Actes de l'atelier GAST – **Gestion et Analyse de données Spatiales et Temporelles**

Cyril De Runz (CReSTIC, Université de Reims Champagne-Ardenne)
Éric Kergosien (GERiiCO, Université Lille 3)
Thomas Guyet (IRISA-Inria/AGROCAMPUS-OUEST)
Christian Sallaberry (LIUPPA, Université de Pau et des Pays de l'Adour)

http://gt-gast.irisa.fr/gast-2017/

Mardi 24 janvier 2017, Grenoble

PRÉFACE

Le troisième atelier "Gestion et Analyse des données Spatiales et Temporelles" (GAST) est associé à EGC'2017. Cet atelier, s'appuyant sur le Groupe de Travail GAST, regroupe des chercheurs, du domaine académique et de l'industrie, qui s'intéressent aux problématiques liées à la prise en compte de l'information temporelle ou spatiale – quantitative ou qualitative – dans leurs processus de gestion et d'analyse de données (méthodes et application de l'extraction, la gestion, la représentation, l'analyse et la visualisation d'informations).

Ces actes regroupent 9 soumissions présentées à l'atelier GAST'2017 :

- Luciano GERVASONI, Martì BOSCH, Serge FENET, Peter STURM, LUM_OSM : une plateforme pour l'évaluation de la mixité urbaine à partir de données participatives ;
- Jérémy LHEZ, Olivier CURé, Profilage sémantique et probabiliste de zones géographiques ;
- Nathalie VALMOND-LEBLANC, Eric KERGOSIEN, Natalia GRABAR, Marta SEVERO, Une approche semi-automatique pour l'extraction d'informations liées aux itinéraires culturels à partir des réseaux sociaux : cas de la Via Francigena ;
- Besma KHALFI, Cyril de RUNZ, Sami FAIZ, Herman AKDAG, Modèles de validation syntaxique et sémantique pour un stockage cohérent de données géospatiales imprécises dans les systèmes NoSQL document ;
- Alain GUEZ, Claire LAGESSE, Francis ROUSSEAUX, Approche pour l'élaboration d'un modèle chronotopique urbain 2 ;
- Françoise BAHOKEN, Vers la représentation cartographique d'une extraction des motifs de flux et mouvements mondiaux ;
- Sylvain BOUVERET, Philippe GENOUD, Kim HOBUS, Danielle ZIEBELIN, The Semantic Evolution of Observation Data: Visualizing spatiotemporal water data ;
- Karim HAMMOUDI, Halim BENHABILES, Mahmoud MELKEMI, Fadi DORNAIKA, Analyse et gestion de l'occupation de places de stationnement par vision artificielle ;
- Thierry PAGÈS, Anne QUESNEL-BARBET, Régis BEUSCAR, Application mobile de visualisation 3D temps réel des compétitions sportives en extérieur (LIS3D) : démonstration.

Ces articles montrent une large étendue des recherches actuelles à des fins de modélisation, d'extraction, d'analyse, ou de visualisation d'information, basées sur les dimensions temporelles et spatiales associées. Nous y trouvons des thématiques et applications aussi différentes que l'exploitation de données participatives, le traitement de flux de données de capteurs et la mise en œuvre de méthodes de profilage, l'extraction d'information spatiale à partir de données issues du réseau social Instagram, le traitement de données géo-spatiales imprécises, le regard pluridisciplinaire (entre anthropologie, urbanisme, paysage et architecture) pour l'analyse de la ville dans le temps, l'analyse (carto)graphique des flux, ou encore, la visualisation de séries de données spatiotemporelles.

Nous espérons que les orateurs, les auditeurs et les lecteurs pourront interagir autour de ces sujets, que les questions et les défis associés à l'information temporelle et spatiale continueront à animer les débats. Nous tenons à remercier tous les auteurs pour leurs propositions d'articles ainsi que les membres du comité de lecture dont les retours ont été de qualité pour l'ensemble des articles. Nous remercions également chaleureusement Mathieu Roche, chargé de recherche au CIRAD, pour son intervention en tant que conférencier invitée à la journée GAST'2017. En espérant que ces articles vous apporteront de nouvelles perspectives "Gestion et Analyse des données Spatiales et Temporelles", nous vous souhaitons une bonne lecture.

