de téléphonie sont la difficulté d'accès aux données ainsi que le manque de précision des localisations, la localisation considérée étant celle le plus souvent celle de l'antenne à laquelle le téléphone est connecté (cf. partie 2.3.1).

1.2.2 AUTRES TYPES DE DONNEES GEOLOCALISEES

Nous pouvons citer ici plusieurs types de données géolocalisées d'intérêt en reprenant la classification de Senaratne et al. (2016) :

- les données de type *textuel* comme Twitter, pour lesquelles la localisation provient d'un GPS associé ou est décrite dans le texte lui-même;
- les données de type *image*, comme FlickR ou Panoramio, dans lesquelles une image est associée à des coordonnées correspondant au sujet de la photo ou à la localisation de l'appareil photo;
- les données de type *enregistrement*, comme Swarm/Foursquare ou Facebook, dans lesquelles un utilisateur se déclare comme présent dans un lieu (lieu touristique, magasin…);
- les *données d'usage* d'équipements localisés, comme par exemple les données d'entrée dans une station de métro de la RATP, les données d'occupation de stations Velib' ou Autolib', ou les données d'occupation d'hébergements d'AirBnB.

La localisation de ces données, quand elle est précise, est souvent issue d'un positionnement par GPS[3]. Les enregistrements peuvent s'effectuer soit via un téléphone portable et des applications installées (Twitter ou applications sportives spécialisées par exemple), soit directement par des appareils GPS dédiés. Dans d'autres cas, une localisation moins précise peut exister via une adresse, ou l'appartenance à une commune, comme par exemple si on observe des sites Web de réservation d'hébergement pour déterminer la fréquentation touristique d'un lieu.

2.3 QUELQUES DEFIS LIES AUX DONNEES SPATIALISEES ISSUES DE LA FOULE

Nous listons ici quelques défis liés à l'usage des données spatialisées générées par la foule. Si la liste des difficultés potentielles peut paraître longue, il s'agit de montrer la prudence

[3] GPS: Global Positioning System, système américain de positionnement par satellite. Terme qui est devenu générique dans le grand public pour désigner tout système de positionnement par satellite, ainsi que les applications de navigation qui s'appuient dessus.

nécessaire lors de l'analyse de ces données. Mais notre présupposé général reste bien que le volume, la précision et la richesse croissante de ces données leur confère un grand potentiel malgré les difficultés listées.

2.3.1 PRECISION DE LOCALISATION

Puisque nous nous intéressons à des données localisées, il faut se poser la question de la précision de cette localisation. Celle-ci est évidemment dépendante de la source des données, parfois inconnue, et la variabilité peut être forte selon les situations considérées. Quelques éléments d'appréciation néanmoins :

- Les GPS grand public (ceux des téléphones) ont maintenant une précision globale de quelques mètres à quelques dizaines de mètres, compatible en général avec l'analyse de mobilité en ville. Par contre, il y a généralement du bruit (des localisations erronées), et on identifie des contextes où le bruit peut être particulièrement fort : à l'intérieur des bâtiments bien sûr (le signal des satellites y est pas ou très mal capté), dans les canyons urbains (Niehoefer et al., 2013) et sous couvert végétal dense (Lewis et al. 2007). Par ailleurs, il faut savoir que l'information d'altitude est beaucoup moins précise que la précision planimétrique.

- Dans des initiatives de cartographie de type OpenStreetMap (OSM), les données, qui sont souvent issues de levés GPS ou de saisie sur photographies (ou importées depuis d'autres sources), ont une précision géométrique de l'ordre de la dizaine de mètres. Nous avons évalué cela sur les données OSM France (Girres et Touya 2010), mais également sur des données saisies en urgence après le tremblement de terre d'Haïti (Document interne IGN 2013).

- Les données de téléphonie fournies par les opérateurs sont par défaut localisées à l'antenne réceptrice, soit une précision de l'ordre de quelques centaines de mètres (voire plus en zone peu dense). NB : si le téléphone est capable de se localiser bien plus précisément, avec son GPS par exemple, cette information n'est par défaut pas transmise aux opérateurs, en tous cas pas sans autorisation explicite de l'utilisateur.

- La localisation associée à une donnée peut être éloignée du sujet de la donnée. Typiquement, les photos sont souvent localisées là où la photo a été prise (où était l'appareil photo) et non sur l'objet visé, même si quelques utilisateurs font l'effort de ce recalage a posteriori dans des applications spécialisées. Par exemple, Zielstra et Hochmair (2013) montrent que les photographies du site Panoramio sont plus souvent recalées sur l'objet photographié que celles du site Flickr. Dans le même ordre d'idée, la localisation du lieu de résidence d'un usager de téléphone mobile peut avoir été estimée à partir de la localisation de son portable, ou alors à partir de son adresse de facturation qui peut être imprécise, erronée, ou éloignée du lieu réel de résidence. De même, il existe plusieurs sortes de localisation de tweets qu'il ne faut pas confondre : la localisation « exact location » correspond à une localisation par GPS faite à la volée par le téléphone utilisé pour diffuser le tweet, mais la localisation « place » correspond à une localisation donnée interactivement (au niveau de la ville) par l'utilisateur et qui peut ne pas correspondre au lieu duquel a été envoyé le tweet, mais par exemple à une localisation par défaut définie par l'utilisateur sur son compte (Leetaru et al. 2013). Les figures ci-dessous montrent la

différence de répartition de tous les tweets sur plusieurs années, selon que l'une ou l'autre localisation est utilisée.

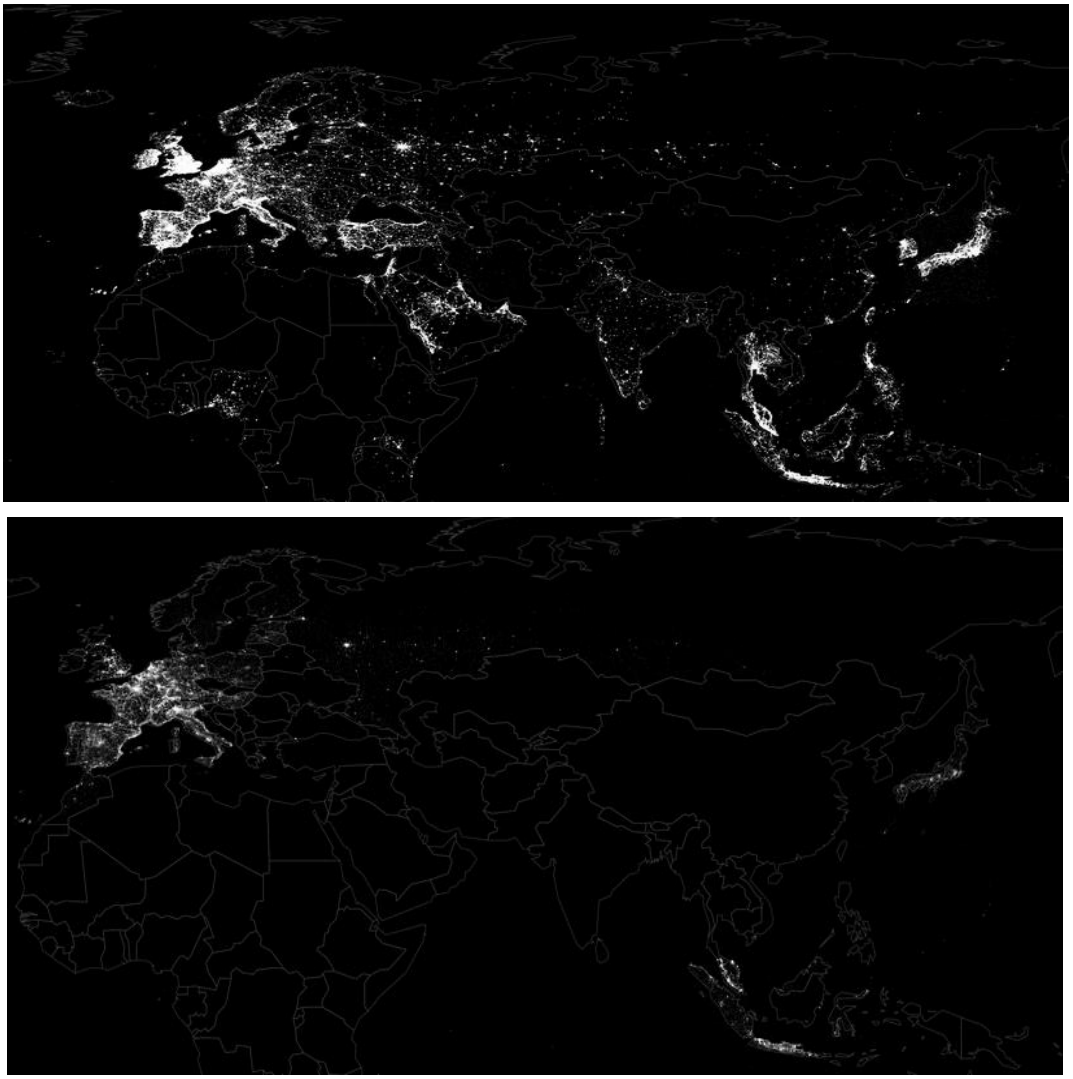


Figure 3. Carte de tweets localisés avec la métadonnée « exact location » (en haut) ou la métadonnée « place » (en bas) [Leetaru et al. 2013].

2.3.2 PERENNITE

De nombreuses initiatives de collecte de données volontaires à travers le développement de plateformes de saisie collaborative de données (OSM, Wikipedia...), d'applications mobiles (FixMyStreet, Sauvages de ma rue...) et de sites internet de partage d'information (RandoGPS, Did You Feel It, Citizen Sky...) ont vu le jour ces dernières années. Certaines connaissent une activité continue, d'autres ont eu un objectif ponctuel lié un événement particulier, tel que le tremblement de terre de Haiti ou le crash d'avion de Malaysia

Airlines, mais d'autres encore ont cessé leurs activités faute de motivation de contributeurs (See et al. 2016). Il a été noté que la motivation des contributeurs, l'esprit de communauté, la satisfaction sociale, la volonté de se rendre utile, ainsi que le retour reçu sont des éléments clés pour assurer la pérennité d'une initiative collaborative (Olteanu-Raimond et al. 2016). Il existe un véritable enjeu à faire perdurer les initiatives de collectes de données au-delà d'effets de mode ou d'effets ponctuels (via une mise en lumière dans un média par exemple). Pour parvenir à cela, des techniques de « gamification » sont parfois mentionnées (motivation par la mise en place de procédés ludiques, tels que des concours entre contributeurs). Par ailleurs, le retour vers les contributeurs, via la mise à disposition de résultats d'analyse, sont essentiel pour la pérennité des contributions.

2.3.3 HETEROGENEITE, COMPLETUDE, REPRESENTATIVITE

Une des principales caractéristiques des données spatialisées générées par la foule est leur l'hétérogénéité spatiale. Typiquement, les données sont en général plus nombreuses en zone urbaine qu'en zone rurale, ne serait-ce que parce que la densité de contributeurs potentiels est directement liée à la densité de population. Cela est attesté par de nombreuses études (Girres et Touya, 2009 ; Neis et Zielstra, 2014; Estima et al, 2014; Ma et al, 2015). Mais de plus, au sein même de zones urbaines denses, on observe une hétérogénéité spatiale, les zones touristiques et dynamiques ayant plus de données que les zones résidentielles ou moins touristiques, là encore parce qu'elles attirent plus de contributeurs (Figure 4 ; Antoniou et Schlieder, 2014; Estima et al, 2014). Si le biais lié à la densité de population peut être maîtrisé, celui lié à l'attractivité des lieux est beaucoup plus difficile à appréhender.

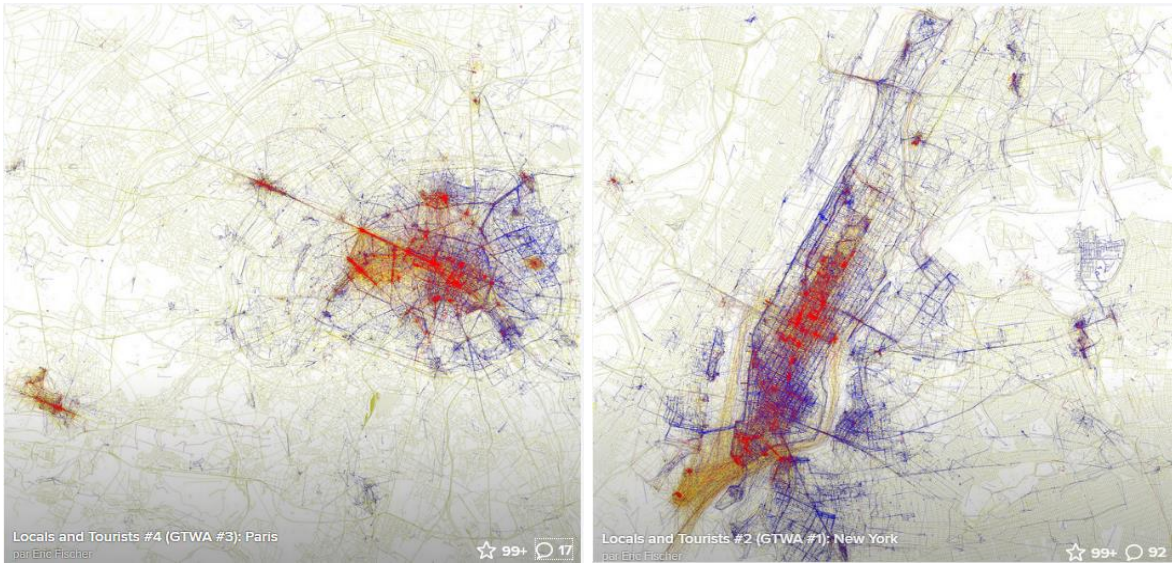


Figure 4. Densité de photos géolocalisées issues de Flickr: Paris à gauche et New York à droite (source : Locals and Tourists, Eric Fischer). La densité est corrélée à l'attractivité des lieux et non à la densité d'habitants.

De plus, si on se concentre sur les données qui ne sont pas nativement géographiques, comme Twitter dans lequel il faut volontairement activer le GPS pour faire des tweets localisés, le problème de la couverture des données est fort (seuls quelques pourcents des tweets sont géolocalisés, cf. ci-dessous). Si la période de temps étudiée est courte et la zone considérée est peu étendue, les données peuvent être peu nombreuses et le risque de non représentativité de l'échantillon considéré est élevé. Par contre, le potentiel en termes de volume de données est tel que les effets de masse permettent de gommer certains biais non systématiques pour l'analyse. Pour que cela se vérifie, il faut donc analyser des volumes importants de données, ce qui implique la nécessité de mettre en place des outils d'analyse automatique, même si leur finesse peut être limitée. La plupart des outils d'analyse de ces données privilégient donc l'automatique à la précision, en postulant que les erreurs d'analyse seront masquées par la quantité de données.