Cette année, l'atelier a été organisé en trois temps : une présentation invitée de Mathieu Roche (UMR TETIS, Montpellier) sur le thème "Analyse des dynamiques spatiales et thématiques par des méthodes de fouille de textes", puis un ensemble de présentations orales des articles retenus pour l'atelier et finalement un temps dédié à la discussion avec l'ensemble des participants. Nous espérons que le lecteur qui n'a pu y assister trouvera toutes les informations dans les articles de ce volume.

Eric Kergosien
Université Lille-3/GERiiCO

Thomas Guyet
Agrocampus-Ouest/IRISA-Inria

Cyril De Runz
Université de Reims
Champagne-Ardenne/CReSTIC

Christian Sallaberry
Université de Pau/LIUPPA

Membres du comité de lecture

Rodéric Béra, Agrocampus-Ouest
Sandro Bimonte, IRSTEA
Géraldine Del Mondo, INSA Rouen
Cyril De Runz, Université de Reims
Thomas Devogele, Université de Tour
Frédéric Flouvat, Université Nouvelle Calédonie
Serge Guillaume, IRSTEA
Thomas Guyet, Agrocampus-Ouest
Eric Kergosien, Université Lille 3
Christine Plumejeaud, Université de La Rochelle
Florence Le Ber, ENGEES
Simon Malinowski, Université Rennes 1
Jean-François Mari, INRA
Nicolas Merger, Université de Savoie
Marta Severo, Université Lille 3
Maguelonne Tesseire, IRSTEA

TABLE DES MATIÈRES

Présentation invitée

Analyse des dynamiques spatiales et thématiques par des méthodes de fouille de textes
Mathieu Roche 1

Articles de l'atelier

LUM_OSM : une plateforme pour l'évaluation de la mixité urbaine à partir de données participatives
Luciano Gervasoni, Martì Bosch, Serge Fenet, Peter Sturm 3

Profilage sémantique et probabiliste de zones géographiques
Jérémy Lhuez, Olivier Curé 25

Une approche semi-automatique pour l'extraction d'informations liées aux itinéraires culturels à partir des réseaux sociaux : cas de la Via Francigena
Nathalie Valmond-Leblanc, Eric Kergosien, Natalia Grabar, Marta Severo 37

Modèles de validation syntaxique et sémantique pour un stockage cohérent de données géo-spatiales imprécises dans les systèmes NoSQL document
Besma Khalfi, Cyril de Runz, Sami Faiz, Herman Akdag 43

Approche pour l'élaboration d'un modèle chronotopique urbain 2
Alain Guez, Claire Lagesse, Francis Rousseaux 51

Vers la représentation cartographique d'une extraction des motifs de flux et mouvements mondiaux
Françoise Bahoken 67

The Semantic Evolution of Observation Data: Visualizing spatiotemporal water data
Sylvain Bouveret, Philippe Genoud, Kim Hobus, Danielle Ziebelin 79

Analyse et gestion de l'occupation de places de stationnement par vision artificielle
Karim Hammoudi, Halim Benhabiles, Mahmoud Melkemi, Fadi Dornaika 91

Application mobile de visualisation 3D temps réel des compétitions sportives en extérieur (LIS3D) : démonstration
Thierry Pagès, Anne Quesnel-Barbet, Régis Beuscar 99

Index des auteurs **104**

Analyse des dynamiques spatiales et thématiques par des méthodes de fouille de textes

Mathieu Roche*

* CIRAD, Montpellier

Résumé

Les dynamiques spatiales et thématiques font l'objet d'un intérêt croissant notamment dans les milieux à enjeux et de tension (épidémiologie, migrations humaines, ...). L'objectif de nos travaux est de proposer un cadre méthodologique permettant l'appréhension de ces phénomènes à partir de données textuelles hétérogènes (récits, articles, réseaux sociaux, etc.). Les analyses produites s'appuient, en général, sur trois types d'informations (thématiques, spatiales et temporelles) qui sont extraites et exploitées par des méthodes de fouille de données et de traitement automatique du langage naturel (TALN). Des cas d'études concrets seront présentés à travers les problématiques issues du projet Mastodons QDoSSI et des travaux liés à la veille en épidémiologie animale (lien web).

`LUM_OSM` : une plateforme pour l'évaluation de la mixité urbaine à partir de données participatives

Luciano GERVASONI[*,**], Martì BOSCH[***]
Serge FENET[*,****] Peter STURM[*,**]

[*]Inria Grenoble – Rhône-Alpes, France
[**]Univ. Grenoble Alpes, Lab. Jean Kuntzmann, Grenoble, France
[***]École Polytechnique Fédérale de Lausanne, Lausanne, Suisse
[****]Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

Résumé. L'augmentation de plus en plus rapide des populations urbaines depuis les dernières décennies, en conjonction avec l'accessibilité croissante de données urbaines massives, motivent de plus en plus de recherches portant sur la soutenabilité du développement urbain.