Quelle que soit la source utilisée, les études des données générées par la foule sont évidemment impactées par le profil de la population active pour produire ces données. La plus grande prudence est donc de mise, surtout si on souhaite estimer des populations absolues à partir de ces données, le redressement des populations estimées étant d'une difficulté forte.

Nous citons ci-dessous quelques éléments pour juger de la représentativité des tweets, données caractéristiques des données produites par la foule. On estime (chiffres en constante évolution et parfois variables selon les estimations) :

- 300 millions d'utilisateurs inscrits sur Twitter, dont 100 millions actifs (chiffres 2015) ;
- avec une forte sur-représentativité des pays anglo-saxons (50% des utilisateurs issus des Etats-Unis en 2012), et des jeunes (73% des utilisateurs ont entre 15 et 25 ans en 2012) ;
- et des utilisateurs diversement actifs (1% des utilisateurs envoient 20% des tweets ; 15% des utilisateurs envoient 85% des tweets).

Si on s'intéresse spécifiquement aux tweets géolocalisés, on constate :

- avoir a disposition relativement peu de tweets : entre 1 et 3% des tweets sont géolocalisés (chiffres variables selon les lieux, périodes et façons de compter) [Leetaru 2013 ; Fuchs et al. 2013] ; soit quand même quelques millions par jour sur la Terre entière ;
- qu'1% des utilisateurs de Twitter produisent 66% des tweets géolocalisés [Leetaru et al. 2013] ;
- que le nombre de tweets est évidemment corrélé à la densité population des lieux concernés ;
- et des attitudes diverses face à la géolocalisation, soit en raison de diversités culturelles sur les questions liées à la préservation de la vie privée par exemple, soit plus pragmatiquement parfois pour des raisons techniques liées par exemple aux applications phares pour tweeter selon les pays. Quelques exemples dans les Figure 5,

Figure 6 et Figure 7 pour s'en convaincre. Quelle que soit la raison de ces biais, ils compliquent la tâche de quantification des populations étudiées.

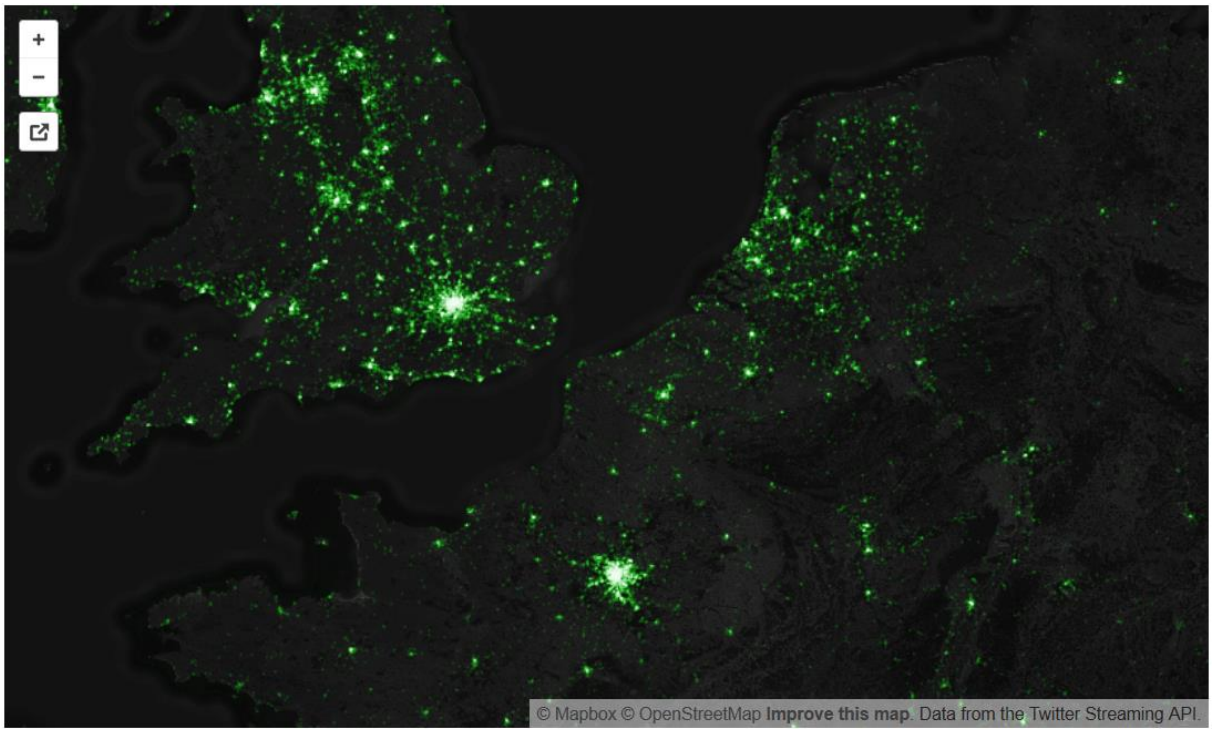


Figure 5. Carte des tweets géolocalisés [Fischer 2013] : la frontière Allemagne / Pays-Bas est bien visible : est-ce dû à des raisons techniques ou à attitudes culturelles différentes dans ces pays vis à vis de la préservation de la vie privée ?

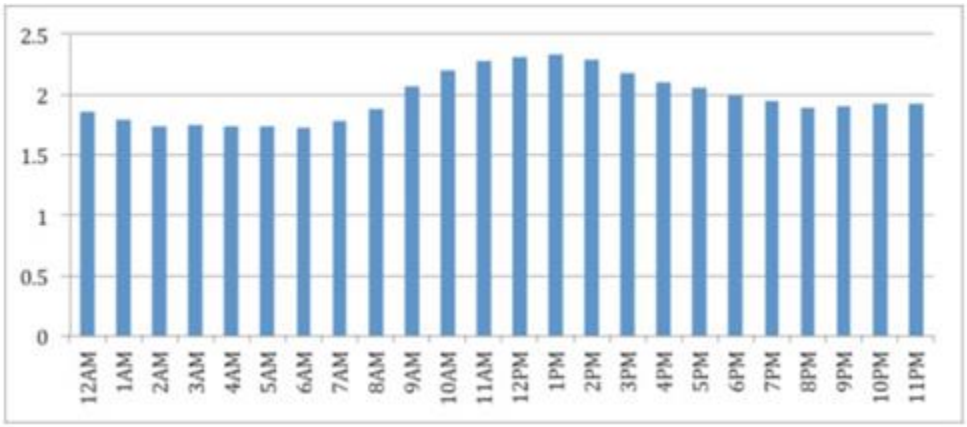


Figure 6. Pourcentage de tweets géolocalisés selon l'heure de la journée (Pacific Standard Time = côte ouest américaine). Selon quelle partie du monde est réveillée, une plus ou moins grande proportion de tweets sont géolocalisés (Leetaru et al. 2013) : la marque de cultures différentes ?

Table 2: Percent georeferenced tweets by language (Twitter Decahose 23 October 2012 to 30 November 2012).			
	Percentage georeferenced tweets	Percentage all tweets	Percentage language georeferenced
English	41.57	38.25	2.17
Spanish	11.16	11.37	1.96
Portuguese	9.50	5.58	3.40
Other	8.39	0.51	32.78
Indonesian	7.33	8.84	1.66
Turkish	3.87	1.80	4.29
French	3.85	2.30	3.35
Arabic	2.81	4.09	1.37
Russian	2.24	1.12	3.98
Italian	1.95	1.31	2.97
Japanese	1.63	11.84	0.27
Dutch	1.40	1.51	1.85
Norwegian	0.76	7.74	0.20
German	0.75	0.66	2.25
Swedish	0.48	0.27	3.63
Thai	0.46	0.48	1.92
Finnish	0.44	0.34	2.62
Polish	0.40	0.34	2.34
Korean	0.36	1.17	0.62
Czech	0.13	0.11	2.35
Danish	0.13	0.09	2.90
Greek	0.11	0.07	3.20
Chinese	0.11	0.09	2.53
Ukrainian	0.09	0.04	4.14
Vietnamese	0.03	0.04	1.80
Persian	0.02	0.03	1.28
Hebrew	0.01	0.01	2.49

Figure 7. Le pourcentage de tweets (dernière colonne) géolocalisés selon la langue (estimée) des tweets ; on notera en particulier que 2,17% des tweets en anglais sont géolocalisés, alors que seulement 0,17% des tweets en japonais le sont. Tableau issu de (Leetaru et al. 2013).

D'autres études soulignent le fait que les réseaux sociaux sont utilisés davantage par des populations jeunes. Par exemple, Lenhart (2009) indique qu'aux États-Unis les trois quarts des personnes entre 18 et 25 ans ont un profil dans un réseau social internet. Ce pourcentage diminue avec les classes d'âges, avec moins de 10 % des personnes ayant plus de 65 ans. Cela peut donc entraîner des biais dans les analyses de données issues de médias sociaux. Par exemple, Siła-Nowicka et al. (2016) s'intéressent aux déplacements individuels. Ils précisent ne pas avoir pris en compte des données issues de média sociaux comme Twitter et Foursquare afin d'éviter un biais par sur-représentativité de populations jeunes. Leur protocole intègre donc un enregistrement des localisations par des traceurs GPS auprès de volontaires. Cependant, cela a entraîné un autre biais car les personnes âgées ont été alors surreprésentées.

De même que pour les tweets, on retrouve des hétérogénéités des populations contributrices pour les données enregistrées par les GPS des téléphones portables. Bergman et Oksanen (2016) étudient les traces GPS de cyclistes enregistrées via l'application Strava (Figure 8). Ils constatent que 50 % des traces sont enregistrées par moins de 5% des contributeurs. Les trajets sont classés en deux types dans leur étude : ceux

qui relient deux lieux distincts et ceux en boucle. Les trajets entre deux points distincts sont plus nombreux et surtout majoritaires en début de matinée. Ces caractéristiques de traces GPS peuvent ensuite être reliées à des types d'usagers. Les cyclistes effectuant des trajets en boucle sont probablement surtout composés de cyclistes de loisirs plutôt que de cyclistes utilitaires, mais les informations liées à ces deux populations sont mélangées.

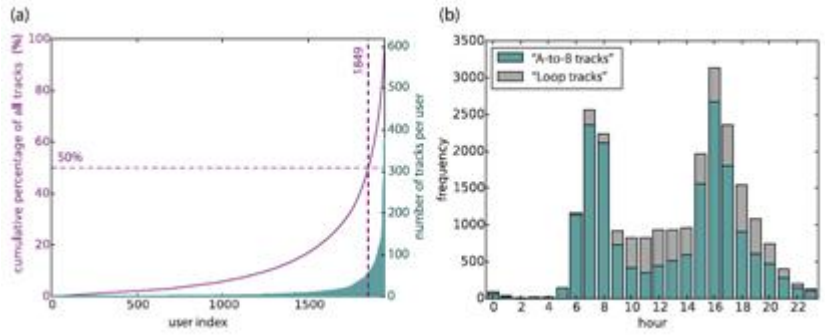


Figure 8. (a) Le nombre de traces GPS enregistrées par individus et (b) le type de trajets (trajet entre deux points distincts ou trajet en boucle) par heure de la journée [Bergman & Oksanen 2016]. Ces graphiques illustrent l'hétérogénéité du nombre de contributions selon les contributeurs et les caractéristiques des traces.

La représentativité des données issues de téléphones portables dépend aussi de l'opérateur dont sont issues les données et du type de données considérée (voir partie 2.2). Les données appartenant aux opérateurs, celles-ci ne représentent donc que la proportion (et les caractéristiques) des personnes abonnées à chaque opérateur (Tatem et al., 2009; Fen Chong, 2012 p.82). Des redressements pour extrapoler à la population globale peuvent être effectués par les opérateurs, mais ceux-ci sont rarement explicités, en raison de leur importance commerciale, et méritent d'être donc évalués d'une manière ou d'une autre en croisant les données avec d'autres sources d'information.

2.3.4 MANIPULATION D'UN VOLUME IMPORTANT DE DONNEES

Le grand volume des données rend difficile la tâche d'analyse en raison du temps de calcul nécessaire. Des techniques de filtrage ou d'échantillonnage sont alors proposées dans la littérature. Le filtrage consiste en général à identifier et à éliminer le bruit ou les données considérées aberrantes (Ivanovic et al., 2016). Iovan et al., (2013) proposent une série d'analyses qui permettent de filtrer des données aberrantes parmi les données issues de téléphones mobiles tels que la recherche de "ping-pong" (allers-retours peu probables) ou l'identification des données redondantes. L'échantillonnage, quant à lui, consiste à sélectionner une catégorie représentative « d'individus » dans la masse des données. Certaines études sélectionnent au hasard un échantillon d'utilisateurs à partir de l'intégralité du jeu de données (González et al., 2008). D'autres essaient d'optimiser cette méthode en prenant en compte, par exemple, uniquement les utilisateurs avec un certain nombre d'événements ou de localisations enregistrés (Tatem et al., 2009; Song et al., 2010; Onnela et al., 2011). Au premier regard, le second choix semble être plus pertinent car avoir plus de traces pourrait rendre l'analyse plus précise. Cependant, Smoreda et al. (2011) ont démontré qu'il existe une corrélation entre le nombre d'appel émis et les déplacements

effectués. Par conséquent, ne sélectionner que les utilisateurs ayant un nombre important de localisations semble introduire un biais dans l'analyse car cela revient à surreprésenter les individus avec une importante mobilité. On peut donc perdre la représentativité de la population si on cherche à optimiser le nombre de données considérées.

Par ailleurs, dans le but de détecter des données particulières, et non de rechercher des données représentatives, des méthodes statistiques d'échantillonnage peuvent être mises en places pour la sélection des zones d'intérêt particulières ou pour la détection de données à valider interactivement car avec des spécificités locales. On rencontre ce type d'approche dans les travaux liés à la collecte volontaire active pour des besoins spécifiques tels que l'estimation de l'occupation du sol (Projet européen Lucas, GeoWiki), l'identification d'espèces (VigiNature), ou la collecte de photos dans le monde (Projet Degree Confluence) par exemple.

2.3.5 RESPECT DE LA VIE PRIVEE

Les données de téléphonie mobile ainsi que les données géolocalisées sont des données sensibles car elles peuvent fournir des informations individuelles de grande précision spatiale et temporelle. L'anonymisation est donc nécessaire. Mais supprimer le nom de la personne associée aux données ne garantit pas l'anonymat, car des analyses et croisement de données peuvent permettre de retrouver les personnes concernées.