Nous présentons dans cet article une plateforme d'analyse et de construction de cartes de mixité urbaine afin d'aider les planificateurs urbains à évaluer des indices de mixité. Dans ce premier travail, nous nous focalisons sur les usages résidentiels et d'activités, qui sont extraits des données collaboratives d'OpenStreetMap. Ils servent à la construction d'une représentation continue de l'indice d'entropie, mais peuvent être généralisés à tout indice. La méthode est appliquée à plusieurs villes, et les premiers résultats sont présentés.

1 Introduction

La population vivant dans des centres urbains a augmenté de façon considérable depuis les années 50, passant de 746 millions à 3,9 milliards en 2014 (United Nations (2014)). Les projections actuelles prévoient que plus de 66% de la population mondiale vivra dans des villes d'ici 2050, contre 30% en 1950. En parallèle avec un léger déclin de la population rurale globale dans les 35 prochaines années, l'augmentation de l'urbanisation devrait ajouter près de 2,5 milliards de personnes dans les centres urbains d'ici 2050. Cette situation, qui impose de nouveaux défis liés à la conception de villes capables d'héberger de telles populations de manière soutenable, nous force à penser les villes du futur sous des aspects économiques, sociaux et environnementaux.

Nous nous concentrons dans cet article sur la formalisation de la mesure de développement mixte et de mixité urbaine (LUM pour *land use mix*) dans les villes, afin notamment d'étudier son impact sur la soutenabilité des villes. La mixité urbaine décrit la coexistence dans un rayon localisé de différents usages du sol urbain (habitations, lieux de travail, lieux de consommation et de loisirs, etc.). Elle est donc associée à une certaine compacité, mais aussi à des aménagements rendant la vie dans les villes plus agréables et permettant de limiter la consommation de carburants fossiles (de par la présence de parcs, de jardins ouvriers ou de pistes cyclables, par exemple).

La soutenabilité urbaine décrit la capacité de la structure urbaine à maintenir son intégrité dans le temps sans épuiser les ressources qui lui permettent d'exister, tout en assurant un service de qualité aux habitants. Il a déjà été prouvé qu'une forte mixité urbaine apporte de nombreux bénéfices en termes de soutenabilité, et les structures urbaines denses et compactes ont été largement étudiées dans la littérature en terme de soutenabilité de développement (Gordon et Richardson (1997); Breheny (1995); Burgess (2000); Neuman (2005)). Bien que le degré nécessaire de compacité reste une question ouverte, un accord large existe sur les impacts positifs de la mixité urbaine, non seulement en termes de soutenabilité de développement, mais aussi en termes sociétaux, de santé publique et d'infrastructure de transports publics.

Afin d'aider les planificateurs urbains, nous proposons une plateforme logicielle permettant de capturer l'état courant de mixité urbaine d'une ville à l'aide de données collaboratives publiques. Cet outil permet de construire une carte de type SIG (système d'information géographique) contenant le degré de mixité de tout point de la ville, avec une résolution arbitraire. L'usage des sols associé à chaque polygone est d'abord extrait des données OpenStreetMap en entrée. Ensuite une représentation spatiale de chaque usage est calculée à l'aide de noyaux statistiques. Enfin, ces représentations génériques sont utilisées pour calculer l'indice final. Même s'il est souvent désirable d'atteindre un degré de mixité le plus élevé possible, nous n'effectuons pas ici d'interprétation numérique de la valeur optimale nécessaire à un développement soutenable. Nous nous focalisons plutôt sur la méthode de calcul d'une représentation spatialisée continue de la carte de mixité, et permettant de calculer n'importe quel indice (nous l'appliquons ici à l'indice d'entropie) pour peu qu'il soit basé sur des données présentes dans la base OpenStreetMap.

Cet article est organisé de la manière suivante. Nous présentons dans la section 2 une vue d'ensemble du contexte de ces travaux. La section 3 présente une rapide analyse des indicateurs de mixité existants. Nous argumentons dans la section 4 l'importance des données collaboratives urbaines dans le cadre des études de mixité, et nous présentons notre plateforme dans la section 5. Finalement, nous illustrons dans la section 6 son application à plusieurs villes, avant de conclure en présentant les travaux futurs envisagés.