Plusieurs articles mentionnent la protection par les compagnies de téléphone des données individuelles. Leur diffusion est d'une part limitée et d'autre part anonymisée. Des méthodes d'anonymisation adaptées aux données de téléphonie mobiles sont proposées selon les études. Une méthode est d'agréger l'information à un niveau spatial, temporel ou des individus. Dans une étude en épidémiologie en Tanzanie de Tatem et al. (2009), les données sont localisées au niveau des zones administratives par abonné. Ceci limite donc l'analyse spatiale à un niveau de détail de six zones pour un pays d'environ 1 million de km². Il est également précisé que les données ne sont conservées que 3 mois par la compagnie de téléphonie, ce qui empêche des analyses de données passées. On retrouve des principes similaires dans Becker et al. (2013). Dans cet article, trois étapes sont suivies : 1) le remplacement des numéros de téléphone par un identifiant effectué par une personne non impliquée dans l'étude ; 2) la suppression de toutes les données non pertinentes pour l'étude comme ici les données socio-démographiques des abonnés ; 3) les résultats publiés le sont à un niveau agrégé, le niveau individuel n'étant publié qu'avec autorisation. Des méthodologies sont également développées avec les opérateurs dans le contexte de directives européennes de respect de la vie privée (Tiru et Ahas, 2012).

Pour plus de détail sur les difficultés de l'anonymisation et les techniques pour la réaliser, voir la partie 4.5 de ce document.

2.3.1 ASPECTS LEGAUX

Les aspects légaux ne sont pas abordés dans ce document, mais il ne faut oublier que les données issues de la foule, même lorsqu'elles sont diffusées sans restriction technique, ne sont pas libres de droits. Elles sont associées à des licences définies par les entités en charge de la collecte et la diffusion de ces données, souvent des associations ou entreprises

commerciales. Certains des travaux qui sont présentés ci-dessous ne respectent probablement pas les licences des données utilisées. Ceci est d'autant plus complexe à évaluer que les travaux s'appuient souvent sur plusieurs sources de données, aux licences hétérogènes, causant de véritables casse-têtes juridiques. Notons aussi que manipuler des données issues de la foule nécessite souvent un enregistrement auprès de la CNIL.

3 EXEMPLES D'ANALYSES DES DONNEES GEOLOCALISEES ISSUES DE LA FOULE

3.1 ETUDE DES COMPORTEMENTS ET DE L'ESPACE PRATIQUE

Le dénombrement des touristes dans un espace, l'analyse de leurs activités et déplacements, ainsi que le suivi de leur durée de séjour reposent souvent sur des enquêtes ou des monographies réalisées sur des groupes de population particuliers (Simon, 2010). En France, l'INSEE coordonne de nombreuses enquêtes portant sur la fréquentation de l'hôtellerie, les vacances, etc. La Direction du tourisme coordonne pour sa part une enquête mensuelle sur le suivi de la demande touristique et une autre sur la présence des visiteurs étrangers (Terrier, 2006). Conjointement à cela, des protocoles d'enquêtes qui consistent à combiner les techniques d'enquête traditionnelles et suivis GPS ont aussi été récemment définis et menés auprès d'individus volontaires (Christophe *et al.*, 2010; Beeco et al., 2013). Les principales différences entre données issues d'enquêtes et données issues d'observations ont été étudiées explicitement dans certaines applications, par exemple dans Bricka et al. (2012) sur des analyses de déplacements.

Ces données issues de la foule peuvent être considérées comme une information à part entière ou comme un complément à des données d'enquêtes, notamment lorsque les données GPS peuvent être associées à des comportements individuels (Beeco et al. 2013). Les données issues de la foule peuvent venir compléter des connaissances issues d'enquêtes ou, au contraire, des enquêtes peuvent venir raffiner l'analyse de données issues de la foule. Par exemple, dans Tatem et al. (2009) une perspective identifiée est de mener des enquêtes terrain auprès de voyageurs afin d'améliorer la compréhension de déplacements types identifiés au préalable par analyse de données de téléphonie.

A titre de premier exemple d'analyse, Olteanu-Raimond et al. (2012) s'intéressent quant à eux à l'estimation de la fréquentation des sites touristiques dans la région parisienne. Les touristes étrangers représentent 1,5 million d'utilisateurs anonymes, porteurs de cartes SIM non françaises. En utilisant des données de téléphones portables en zone parisienne et analysant la charge des antennes relais, Olteanu-Raimond et al (2012) confirment des connaissances préalables sur les pratiques touristiques selon lesquels les touristes en région parisienne faisant de courts séjours se focalisent sur l'hyper-centre touristique, tandis que ceux qui allongent la durée de séjour élargissent leurs pratiques de l'espace à la périphérie (Freytag, 2010) - cf. Figure 9. Plus précisément, la méthode d'estimation consiste d'abord à définir les lieux de stationnements. Ces derniers sont estimés a priori à partir des sites touristiques recensés par l'office de tourisme parisien et de la couverture définie par le réseau de télécommunication en région parisienne. Ainsi, pour chaque site touristique, le

lieu de stationnement est défini par la surface délimitée par la cellule de Voronoi[4] contenant le site touristique ainsi que les cellules voisines. Les critères utilisés pour identifier les lieux de stationnement empruntés par les trajectoires individuelles journalières des touristes sont la durée minimale de stationnement et le nombre minimum de points consécutifs inclus dans l'étendue du lieu de stationnement. Les horaires d'ouverture des sites touristiques ont été également pris en compte. Les estimations obtenues à partir de données de téléphones mobiles, ont été ensuite comparées aux estimations données par l'Office de Tourisme. Les auteurs montrent qu'il y a deux tendances : i) les vingt premiers sites touristiques les plus fréquentés selon l'office de tourisme ainsi que les sites moins fréquentés se retrouvent dans des déciles similaires, voire voisins; 2) les sites touristiques situés au milieu du classement du CRT se retrouvent dans des déciles plus éloignés (Figure 10).

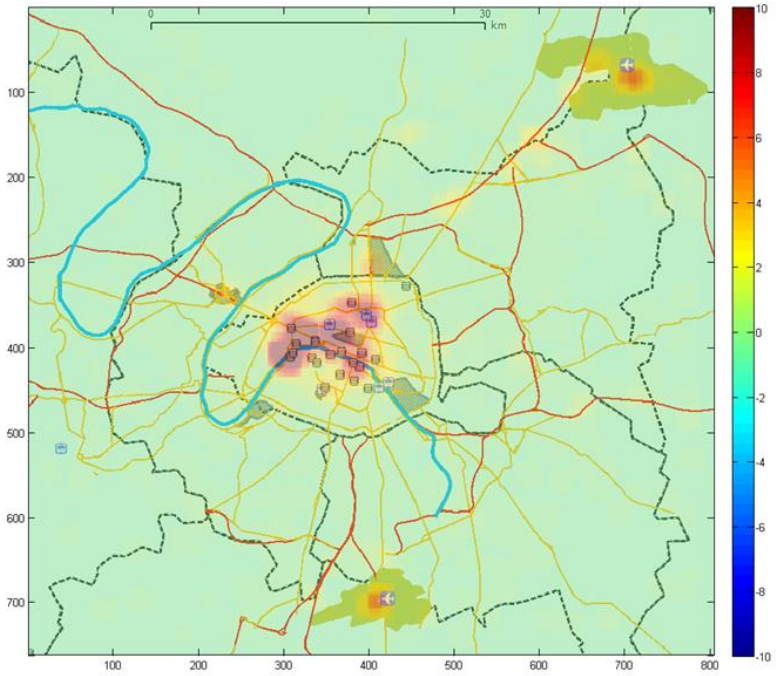


Figure 9. Densité de présence des excursionnistes (ici : touristes étrangers passant une journée en Ile-de-France) calculée à partir de données de téléphonie (Olteanu-Raimond et al. 2012).

[4] La cellule de Voronoï d'une antenne est ici une estimation de la zone qu'elle couvre.

Tableau 3. Comparaison des classements en déciles de la fréquentation des sites touristiques selon les deux types de données

Site	Décile Référence	Décile GSM	Comparaison
Disneyland Resort Paris	10	10	Même classement
Sacré-Cœur de Montmartre	10	10	
Tour Eiffel	10	10	
Notre-Dame de Paris	10	10	
Musée du quai Branly	10	10	
Centre Pompidou	10	10	
Le PANTHÉON	9	9	
Arc de Triomphe	9	9	
Musée national du moyen age	8	8	
Institut du monde arabe	8	8	
Domaine de Versailles	10	9	Classement proche
Cité des sciences et de l'industrie	10	9	
Musée d'art moderne de la ville de Paris	9	10	
Musée Rodin	9	10	
Musée Carnavalet	9	10	Classement différent
Musée du Louvre	10	7	
Musée national d'histoire naturelle	10	6	
Musée d'Orsay	10	2	

Figure 10. Comparaisons entre sources de données issues de téléphonie et des données issues de l'Office de Tourisme de Paris (Olteanu-Raimond et al. 2012).

Beeco et al. (2013) ont établi quant à eux un protocole de suivi individuel des touristes en menant à la fois des enquêtes ciblées et des analyses de traces GPS issus de traceurs GPS fournis aux visiteurs dans un site naturel aux États-Unis. Leur but est de lier le comportement type des touristes identifié par enquête à leurs déplacements réels. Pour cela ils ont identifié des types de trajectoires de déplacement parmi les traces étudiées, et ils ont cherché à déterminer si une corrélation peut être établie entre ces types de trajectoires et une typologie déterminée d'après enquête (ici planificateur vs. explorateur). Une des conclusions de l'article est une relative similarité dans les types de trajectoires pour tous les types de touristes, notamment au niveau des lieux de passages fréquentés et des routes empruntées. La Figure 11 présente la typologie ainsi qu'une carte des lieux de pause correspondant majoritairement à des points d'intérêt de tous les touristes suivis. Dans ce cas, les données GPS servent à questionner la pertinence de la typologie choisie pour l'enquête.

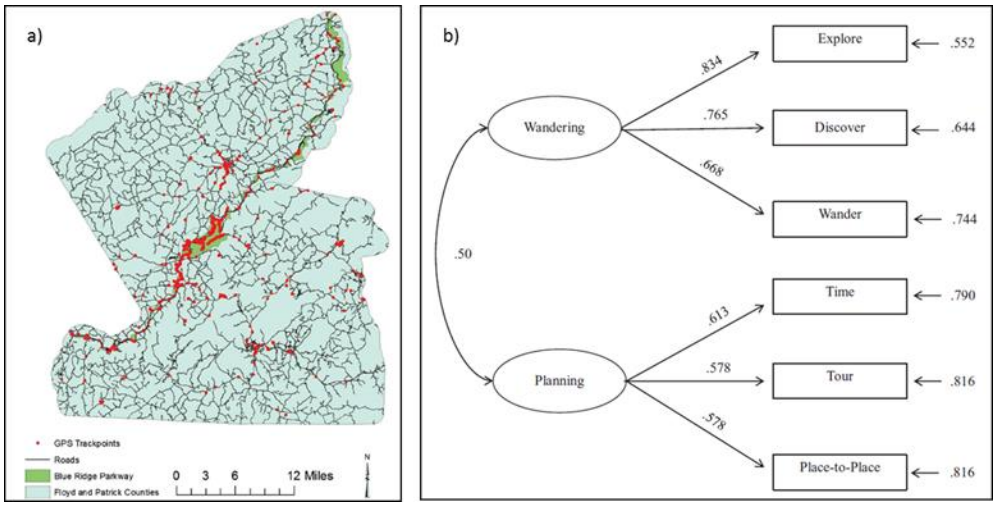


Figure 11. a) Localisations des pauses pour tous les groupes de touristes et b) la typologie de touristes mise en correspondance avec les déplacements enregistrés [Beeco et al. 2013]. L'étude ne montre pas de différence de comportement selon les catégories vis à vis des espaces visités et des routes empruntées.

D'autres recherches portent sur la distinction des comportements entre différents usagers d'un même espace. Par exemple, des différences de comportements peuvent être recherchées entre touristes et résidents (pour un exemple dans les Alpes, voir Bourdeau 2012) ainsi qu'entre touristes pratiquant des activités différentes et possiblement conflictuelles (voir par exemple Reis et Higham 2009). Ahas et al. (2007) étudient pour leur part le tourisme saisonnier en Estonie à partir de données de téléphones mobiles. En analysant la temporalité et la récurrence des données de téléphones mobiles, les auteurs parviennent à distinguer les déplacements résidentiels de ceux des saisonniers dans une zone située au bord de la mer. Les données de déplacements permettent donc ici de catégoriser les individus, et donc d'estimer les populations concernées ou de segmenter les populations analysées avant d'analyser plus en avant les comportements. Des connaissances ont par exemple été acquises dans ce même contexte sur les lieux de destinations privilégiées par les touristes ou les conséquences des conditions météorologiques sur leurs séjours (Figure 12 et Figure 13; Ahas, 2008).

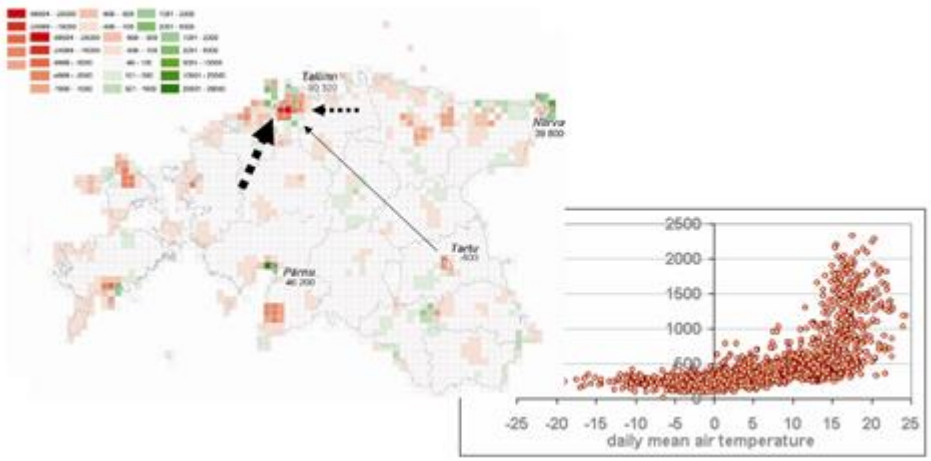


Figure 12. Probabilité de réaction à la météo des touristes en Estonie (changement de mobilité en vert ou absence de changement en rouge). Les changements de programme dus à la météo sont moins probables dans les grandes villes qu'ailleurs (Tallinn, Pärnu) [Ahas, 2008].