2 Contexte

2.1 Motivation

Un motif migratoire évident des zones rurales vers des centres urbains de taille croissante peut être observé dans l'histoire récente, et ce phénomène devient de plus en plus important depuis la fin du 19ème siècle. Ce processus, présent de plus dans tous les pays, a impliqué une augmentation constante de la taille des populations urbaines : plus de la moitié de la population humaine vit maintenant dans les villes, et ces dernières font face à des problèmes croissants pour continuer à assurer aux habitants les services essentiels et une bonne qualité de vie. Cette concentration impose donc de nouvelles façons de concevoir les villes, ces systèmes complexes dynamiques avec de nombreux niveaux d'interaction, de manière à assurer à la fois leur soutenabilité et une consommation efficace des ressources.

Cependant, plus récemment, l'accessibilité grandissante de données urbaines parfois massives rend possible l'analyse des interactions entre les habitants et l'infrastructure de leur ville. Ainsi, la compréhension des boucles de rétroaction complexes liant les citoyens et le

développement de leur cité pourrait nous permettre de résoudre plus efficacement les questions de soutenabilité future. En conséquence, des efforts considérables et récents visent à comprendre les conditions de la soutenabilité urbaine en se basant sur ces nouvelles sources de données.

Dans ce contexte, le présent travail a été motivé par trois observations :

- Tout d'abord, et alors que les questions de soutenabilité deviennent un sujet central dans plusieurs domaines scientifiques, nous pensons que les outils numériques, et plus spécifiquement la construction de modèles numériques à partir de processus de fouille de données issues du monde réel, peuvent aider à mieux comprendre la dynamique des systèmes socio-économiques complexes. Nous espérons que ces outils permettront de rendre plus efficaces les décisions prises pour améliorer la soutenabilité et la résilience des villes du future.
- Ensuite, plusieurs indicateurs pour mesurer la mixité urbaine existent (voir Song et al. (2013)). Ils peuvent tous se catégoriser en mesures divisionnelles ou intégrales ; les premières étant plus expressives que les secondes en délivrant une valeur pour chaque zone considérée. Cependant, ils sont tous sensibles à l'échelle d'agrégation utilisée pour calculer l'indicateur. Nous pensons qu'un bon indicateur devrait être indépendant de l'échelle de l'analyse.
- Enfin, même si les données participatives voient leur quantité et leur qualité croître régulièrement, il y aura toujours des données manquantes pouvant impacter les mesures sus-citées. En conséquence, il est important d'utiliser des outils statistiques permettant de les prendre en compte afin de mieux capturer les structures urbaines sous-jacentes.

2.2 Brefs rappels historiques

Avant le 20ème siècle, les usages urbains mixtes étaient une tendance naturelle étant données les contraintes de proximité géographique des activités journalières imposées par la rareté des possibilités de transport. À partir du début du 20ème siècle, d'abord aux États-Unis, les pratiques de zonage commencèrent à assigner des usages uniques aux différentes parcelles, induisant une ségrégation nouvelle entre les zones d'habitation et les zones d'activités, et imposant un usage intensif de la voiture pour se déplacer. Cette tendance est devenue particulièrement forte des années 1910 aux années 1950, où les zones mixtes étaient devenues rares, et où le développement ségrégé était la norme. Quelques années plus tard, à partir des années 1960 à 1970, et après avoir été négligés pendant des décennies, les usages mixtes réapparurent, apportant de nouveau leurs avantages dans plusieurs aspects de la société.

Plus récemment, depuis le début des années 1990, le développement mixte a ré-émergé en tant que concept essentiel de la planification urbaine. À titre d'exemple, le *Congress for the New Urbanism* (Leccese et McCormick (2000), `http://www.cnu.org`) milite pour des voisinages "pedestrian-friendly, and mixed-use" et le *Smart Growth Network* (Smart Growth Network (2006)) inclut les usages urbains mixtes dans ses 10 principes. Le concept de *Transit-Oriented Development* et plusieurs grandes agences de transport supportent de leur côté les usages mixtes (Cervero et al. (2002)). Afin d'atteindre de hauts niveaux de développements mixtes, tous ces acteurs recommandent à la fois l'intensité et la diversité des usages des sols urbains, ainsi que l'intégration des zones auparavant ségrégées.