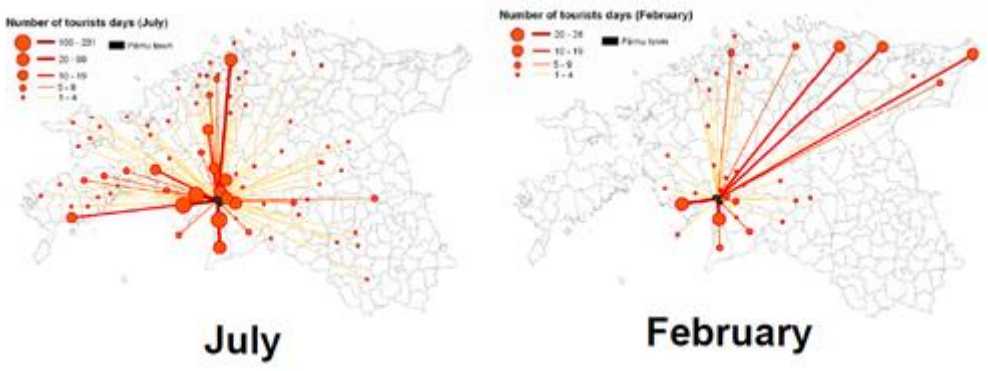


Figure 13. Déplacements de touristes restant plus de 3 nuits consécutives à Pärnu en Estonie (ville située au sud). Les mobilités des touristes dépendent du mois de l'année (ici différences entre juillet et février) [Ahas, 2008]

D'autres travaux s'intéressent à distinguer l'espace occupé par les touristes et les résidents dans des villes touristiques en utilisant d'autres données, comme des données issues du site de partage de photographies FlickR (Girardin et al., 2008). Ceux-ci utilisent les diverses informations disponibles sur les utilisateurs de FlickR (lieu de résidence) et sur leurs contributions (date et heure de la photo) pour analyser les comportements et pour identifier les touristes: par exemple, un utilisateur présent peu de temps dans un lieu loin de son lieu de résidence est classé comme touriste. Kadar et Gede (2013) ont proposé une adaptation de cette méthodologie pour visualiser et analyser les comportements des touristes et résidents par leurs contributions Flickr (Figure 14). Ils ont par exemple identifié par ce moyen un musée qui attirait plus de locaux que de touristes, ce qui a été vérifié a posteriori en croisant avec d'autres informations.

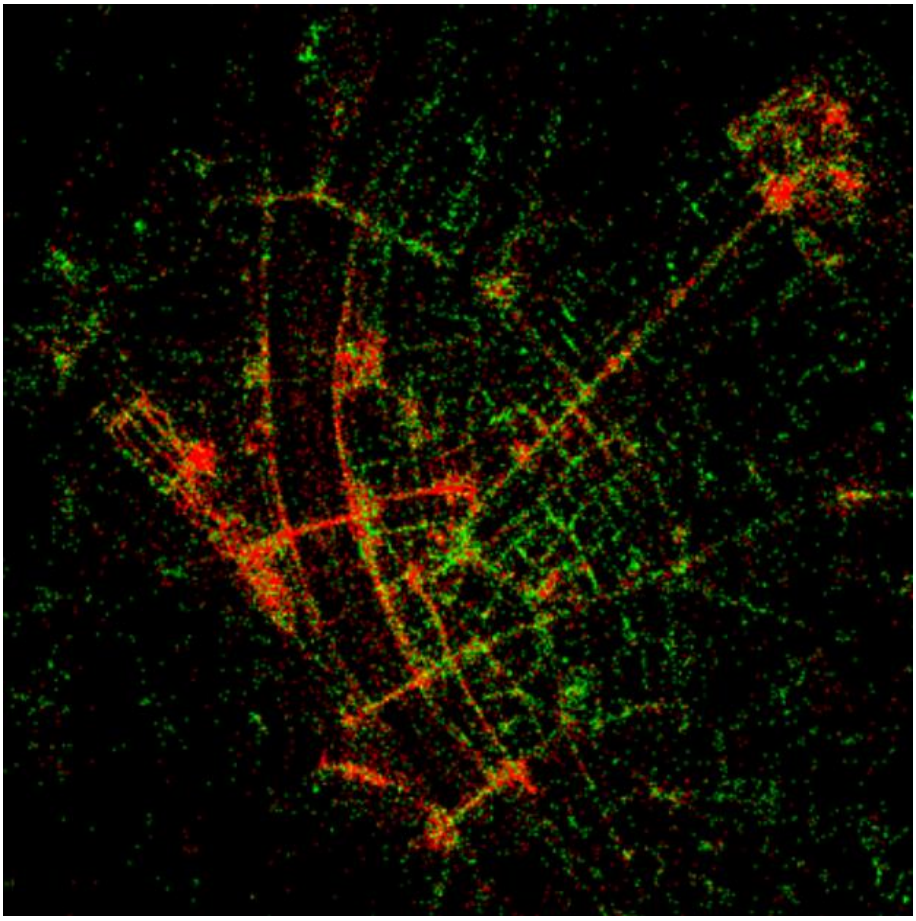


Figure 14. Visualisation des contributions dans Flickr de touristes (en rouge) et de résidents (en vert), à Budapest (Kadar et Gede 2013).

Les données de transport de type comptage, comme le taux d'occupation des stations Velib' ou Autolib' à Paris ont également été étudiées par des chercheurs pour en extraire des informations sur les dynamiques de déplacements urbains. Chabchoub et Fricker (2014) ont proposé des méthodes de clustering pour identifier les stations Velib' les plus souvent vides, pleines ou avec le plus de rotation (Figure 15). Des analyses similaires et concordantes ont été proposées par Côme et Oukhellou (2014), qui ont en plus étudié les corrélations spatiales entre le type d'utilisation des Velib' et la géographie des lieux où sont situés les stations. Lorsque les données sont plus riches, avec par exemple des informations sur les origines et destinations des déplacements de vélos, des analyses de flux sont possibles. Ainsi, Wood et al. (2011) proposent des méthodes pour visualiser et analyser des flux à partir des matrices origines/destinations du réseau de vélos en partage à Londres.

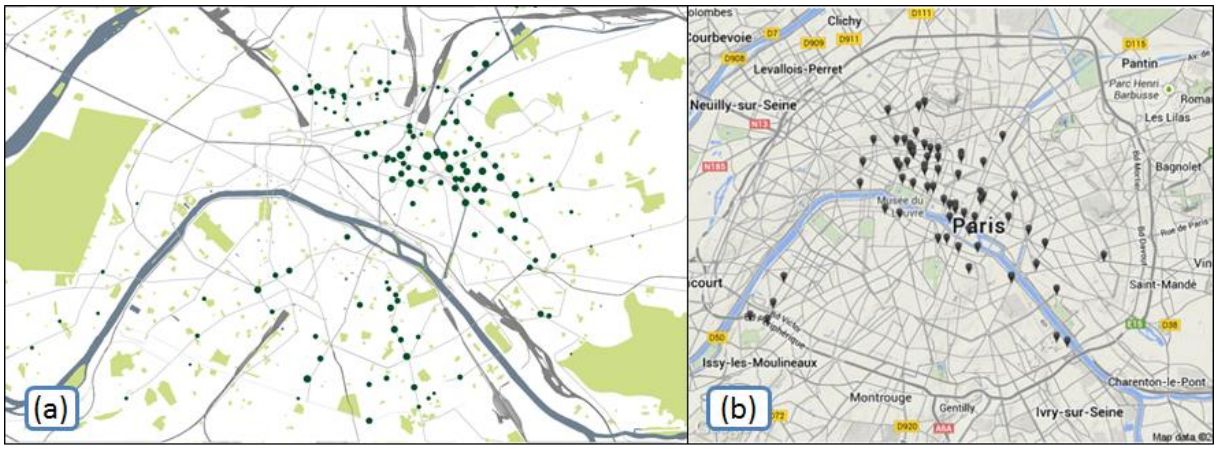


Figure 15. (a) Identification des stations Velib ayant le plus de "sorties" nocturnes (Côme et Oukhellou 2014) (b) stations souvent pleines (Chabchoub et Fricker 2014).

Relativement à une autre source d'information, une étude de Agryzkov et al. (2016) à partir des données de Foursquare montre l'intérêt et la difficulté des données issues de réseaux sociaux pour analyser l'espace et son usage. Les lieux d'intérêt du centre-ville de Murcie (Espagne) tels que recensés dans Foursquare ont été comparés avec des observations et des relevés sur le terrain. Le résultat est un nombre plus élevé de restaurants et de lieux de loisirs dans les données issues du réseau social que dans les observations terrain. Cela peut en partie s'expliquer entre autres par la non actualisation des données (lieux fermés entre l'étude des données du réseau et l'enquête terrain). A l'inverse des magasins existants ne sont pas représentés via le réseau social. Cela signifie que la liste de lieux géolocalisés de manière passive, ici par une application dont le but est d'indiquer la fréquentation d'un lieu et un avis, est incomplète. Cela pourrait également signifier au contraire que certains endroits peuvent être découverts comme lieux d'intérêt par l'usage alors qu'ils ne correspondent pas forcément à une description physique (par exemple des lieux de rassemblement).

3.2 DETECTION ET ANALYSE D'EVENEMENTS

Certains travaux de recherche se sont intéressés à la détection ou l'analyse d'événements particuliers à partir de données spatialisées issues de la foule. Il peut s'agir de la détection d'événements imprévus tels que les catastrophes naturelles (Lu et al., 2012; Bagrow et al. 2010) ou de l'analyse de la réponse à des événements prévus. Les travaux peuvent s'intéresser à des événements de manière générale, quel que soit leur type (Traag et al., 2011; Ferrari et al. 2014; Trasarti et al., 2015) ou s'intéresser à des événements spécifiques tels que des événements culturels (Smoreda et al, 2010), sportifs (Pulselli et al., 2008), ou de loisirs (Girardin et al. 2008; Calabrese et al., 2010).

Pour identifier des événements, de nombreux chercheurs analysent les séries temporelles définies à partir de l'activité téléphonique des antennes. Les séries temporelles mesurent soit la variation de l'activité téléphonique (Andrienko et al., 2013) soit la variation de la densité de présence (Trasarti et al., 2015) de certaines zones choisies a priori (musées,

stade, etc) ou détectées a posteriori. Ces dernières sont caractérisées par un changement de signal par rapport à une situation dite “normale” (Figure 16). Quelques exemples ci-après.

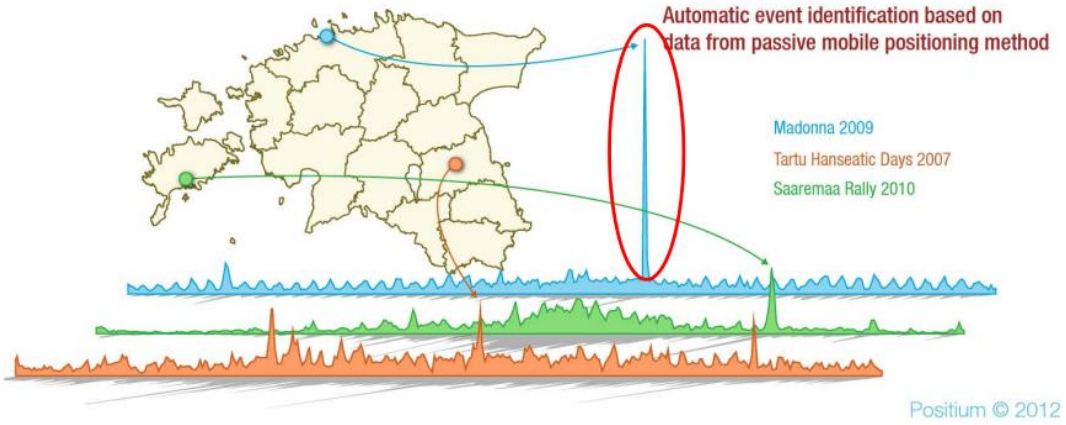


Figure 16. Exemple de détection d’événements à partir de données de téléphones mobiles (d’après Ahas, 2012)

Bagrow et al., (2011) s’intéressent à estimer si on peut détecter à partir de données de téléphones portables des événements de type “catastrophe” (attaque à la bombe, crash d’avion, orage, etc.). Des événements témoins ont d'abord été localisés dans le temps et dans l’espace en analysant les données média et réseaux sociaux de 2007 à 2009. Ensuite l’activité téléphonique (nombre d’appels agrégé à l’antenne) a été analysée. Les auteurs montrent que l’activité téléphonique autour des événements catastrophiques est différente de celle autour des événements “normaux” de types culturels. Comme nous nous pouvons le remarquer dans la Figure 17, pour les événements catastrophiques il n’y a pas de décalage temporel entre le moment où l’événement se produit et l’augmentation de l’activité téléphonique qui est très rapide, contrairement aux événements culturels. Le seul événement qui est exclu de cette règle dans cette figure est le crash d’avion. Ce type d’événement, se produisant en général, dans des endroits plus isolés nécessite une certaine période de temps pour que la nouvelle se diffuse et que les individus commencent à appeler.

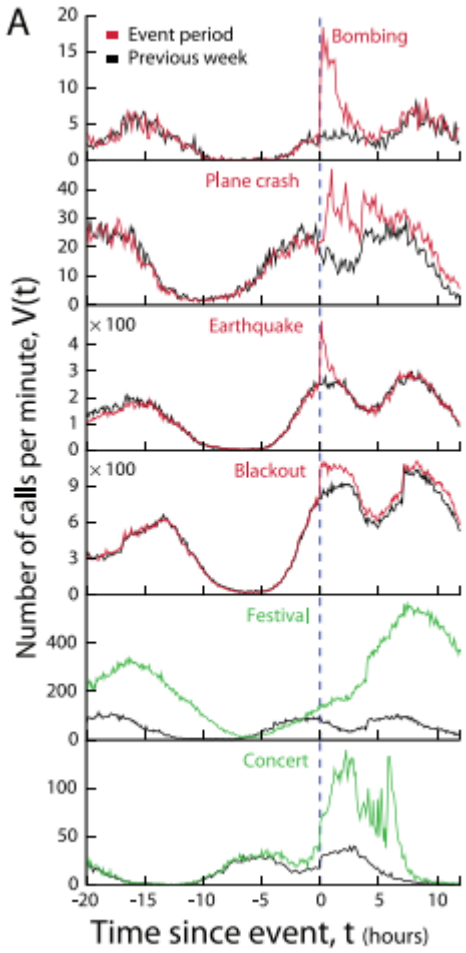


Figure 17. Séries temporelles illustrant l'activité téléphonique autour d'événements spécifiques (d'après Bagrow et al., 2011).

Twitter est également utilisé comme une source d'information pour détecter en temps réel des évènements (McEachren et al 2010), et notamment des catastrophes naturelles, pour lesquelles on utilise à la fois la localisation des tweets et le texte de ces tweets. Par exemple, Sakaki et al (2010) et Liu et al (2012) proposent des méthodes pour détecter en temps réel des tremblements de terre en analysant les tweets publiés par les victimes (Figure 18). Certains travaux étudient également comment les utilisateurs de Twitter s'auto-organisent pour aider les victimes lors d'une catastrophe ou d'un attentat (Starbird et Palen 2011).

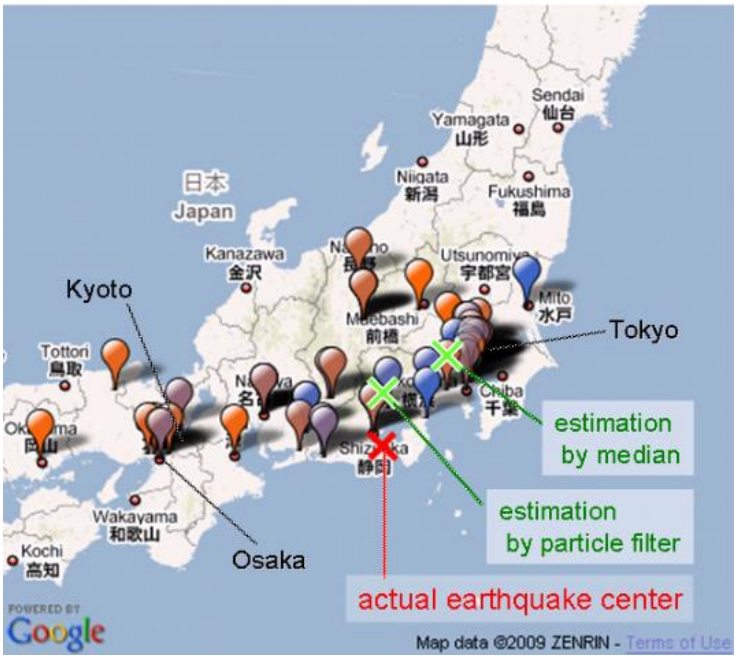


Figure 18. Estimation de la localisation de l'épicentre d'un tremblement de terre par analyse de tweets (Sakaki et al. 2010)

Calabrese et al. (2010) estiment quant à eux le nombre de participants et leur origine pour 52 événements qui ont lieux en 6 lieux à Boston (différents et suffisamment éloignés pour éviter l'ambiguïté entre des événements proches). A partir de données de téléphonie mobile, les auteurs identifient les lieux de domicile des individus ainsi que leurs arrêts. L'analyse spatio-temporelle des arrêts dans les lieux des événements ont permis d'estimer la fréquentation des événements, ainsi que d'étudier les distances entre lieu de résidence et lieu de l'événement, en fonction du type d'événement (Figure 19).

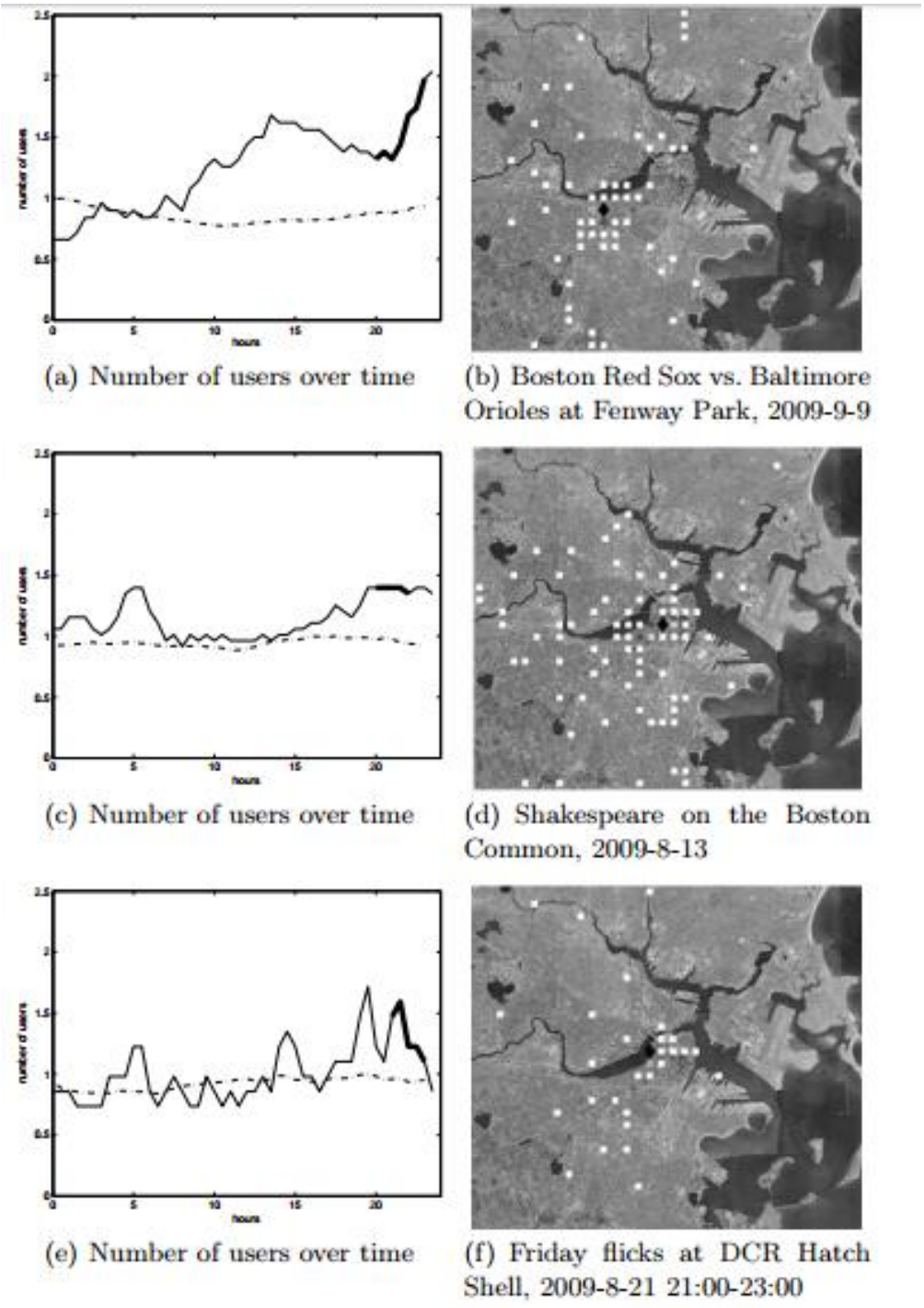


Figure 19. Estimation du nombre d'utilisateurs par événement et leurs lieux d'origine (d'après Calabrese et al., 2010)

3.3 ESTIMATION DES FLUX ORIGINES/DESTINATIONS

La construction des flux origines/destinations (O/D) est réalisée traditionnellement à partir de données issues d'enquêtes par comptages ou d'enquêtes rétrospectives. Les flux représentent des motifs de déplacement observés, comme les flux domicile-travail par exemple. En France par exemple, l'INSEE publie plusieurs données de flux O/D : celles construites à partir du recensement de population et le fichier MOBPRO 2007 ou l'Enquête Globale Transport – EGT, pour ne citer que celles-ci.

Avec l'arrivée de données issues de la foule de nouvelles approches ont également été mises en œuvre pour inférer des matrices de flux O/D à partir de données issues de capteurs mobiles tels que les téléphones (cf. Figure 20). Diverses approches existent. Friaz-Martinez et al. (2012) proposent une méthode qui permet de calculer des flux en utilisant directement les traces des téléphones portables. La proposition de Byeong-Seok et al. (2005) s'appuie sur des durées de présence en un lieu défini par l'aire de couverture de l'antenne réceptrice. L'approche de Calabrese et al. (2011) et Wang et al., (2013) consiste, quant à elle, à identifier les trajets et les arrêts qui composent chacune des trajectoires. Caceres et al. (2007) et Giannotti et al. (2011) proposent de découper les trajectoires en utilisant une fenêtre temporelle et de calculer les flux O/D en prenant le premier et le dernier point de la sous-trajectoire déterminée par le filtrage temporel. Inspirée par cette approche, Bahoken et Olteanu-Raimond (2013) proposent de prendre en compte les hétérogénéités spatio-temporelles des données de téléphones portables en prenant en compte les deux premiers/derniers points de la sous-trajectoire et en modélisant la notion d'appartenance aux voisins.

Dans sa thèse, Fen-Chong, (2012), quant à elle, propose deux méthodes de calcul de flux à l'échelle de cellules Voronoi et en utilisant les données issues de téléphones mobiles. Une première méthode consiste à compter les flux sortant et entrant entre toutes les cellules Voronoi; pour faciliter la lecture de flux elle s'appuie sur des techniques de filtrage et lissage. La deuxième méthode permet d'identifier les flux préférentiels. En s'appuyant sur la première loi de la géographie de Tobler "Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés", seules les cellules de Voronoi voisines sont prises en compte pour estimer le flux préférentiel d'une cellule Voronoi donnée. Pour diminuer l'effet produit par la structure du réseau de téléphonie, Fen-Chong (2012) fait la distinction entre le flux théorique basé sur une mesure de probabilité et le flux observé en prenant en compte que les flux impliquant les cellules voisines. L'intensité du flux préférentiel entre deux cellules Voronoi est calculée en fonction des flux observés et les flux théoriques.

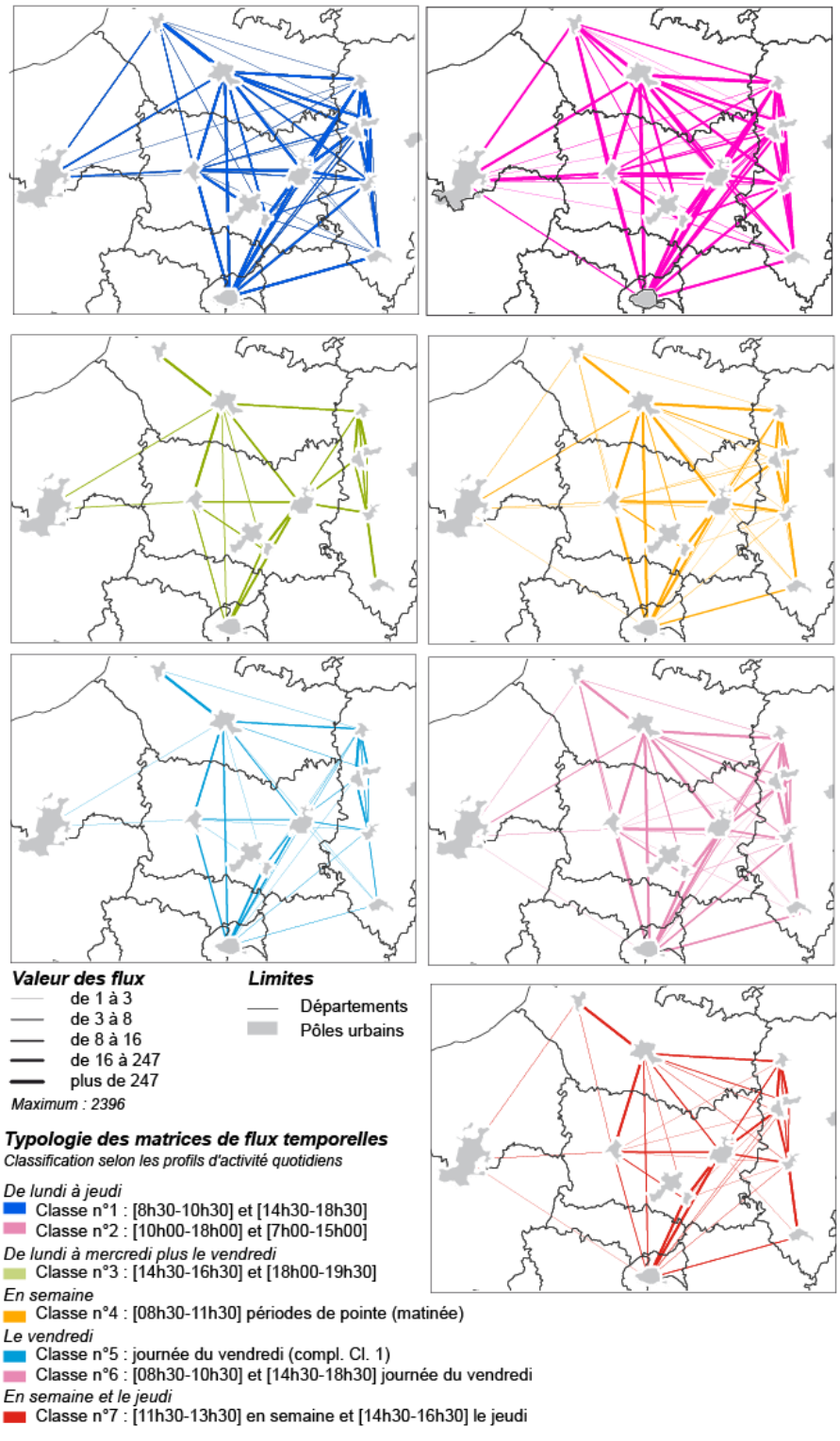


Figure 20. Estimation de flux origines/destinations à partir de données de téléphonie, à différentes heures de la journée, regroupées selon leurs similarités (Olteanu-Raimond et al. 2013)

Un aspect important pour l'analyse des flux est la définition des zones de référence des origines et destinations. Elles peuvent être discontinues (zones telles que les villes ou les aires urbaines) ou continues (une grille rectangulaire, les cellules Voronoï du réseau de télécommunication), connues a priori ou inférées (arrêts dans les trajectoires) à partir des déplacements. En fonction de ces caractéristiques les méthodes d'estimations proposées dans la littérature peuvent être plus ou moins fiables. Etant donné les hétérogénéités spatio-temporelles des téléphones portables, l'estimation de matrices de flux O/D peut être très sensible lorsque les zones OD sont discontinues (Olteanu-Raimond et al, 2013; Wang et al. 2013).

Un deuxième aspect à fixer est le choix de la temporalité. L'existence de données de téléphonie en continu permet d'agréger selon différentes résolution temporelles. En fonction de la manière dont la matrice de flux OD a été construite, les valeurs des flux peuvent être dans une fenêtre temporelle donnée, définie a priori, et sur un espace défini par les lieux d'origine et de destination (Caceres et al., 2007 ; Giannotti et al., 2011 ; Bahoken et Olteanu-Raimond, 2013) ou journalière, où les flux OD sont datée par temps de départ de la zone d'origine et temps d'arrivée dans la destination (Calabrese et al., 2011; Wang et al. 2013). Dans les deux cas, il s'agit donc d'une matrice de flux OD dynamique, la différence étant due au fait que le filtre temporel est appliqué avant ou après l'inférence des flux.

Notons que les méthodes d'inférence des flux O/D à partir de traces de téléphones portables capturent les déplacements de tous les individus, quel que soit leur moyen de transport. Apporter plus d'information sur les modes de déplacement, comme distinguer les déplacements motorisés et non motorisés, est souvent considéré comme une difficulté.

3.4 IDENTIFICATION DE REPERES INDIVIDU-CENTRES / ESPACE PERÇU

Les données générées par la foule permettent également de déduire la perception de l'espace géographique par les contributeurs. Les données de type enregistrement dans un lieu (ou "check-in" en anglais), comme celles contribuées dans Foursquare (devenu Swarm) ou Facebook, peuvent permettre de définir les lieux les plus importants dans la perception de l'espace par un utilisateur ou par un profil d'utilisateurs. Ainsi, Quesnot et Roche (2015a et b) proposent de définir en plus de métriques de saillance visuelle pour les points de repère utilisés en navigation, des métriques de saillance sémantique, basées sur les enregistrements Swarm et la popularité des lieux dans ce réseau social.

Hollenstein et Purves (2010) utilisent eux Flickr pour définir une délimitation de lieux vernaculaires, comme le centre-ville d'une agglomération. Pour cela, ils analysent les "tags" associés aux photos de Flickr et y cherchent des mots clés comme "downtown", "innercenter" ou "citycenter". La figure suivante illustre le type de résultat que ces méthodes peuvent générer.

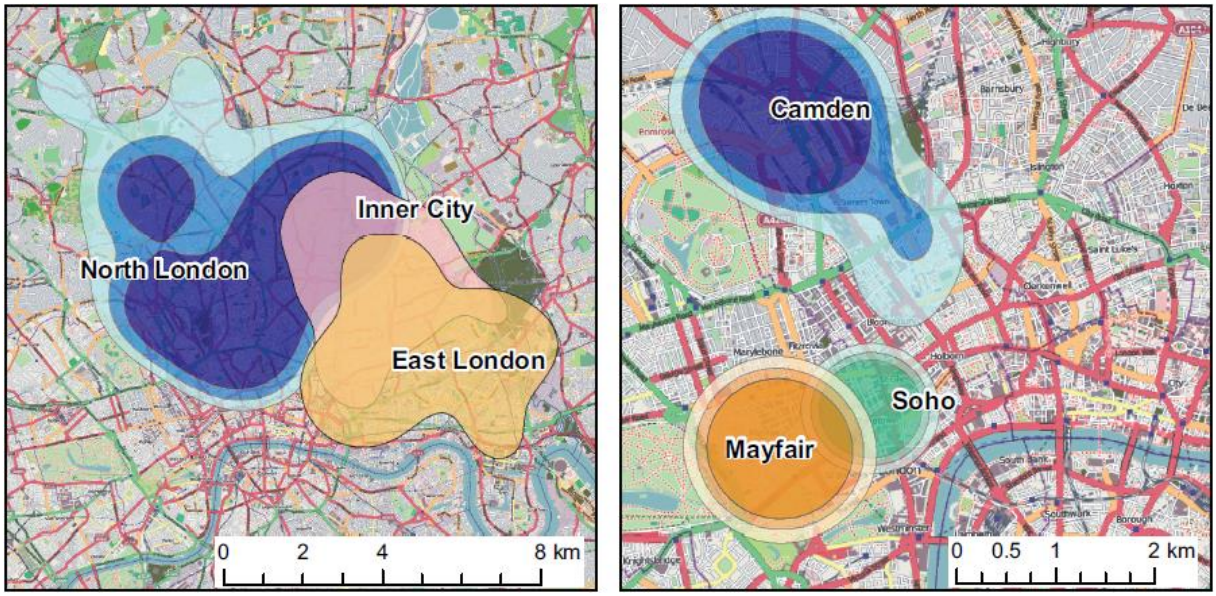


Figure 21. Identification de lieux perçus à Londres par analyse des tags associés aux photographies géolocaliséees de Flickr (Hollenstein et Purves 2010).

4 QUELQUES QUESTIONS METHODOLOGIQUES OU TECHNIQUES POUR L'ANALYSE DES DONNEES

4.1 ANALYSE DE DENSITE

Etant donné la nature épisodique des données issues de la foule ("episodic movement data", Andrienko, et al. 2012) et le fait que les données de téléphonie ne sont mises à disposition que de manière agrégée en raison de contraintes d'anonymisation de données, de nombreux études s'intéressent à la densité spatiale de présence, et à l'analyse de l'évolution de cette densité au cours du temps.

A partir de données individuelles (ou pré-agrégées à des antennes par exemple), des cartes de densité peuvent être réalisées, auquel cas une attention particulière doit être attribuée aux méthodes de lissage employées. Par exemple, pour déterminer l'espace occupé par les touristes étrangers dans l'hypercentre de Paris, Fen-Chong (2012) propose une méthode de lissage et étudie l'impact des différents lissages effectués sur les phénomènes observés (Figure 22). Pour calculer la densité de présence, l'espace peut aussi être divisé selon une grille régulière (Calabrese et al. 2007; Reades et al. 2009) ou irrégulière (Sevtsuk 2008; Vierira et al. 2010). Notons aussi que le phénomène observé peut être mesuré de différentes manières, sensiblement différentes et à choisir fonction de l'objectif de l'étude : nombre de personnes, nombre d'éventements, charge des antennes...

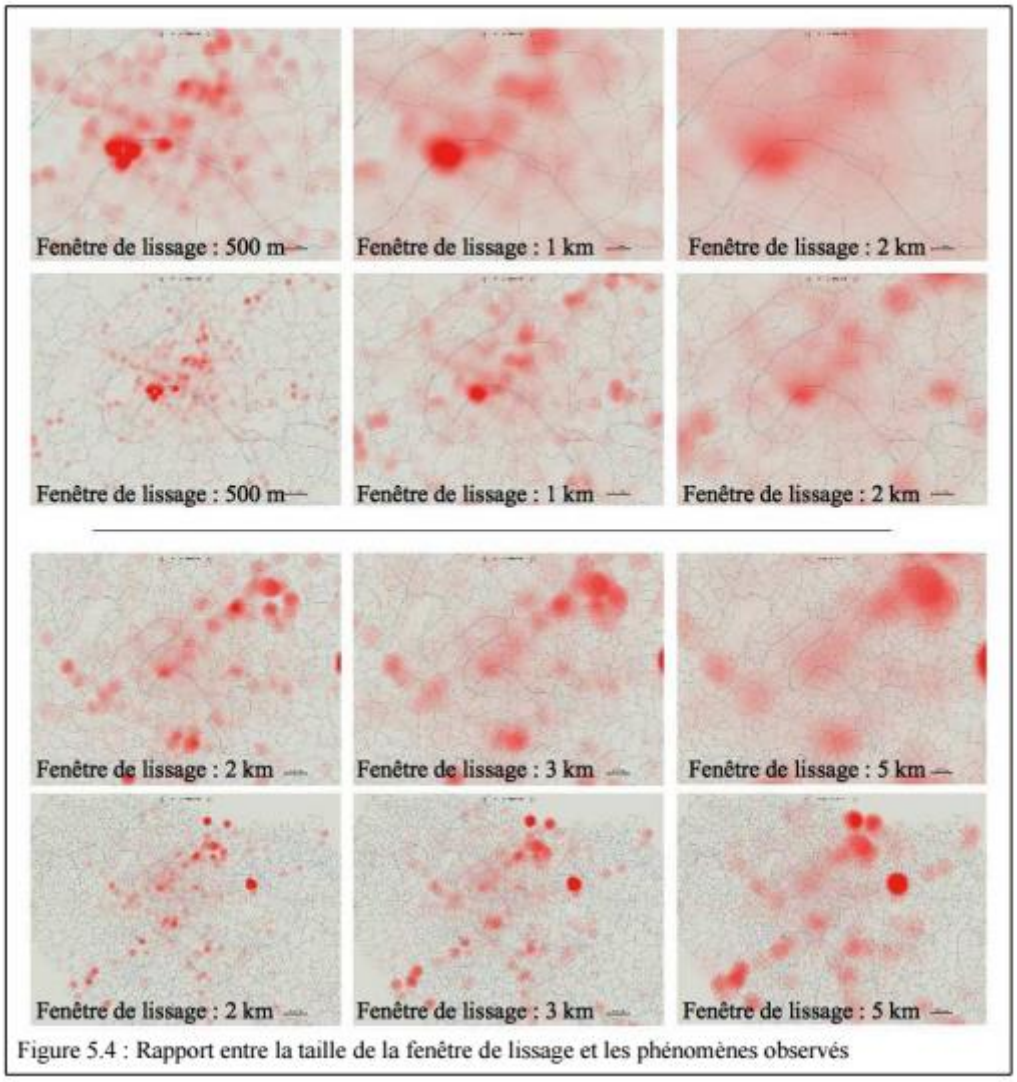


Figure 22. Espace occupé par les touristes étrangers à Paris déterminé par différents lissages de la charge des antennes relais (d'après Fen-Chong, 2012).

Certaines méthodes cherchent à définir les zones de forte densité sans a priori sur les limites des zones recherchées. En utilisant des données de téléphonie de Téléfonica de Madrid, Vieira et al. (2010) proposent par exemple une méthode nommée DAD-MST (Dense area Discovery-Maximum Spanning Tree) basée sur l'arbre couvrant de poids maximal pour détecter les zones denses. A partir de la tessélation de Voronoi (un pavage de l'espace estimant l'aire de couverture de chaque antenne téléphonique), l'algorithme cherche à détecter des zones de densité forte, disjointes, où l'activité téléphonique atteint un seuil maximum dans une période donnée.

Dans l'objectif de montrer un lien entre l'usage de téléphone mobile et les pratiques spatiales, Reades et al. (2009) proposent et comparent deux méthodes pour estimer la densité de présence d'individus à partir de données agrégées de téléphones portables. La première méthode consiste à rastériser la zone d'étude en 2115 pixels (500 m x 500 m). La densité de présence par maille considérée est alors la somme d'Erlang normalisée par la

taille du pixel. La deuxième méthode consiste à modéliser des patterns d'usage de téléphones portables en utilisant les vecteurs propres. Chaque maille est décrite par une matrice de dimension N*M (N = le nombre de jours d'observation et M = la fréquence d'enregistrement - toutes les 15 minutes dans cette étude). Chaque élément de la matrice correspond à la valeur d'Erlang normalisée par la valeur moyenne d'Erlang de la maille rapportée à la journée. A partir de cette matrice, on calcule la matrice de covariance puis les vecteur et valeurs propres. Les auteurs ont démontré que les deux premiers vecteurs propres suffisent pour reconstituer la signature principale de la maille. La figure suivante montre la densité de présence par maille en utilisant la première méthode (à gauche) et la deuxième méthode (à droite) à différents pas de temps : 1am, 5 am, 9 am, 5 pm, 9 pm..

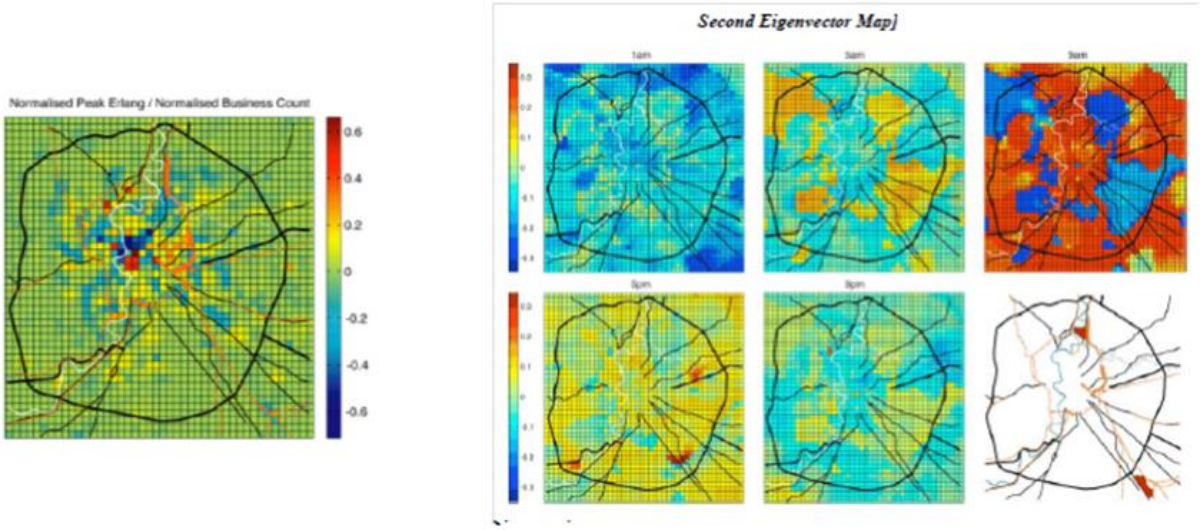


Figure 23. Valeurs d'Erlang normalisées à l'échelle de la cellule (à gauche) et deuxième vecteur propre à différents pas de temps (d'après Reades **et al.** *2009)*

4.2 RECALAGE ET SEGMENTATION DE TRAJECTOIRES

Les données n'étant pas nécessairement localisées avec une très grande précision (cf. partie 2.3.1), il est parfois nécessaire de recaler ces données sur une certaine référence spatiale, afin d'agréger et d'analyser les données. Par exemple dans Bergman et Oksanen (2016), les trajectoires GPS des cyclistes sont recalées aux voies de transport cartographiées dans OpenStreetMap. Dans ce cas, des méthodes dites d'appariement sont nécessaires. L'appariement est un problème qui peut être complexe, car recaler des données sur la route la plus proche n'est parfois pas la meilleure solution. Diverses méthodes de recalage existent donc (Figure 24).

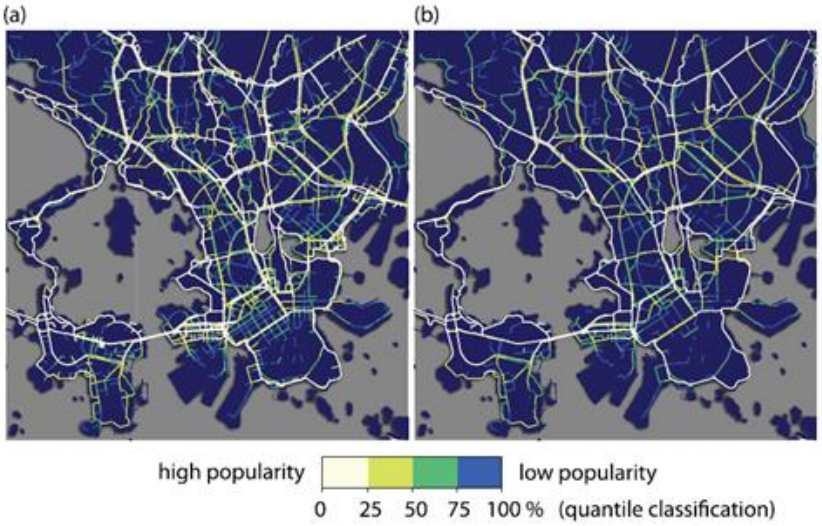


Figure 24. Popularité des tronçons empruntés (nombre de cyclistes) suite à l'appariement par deux algorithmes différents (a) geomeric point-to-curve (b) Hidden Markov Model-based matching. L'algorithme HMM a une approche plus globale de l'appariement et donne de meilleurs résultats par rapport à l'algorithme dit géométrique qui a une approche locale de l'appariement [Bergman & Oksanen 2016)].

Par ailleurs, afin de caractériser l'utilisation de l'espace, une approche consiste à décomposer des trajectoires en segments homogènes. La segmentation des trajectoires s'appuie sur des caractéristiques géométriques ainsi que sur le contexte spatio-temporel. Siła-Nowicka et al. 2016 s'intéressent, à partir de données de traceurs GPS, à la détection automatique de lieux importants pour les personnes suivies. Ces lieux correspondent à la maison, au travail, à des endroits de sociabilisation. Un enjeu est alors de définir des méthodes d'analyse spatiale qui permettent d'extraire automatiquement ces lieux à partir des traces GPS (Figure 25).

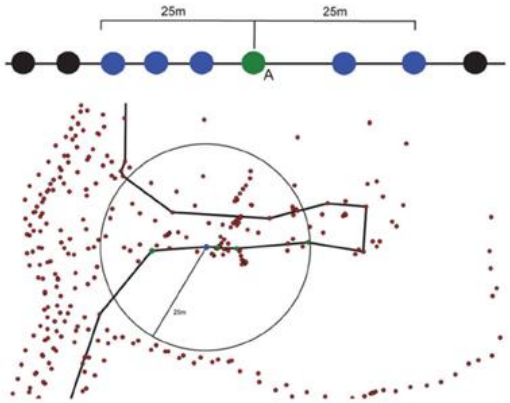


Figure 25. Détection de clusters de points selon la méthode STKW (Spatio-temporal Kernel Window). La détection de ces clusters permet ensuite de déterminer des lieux importants pour les personnes [Siła-Nowicka et al. 2016]

4.3 ANALYSE DE TEXTE

Les données de type textuel, comme celles issues de Twitter, ont la particularité de posséder parfois une géolocalisation par coordonnées, souvent issue du GPS du smartphone d'où à été créée la donnée, et parfois une autre localisation souvent implicite dans le texte. Par exemple, il existe des tweets localisés à Roland Garros, des tweets contenant "Roland Garros" dans leur texte, et parfois les deux. Mais il n'est pas évident que l'on tweete à propos du lieu où on est physiquement. Hahmann et al. (2014) ont étudié la relation entre lieu du tweet et contenu du tweet pour différents types de lieux. Ils montrent que cette corrélation lieu/contenu est très variable en fonction du type de lieu : par exemple, elle est assez importante pour les gares et très faible pour les supermarchés. De la même manière, Leetaru et al. (2013) montrent que peu de tweets sont explicitement localisés par les contributeurs et qu'une grande partie de la géographie de Twitter est contenue dans le texte des messages. Leur analyse, pour passer à l'échelle, nécessite alors l'utilisation de techniques de traitement automatique du langage. De nombreuses techniques de reconnaissance automatique de lieux nommés dans les textes existent. Elles trouvent néanmoins souvent leurs limites dans les textes très court, où les éléments de contexte sont peu nombreux pour interpréter les mots utilisés. Ils s'avèrent donc assez difficile à mettre en œuvre sans un taux d'erreur significatif dans des données du type de celles de Twitter ou de descriptions associés à des photographies. Comme déjà illustré (Figure 21), ces techniques peuvent quand même être mises en œuvre dans des gros volumes de données, le nombre masquant une partie du bruit.

4.4 ANALYSE VISUELLE

Notons que l'analyse des données peut être automatique ou non. Des outils d'analyse visuelle ont été développés ces dernières années dans le but de mieux détecter et analyser les événements particuliers dans des données volumineuses. Citons à titre d'exemple, l'équipe de G. Andrienko qui propose une série de méthodes et d'outils via le logiciel Common GIS (http://geoanalytics.net/and/).

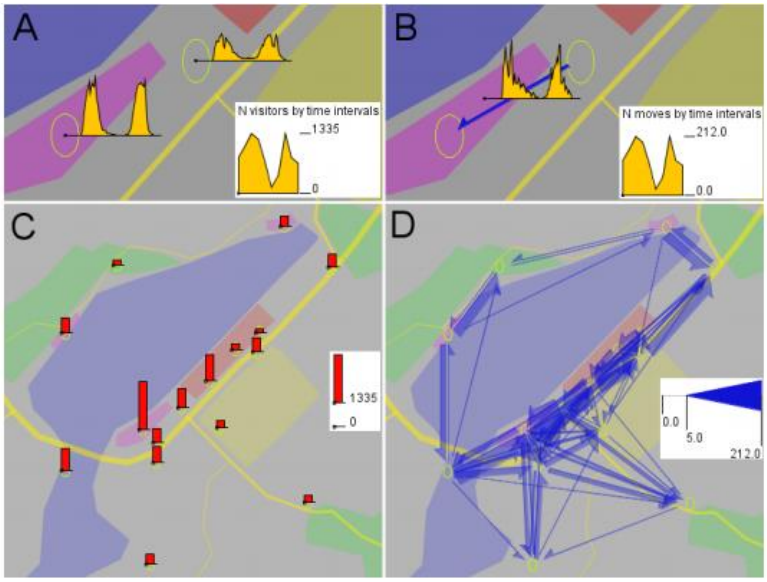


Figure 1. Different views on aggregated movement data. A, B: Time series associated with places (A) and links (B). C, D: spatial situations in terms of presence (C) and flows (D).

Figure 26. Exemple d'outils d'analyse visuelle pour les données agrégées (Andrienko et al. 2012)

Au sein du projet UrbanMobs (http://www.urbanmobs.fr/fr/), des chercheurs de l'opérateur Orange proposent aussi des méthodes pour identifier et visualiser les évènements ponctuels tel que la fête de la musique, le nouvel an, des concerts ou des matchs de football. Les données issues de téléphones portables sont pour cela agrégées par antenne et par minute.

4.5 TECHNIQUES D'ANONYMISATION

La question de l'anonymisation de données individuelles est rendue difficile par le fait que la simple suppression des étiquettes, des attributs permettant de relier immédiatement des données à un individu (nom, numéro de téléphone), ne garantit pas l'élimination du danger d'atteinte à la vie privée. En effet, Sweeney (2002) a montré qu'un individu malveillant pouvait utiliser certaines informations contenues dans les données et les croiser avec d'autres bases de données libres ou avec d'autres connaissances préalables pour retrouver le propriétaire de ces données. Dans cet article est définie la notion de quasi-identifiant, un ensemble d'attributs susceptibles de conduire à la ré-identification des propriétaires de données publiées dans une base. Dans le cas de bases de données tabulaires, les quasi-identifiants sont des attributs tels que la date de naissance, le sexe et le code postal, un triplet contenant ces informations pouvant avoir un unique individu pour antécédent ; d'autres types de données sont susceptibles d'être des quasi-identifiants selon les circonstances. En tenant compte de ce risque de ré-identification, une stratégie possible pour l'anonymisation de données est le k-anonymat, dont le principe est que chaque entrée d'une base de données doit prendre les mêmes valeurs que k-1 autres entrées dans les champs susceptibles d'être des quasi-identifiants. Cette propriété peut être obtenue en généralisant suffisamment les valeurs prises par ces attributs, voire en supprimant les entrées qui ne peuvent être anonymisées par généralisation. Ce traitement est approprié

pour la publication en une seule fois d'une base de données préalablement constituée et anonymisée en bloc (anonymisation dite offline), par opposition à une base publiée au fur et à mesure de sa constitution.

Dans le cas de l'anonymisation de bases de trajectoires, la définition des quasi-identifiants, et plus généralement la modélisation des connaissances d'un adversaire n'est a priori pas naturelle, car il s'agit souvent de la connaissance des positions de certains individus à certaines dates, connaissance qui peut être acquise par l'observation ou l'exploitation d'autres types de ressources. Cette modélisation de la connaissance adverse détermine les critères que doit respecter une base anonymisée et a un rôle crucial dans le choix de la méthode d'anonymisation et des traitements qu'on peut faire subir aux données. En conservant le formalisme des quasi-identifiants, on peut supposer que la connaissance d'un adversaire potentiel reste cantonnée à une partie d'une partition de l'espace (Terrovitis et al. 2008), ou au contraire que chaque trajectoire peut-être ré-identifiée par des éléments différents et possède son propre quasi-identifiant (Yarovoy et al. 2009). Si dans le premier cas la mise en œuvre de l'anonymisation diffère peu des méthodes utilisées pour les bases de données tabulaires, dans le deuxième cas on modifie l'ensemble des trajectoires de sorte que chacune ne puisse être distinguée de k-1 autres trajectoires en considérant seulement les informations contenues dans son quasi-identifiant individuel. Ce premier traitement ne crée pas des clusters disjoints de trajectoires, et une analyse globale des liens possibles entre des positions observées et la base anonymisée peut permettre de ré-identifier certains éléments. Yarovoy propose une nouvelle opération pour rendre impossible une telle inférence. D'autres modèles cherchent à s'affranchir de la définition explicite de quasi-identifiants et cherchent à constituer des clusters de trajectoires proches, en se fondant sur les imprécisions inhérentes aux données géo-localisées pour modifier des trajectoires et les rendre indistinguables les unes des autres (Abul et al. 2008), ou en regroupant des trajectoires proches et en calculant des parallélépipèdes spatio-temporels regroupant les points de ces trajectoires (Nergiz et al. 2008).

5 REFERENCES

Abul, O., Bonchi, F., and Nanni, M. (2008) Never Walk Alone: Uncertainty for anonymity in moving objects databases. Proc. of the 24nd IEEE Int. Conf. on Data Engineering (ICDE'08).

Agryzkov T., Martí P., Tortosa L., Vicent J.F. (2016) Measuring urban activities using Foursquare data and network analysis: a case study of Murcia (Spain). International Journal of Geographical Information Science, doi:10.1080/13658816.2016.1188931

Ahas R., Aasa A., Roose A., Mark Ü, Silm A. (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study, Tourism Management, 29, p 469-486

Ahas R., 2008, Using mobile positioning data for mapping space-time behavior and developing LBS: Experiences from Estonia. Communication à CartoTalk, May, 8, 2008, lien: http://cartography.tuwien.ac.at/cartotalk-by-rein-ahas/

Ahas, 2012, PASSIVE ANONYMOUS MOBILE POSITIONING DATA FOR TOURISM STATISTICS, In 1th Global Forum on Tourism Statistics http://www.congress.is/11thtourismstatisticsforum/presentations/Rein_Ahas.pdf

Ahas, R., A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru & M. Zook (2015): Everyday space – time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn, International Journal of Geographical Information Science, DOI:10.1080/13658816.2015.1063151.

Antoniou, V., Morley, J. and Haklay, M., 2009. The role of user generated spatial content in mapping agencies. Proceedings of GISRUK conference.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T. et D. Hecker. (2012) Visual analytics for understanding spatial situations from episodic movement data. KI - Kunstliche Intelligenz, pp. 241–251.

Andrienko , G., N. Andrienko , G. Fuchs , A.-M. Olteanu-Raimond and J. Symanzik (2013) Extracting Semantics of Individual Places from Movement Data by analyzing Temporal Pattrens of visits, 21 st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.

Bagrow JP, Wang D, Baraba´si A-L (2011) Collective Response of Human Populations to Large-Scale Emergencies. PLoS ONE 6(3): e17680. doi:10.1371/journal.pone.0017680.

Bahoken, F., Olteanu-Raimond AM, (2013) Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement, In proceedings of ICC.

Byeong-Seok Y, Kyungsoo C (2005) Origin-destination estimation using cellular phone as information. Journal of the Eastern Asia Society for Transportation Studies, 6:2574–2588.

Becker R., Cáceres R., Hanson K., Isaacman S., Loh J.M., Martonosi M., Rowland J., Urbanek S., Varshavsky A., Volinsky C. (2013) Human Mobility Characterization from Cellular Network Data. Communications of the ACM, Volume 56(1), p 74-82, doi:10.1145/2398356.2398375

Beeco J.A., Huang W.-J., Hallo J. C., Norman W.C., McGeehee N.G., McGee J., Goetcheus C., 2013, GPS Tracking of Travel Routes of Wanderers and Planners. Tourism Geographies: An International Journal of Tourism Space, Place and Environment, Volume 15(3), p. 551-573, DOI: 10.1080/14616688.2012.726267

Bergman C., Oksanen J., 2016, Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing. Transactions in GIS

Bourdeau P., 2012, Visiting/living (in) the Alps: towards a tourist-residential convergence?. Mauro Varotto; Benedetta Castiglioni. Di chi sono le Alpi? : appartenenze politiche, eco-nomiche e culturali nel mondo alpino contemporaneo, Padova University press, p.195-204

Bricka S.G., Sen S., Paleti R., Bhat C.R. (2012) An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation Research Part C, p. 67-88.

Brando et Bucher 2010. Quality in User Generated Spatial Content: A Matter of Specification. Proceedings of the 13th AGILE International Conference on Geographic Information Science 2010, Guimarães, Portugal

Brando 2013. Un modèle d'opérations réconciliables pour l'acquisition distribuée de données géographiques. Thèse de doctorat de l'université Paris Est, soutenue le 5 avril 2013

Bruns, A. 2008 Blogs, Wikipedia, Second Life and Beyond: From Production to Produsage , New York: Peter Lang

Caceres N., Wideberg J., Benitez F., (2007) Deriving origin-destination data from a mobile phone network, Intelligent Transport Systems, IET vol.1, n° 1, p. 15-26.

Calabrese , F., Collona, M., Lovisolo, P., Parata, D., et Ratti, C. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. Intelligent Transportation Systems, IEEE Transactions on 12, 1 (2007), 141-151.

Calabrese F., Pereira F., Di Lorenzo G., Liu L., Ratti C., (2010) The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events, Actes de 8th International Conference on Pervasive Computing, pp.22-37.

Calabrese F., Di Lorenzo G., Liu L., Ratti C. (2011) Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area, IEEE Pervasive Computing, vol.10, n°4, p.36-44

Campagna M., Floris, r., Massa, P., Girsheva A., and Ivanov K. The Role of Social Media Geographic Information (SMGI) in Spatial Planning. In Planning Support Systems and Smart Cities , Lecture Notes in Geoinformation and Cartography, S. Geertman et al. (eds.), DOI 10.1007/978-3-319-18368-8_3.

Chabchoub Y, Fricker C (2014) Analyse des trajets Vélib par clustering. Actes de l'atelier Clustering et Co-clustering, Extraction et Gestion de Connaissances (EGC), Janvier 2014.

Christophe S., Davoine P-A., Jambon F., André-Poyaud I., Chardonnel S., Lutoff C., Ahmed L., « Acquisition de connaissances sur les déplacements des individus dans un contexte de risques naturels. Protocole d'enquête à l'aide des technologies mobiles », Actes de Conférence SAGEO 2010, Toulouse.

Côme E., L. Oukhellou. (2014) Model-based count series clustering for Bike-sharing system usage mining, a case study with the Vélib' system of Paris. ACM TIST 5(3)

Couronné, T., A.-M. Olteanu-Raimond and Z. Smoreda (2011) Looking at spatio-temporal city dynamics through mobile phone lenses, IEEE International Conference of Network of the Future

Estima, J., Fonte, C.C., Painho, M., 2014. Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database, in: Proceedings of the 17th AGILE International Conference on Geographic Information Science. In Proccedings of 17th AGILE International Conference on Geographic Information Science, Castellon, Spain.

Fen-Chong, J. (2012) Organisation spatio-temporelle des mobilites revelees par la telephonie mobile en Ile-de-France. G´eographie. Universit´e Panth´eon-Sorbonne - Paris I, 2012. Fran¸cais. <tel-01004704>.

Ferrari, L., Mamei, M. & Colonna, M. J Ambient Intell Human Comput (2014) 5: 265. doi:10.1007/s12652-012-0169-0.

Fischer, F., 2012 VGI as big data: A new but delicate geographic data source. GeoInformatics, 3, p. 46–47.

Freytag, T., « Déjà-vu: tourist practices of repeat visitors in the city of Paris », Social Geography, n°5, 2010, pp. 49-58.

Frias-Martinez V., Soguero C., Frias-Martinez E. (2012) Estimation of Urban Commuting Patterns Using Cellphone Network Data, Actes du colloque ACM SIGKDD Workshop on Urban Computing, Beijin, China.

Fuchs, G., Andrienko, G., Andrienko, N. & Jankowski, P. (2013). Extracting Personal Behavioral Patterns from Geo-Referenced Tweets. Paper presented at the 16th AGILE Conference on Geographic Information Science, 14 - 17 May 2013, Leuven, Belgium.

Gao J.-H., Hsueh Y.-H., 2014, Exploring the Relationship between Traveler Types and Travel Route Types. International Journal of Basic & Applied Sciences IJBAS-IJENS Vol:14 No:02, p. 48-56

Gao S., Liu Y., Wang Y., Xiujun M. (2013) Discovering Spatial Interaction Communities from Mobile Phone Data. Transactions in GIS, 17(3): 463–481, doi: 10.1111/tgis.12042

Giannotti F., Nanni M., Pedreschi D., Pinelli F., Renso C., Rinzivillo S., Trasarti R. (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data, International Journal on Very Large Data Bases, vol. 20, n° 5, p.695-719.

Girardin F.,Calabrese F., Ratti C., Blat J., « Digital Footprinting: Uncovering Tourists with User-Generated Content », IEEE Pervasive Computing, vol. 7, n°4, 2008, pp.36-43.

Girres, J.-F. and G. Touya 2010. Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS, Volume 14, Number 4, page 435-460

Goodchild 2007: Citizens as sensors: the world of volunteered geography. In: GeoJournal.69 (4), 2007: pp. 215–217.

Goodchild, M. F., and Glennon, J. A. 2010. Crowdsourcing geographic information for disaster response: A research frontier. International Journal of Digital Earth, 3(3), 231–241

González MC, Hidalgo CA, Barabási AL, 2008, Understanding individual human mobility pat-terns, Nature 453: 779–782.

Hahmann, S., R. Purves, and D. Burghardt (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science 9, 1-36.

Hollenstein, L. and R. Purves (2010). Exploring place through user-generated content: Using flickr tags to describe city cores. Journal of Spatial Information Science 1, 21-48.

Iovan C., Olteanu-Raimond, A-M, T. Couronné and Z. Smoreda, 2013, Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies, Geographic Information Science at the Heart of Europe, pp. 247--265, Springer.

Ivanovic S., Olteanu-Raimond A-M, Mustière S., Devogele T., Detection of outliers in crowdsourced data, In Proceedings of the 12th International symposium on "Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 5-8 July, Montpellier, France.

Kadar, B. and M. Gede (2013). Where do tourists go? visualizing and analyzing the spatial distribution of geotagged photography. Cartographica: The International Journal for Geographic Information and Geovisualization 48 (2), 78-88.

Krumm, J., Davies, N., and Narayanaswami, C. 2008. User-generated content. IEEE Pervasive Computing, 7 (4), 10–11.

Le Corre N., Le Berre S., Brigand L., Peuziat I., 2012, Comment étudier et suivre la fréquentation dans les espaces littoraux, marins et insulaires ? De l'état de l'art à une vision prospective de la recherche. EchoGéo [En ligne], 19 | 2012, mis en ligne le 10 février, DOI : 10.4000/echogeo.12749

Lenhart A. (2009) Adults and social network Web sites. Pew Internet and American Life Project. <http://www.pewinternet.org/Reports/2009/Adults-and-Social-Network-Websites.aspx/> Accessed 19.03.09

Leetaru K.H, Wang S., Cao G., Padmanabhan G., Shook E. 2013. Mapping The Global Tweeter Heartbeat : The Geography of Tweeter. In First Monday, http://firstmonday.org/article/view/4366/3654

Liu, Sophia B., Beau Bouchard, Daniel Bowden, Michelle Guy, and Paul Earle. (2012). USGS Tweet Earthquake Dispatch (@USGSted): Using Twitter for Earthquake Detection and Characterization. Poster presented at the American Geophysical Union (AGU) 2012 Annual Meeting for the "Citizen Empowered Seismology" session in San Francisco, CA on December 4, 2012.

Louail T., Lenormand M., Ros O.G.C., Picornell M., Herranz R., Frias-Martinez E., Ramasco J.J., Barthelemy M. (2014) From mobile phone data to the spatial structure of cities. Scientific Reports, 4 : 5276, doi:10.1038/srep05276

Lu X., Bengtsson L., Holme P. (2012) Predictability of population displacement after the 2010 Haiti earthquake, PNAS, Volume 109(29), p 11576–11581, doi: 10.1073/pnas.1203882109

Ma, D., Sandberg, M., Jiang, B., 2015. Characterizing the heterogeneity of the OpenStreetMap data and community. ISPRS International Journal of Geo-Information 4, 535–550. doi:10.3390/ijgi4020535

MacEachren, A. M., A. C. Robinson, A. Jaiswal, S. Pezanov, A. Savelyev, J. Blanford, and P. Mitra (2011, July). Geo-Twitter analytics: Application in crisis management. In 25th International Cartographic Conference.McKenzie, G. and K. Janowicz (2014). Coerced geographic information: The not-so-voluntary side of user-generated geo-content. In Eighth International Conference on Geographic Information Science.

Nergiz, E., Atzori, M., and Saygin, Y. (2008) Towards trajectory anonymization: a generalization-based approach. Proc. of ACM GIS Workshop on Security and Privacy in GIS and LBS.

Niehoefer, B.. Schweikowski F. and Wietfeld C., 2013, Smart Constellation Selection for Precise Vehicle Positioning in Urban Canyons using a Software-Defined Receiver Solution, In proceeding of 20th IEEE Symposium on Communications and Vehicular Technologies, Namur, Belgium, Nov. 2013.

Neis, P., Zielstra, D., 2014. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. Future Internet 6, 76–106. doi:10.3390/fi6010076

Olteanu-Raimond, A.-M., T. Couronné, J. Fen-Chong and Z. Smoreda (2012); Modélisation des trajectoires spatio-temporelles issues des traces numériques de téléphones portables. Le Paris des visiteurs, qu'en disent les téléphones mobiles ?, Revue Internationale de Géomatique, vol. 22, n. 3, pp. 413--437, doi:doi:10.3166/rig.22.413-437

Olteanu-Raimond, A.-M., Bahoken, F., Couronné, T., Smoreda, Z., 2013, Proposition de matrices de flux temporelles issues de l'activité d'individus mobiles, Actes du colloque International d'Analyse Spatiale et de Géomatique (SAGEO'2013), Brest, France, 23-26 septembre 2013.

Olteanu-Raimond, A.-M., Hart, G., Foody, G. M., Touya, G., Kellenberger, T. and Demetriou, D. (2016), The Scale of VGI in Map Production: A Perspective on European National Mapping Agencies. Transaction in GIS. doi:10.1111/tgis.12189

Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA, 2011 Geographic constraints on social network groups, PLoS ONE 6(4): e16939

Quesnot, T. and S. Roche (2015a). Quantifying the significance of semantic landmarks in familiar and unfamiliar environments. In S. I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, and S. Bell (Eds.), Spatial Information Theory, Volume 9368 of Lecture Notes in Computer Science, pp. 468-489. Springer International Publishing.

Quesnot, T. and S. Roche (2015). Platial or locational data? toward the characterization of social location sharing. In System Sciences (HICSS), 2015 48th Hawaii International Conference on, pp. 1973-1982. IEEE.

Reades, J., Calabrese, F., et Ratti, C. Eigenplaces: analysing cities using the space–time structure of the mobile phone network, Environment and Planning B: Planning and Design 36, 5 (2009), 824–836.

Reis, AC & Higham, JES 2009, 'Recreation conflict and sport hunting: moving beyond goal interference towards social sustainability', Journal of Sport and Tourism, vol. 14, no. 2-3, pp. 83-107.

Sakaki, T., Okazaki, M. and Matsuo, Y. 2010: Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web: ACM, 851-860.

Schneider, C. M., Belik, V., Couronne, T., Smoreda, Z., and Gonzalez, M. C. 2013. Unravelling daily human mobility motifs. J. R. Soc. Interface 10, 84

See, L.; Mooney, P.; Foody, G.; Bastin, L.; Comber, A.; Estima, J.; Fritz, S.; Kerle, N.; Jiang, B.; Laakso, M.; Liu, H.-Y.; Milčinski, G.; Nikšič, M.; Painho, M.; Pődör, A.; Olteanu-Raimond, A.-M.; Rutzinger, M. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS Int. J. Geo-Inf. 2016, 5, 55.

Senaratne, H., A. Mobasheri, A. L. Ali, C. Capineri, and M. M. Haklay (2016, May). A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science, 1-29.

Sevtsuk, A. (2008) Explorations into Urban Mobility Patterns Using Aggregate Mobile Network Data. Working paper # TSI-SOTUR-08-03, unpublished, MIT Portugal Program.

Simon G., Pratiques touristiques dans la métropole parisienne. Une analyse des mouvements intra-urbains. Thèse de doctorat, Université de Paris Est, 2010.

Song C., Qu Z, Blumm N, Barabási AL, 2010, Lim-its of Predictability in Human Mobility. Science 327, 1018 (2010); DOI: 10.1126/science.1177170.

Shoval N., Isaacson M., 2006, Tracking tourists in the digital age, Annals of Tourism Research, Vol. 34, No. 1, pp. 141–159, doi:10.1016/j.annals.2006.07.007

Siła-Nowicka K., Vandrol J., Oshan T., Long J.A., Demšar U., Fotheringham A.S., 2016, Analysis of human mobility patterns from GPS trajectories and contextual information, International Journal of Geographical Information Science, Volume 30 (5), p. 881-906, DOI: 10.1080/13658816.2015.1100731

Smoreda Z, Olteanu-Raimond AM, Couronné T (2013) Spatiotemporal data from mobile phones for personal mobility assessment, In Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), Transport Survey Methods: Best Practice for Decision Making, Emerald Group Publishing, London.

Starbird, K. and Palen, L. 2011: Voluntweeters:" Self-Organizing by Digital Volunteers in Times of Crisis. Conference on Computer Human Interaction (CHI 2011), Vancouver, BC, Canada: ACM.

Stefanidis, A., Crooks, A., and Radzikowski, J. 2013. Harvesting ambient geospatial information from social media feeds. GeoJournal, 78 (2), 319 –338

Sweeney, L. (2002) k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty Fuzziness and Knowledge-based Systems 10, 5.

Tatem A.J., Qiu Y., Smith D.L., Sabot O., Ali A.S., Moonen B., 2009. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents, Malaria Journal, 8:287 doi:10.1186/1475-2875-8-287

Terrovitis, M., and Mamoulis, N. (2008) Privacy preservation in the publication of trajectories. Proc. of the 9th Int. Conf. on Mobile Data Management (MDM'08).

Tiru M., Ahas R., 2012, Passive Anonymous Mobile Positioning Data for Tourism Statistics, Actes du 11th Global Forum on Tourism Statistics, November 14-16, 2012, Harpa, Reykjavik, Islande.

Traag, V., Browet, A., Calabrese, F., et Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference. In Privacy, security, risk and trust (passat). IEEE third international conference on on social computing (socialcom). 625 –628.

Trasarti , R., A.-M. Olteanu-Raimond, M. Nanni, T. Couronné, B. Furletti , F. Giannotti , Z. Smoreda and C. Ziemlicki (2015) Discovering urban and country dynamics from mobile phone data with spatial correlation patterns, Telecommunications Policy, vol. 39, n. 3--4, pp. 347–362,doi:10.1016/j.telpol.2013.12.002

Vieira, M.R., Frias-Martinez, V.O.N. et Frias-Martinez, E. (2010) Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics. In Proceedings of Second International Conference of Social Computing (Minneapolis, MN, USA, 2010), pp. 241-248.

Wang, M.H, Schrock, S., Vander Broek, N. and Mulinazzi, T., Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data, International Journal of Intelligent Transportation Systems Research, Vol 11, N°2, 2013, pp. 76-86.

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., and Gonzalez, M. C. (2012). Understanding road usage patterns in urban areas. Scientific Reports 2, Article number: 1001, doi:10.1038/srep01001.

Wesolowski A., Buckee C.O., Bengtsson L., Wetter E., Lu X., Tatem A.J. (2014) Commentary: Containing the Ebola Outbreak - the Potential and Challenge of Mobile Network Data. PLoS Currents, Published online 2014 September 29, doi: 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.

Williams NE, Thomas TA, Dunbar M, Eagle N, Dobra A (2015) Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. PLoS ONE 10(7): e0133630. doi:10.1371/journal.pone.0133630.

Wood, J., A. Slingsby, and J. Dykes (2011). Visualizing the dynamics of london's bicycle hire scheme. Cartographica 46 (4), 239-251.

Yarovoy, R., Bonchi, F., Lakshmanan, L. V. S., and Wang, W. H. (2009) Anonymizing moving objects: How to hide a MOB in a crowd? Proc. of the 12th Int. Conf. on Extending Database Technology (EDBT'09).

Zielstraa D. and Hochmaira H. H. 2013. Positional accuracy analysis of Flickr and Panoramio images for selected world regions, Journal of Spatial Science, vol. 58, issue 2, pp. 251-273

Zook, M., Graham, M., Shelton, T. and Gorman, S. (2010), Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. World Medical & Health Policy, 2: 7–33. doi: 10.2202/1948-4682.1069