2.3 Les impacts des usages urbains mixtes

Dans cette section, nous décrivons rapidement les principaux effets liés à la soutenabilité des villes en présence des différents type de développements. Les relations entre soutenabilité et mixité des usages, qu'elles soient directes ou indirectes, ont été étudiées de manière intensive dans la littérature :

- *Étalement urbain :* Le concept d'étalement urbain fait référence au processus d'extension spatiale de la population s'éloignant des zones centrales pour former des communautés satellites éparses, mono-fonctionnelles, et souvent très dépendantes de la voiture. La sub-urbanisation, c'est-à-dire la création de zones de banlieues éloignées des centre-villes, s'est accélérée durant tout le dernier siècle. En conséquence, une faible mixité urbaine est une des caractéristiques distinctives associées au phénomène d'étalement urbain, qu'il soit résidentiel ou industriel.
  Dans la littérature, l'étalement urbain a été largement associé à des effets négatifs en termes d'environnement, de santé, de relations sociales et d'économie (Squires (2002)). Plus récemment, et en conjonction avec l'apparition du concept de développement soutenable, les planificateurs urbains ont commencé à prendre de plus en plus en compte les conséquences de l'étalement urbain (Johnson (2001)), même si ces conséquences négatives avaient déjà été identifiées au début des années 1970.
  Le manque de planification coordonnée de l'usage de sols et ses conséquences négatives ont commencé à clairement apparaître comme des composants clés de l'étalement urbain (Nelson et Duncan (1995); Ewing (1995); Sierra Club (1998); United States Department of Housing and Urban Development (1999)).
- *Transport :* En termes de transports publics, les zones étalées ne sont pas soutenables lorsqu'elles sont associées à des indices de mixité faibles (Cervero et al. (2002)). En effet, des zones étalées et largement dispersées imposent de se déplacer sur de grandes distances pour des faibles demandes, abaissant l'efficacité des infrastructures de transport.
- *Santé :* Il a été montré que la mixité d'usages a un impact fort sur les activités physiques des habitants (Christian et al. (2011)). Cela implique que la conception *a priori* de la structure urbaine peut directement impacter l'activité et le bien-être physique des habitants, ce qui constitue actuellement un point important lié à la santé des habitants urbains : un taux de mixité élevé est associé à des déplacements à pied plus importants (Christian et al. (2011)) et implique une meilleure santé et une baisse de l'obésité (Frank et al. (2004)).
- *Dépendance à la voiture :* Le type de déplacement, en considérant le mode et la distance à parcourir, a été fortement corrélé au type de mixité urbaine dans Cervero (1996); Kockelman (1997). En effet, la présence de zones à usages commerciaux proches a été associée à la fois à des temps de parcours raccourcis et à une faible nécessité de posséder un véhicule personnel (Cervero (1996)), ce qui implique un impact environnemental positif en termes d'émission de polluants. Les études de Newman et Kenworthy (2006) sur la conception de structures urbaines réduisant la dépendance à la voiture ont déterminé qu'une dépendance plus basse implique d'avoir une valeur de densité urbaine (i.e. nombre d'emplois et de résidences par hectare) au delà d'un seuil minimal, ce qui est directement relié à la notion de mixité urbaine élevée.

3 Les limites des indicateurs existants

Les propositions de méthodes pour mesurer la mixité urbaine abondent dans la littérature. Une revue presque exhaustive (Song et al. (2013)) examine notamment quelles sont les forces et les faiblesses de chacune. La pertinence d'une mesure de mixité donnée repose directement sur deux aspects. D'abord, elle dépend du contexte d'application, dans lequel l'usage prévu pilote fortement les caractéristiques de la mesure. Ensuite, les données utilisées pour la calculer conditionnent directement sa performance : la présence ou l'absence de bruit, le biais impliqué par la source de données, leur quantité, leur ancienneté, sont autant d'aspects à prendre en compte.

Les mesures de mixité urbaine sont largement inspirées des métriques d'écologie du paysage et des statistiques spatiales (Raines (2002)). Plusieurs indices, tels que l'indice d'Atkinson, l'indice de partitionnement, l'indice de dissimilarité, l'indice d'exposition ou l'indice de Gini ont été proposés. Par ailleurs, plusieurs métriques sont basées sur la définition de l'entropie de Shannon.

Les données utilisées pour calculer ces indices sont rarement discutées. Elles sont généralement obtenues lors de l'étude d'une certaine région d'intérêt, et sont donc souvent fortement biaisées par la disponibilité des données relatives à cette région. D'un côté, de nombreux jeux de données sont privés et réduisent fortement la reproductibilité des résultats. D'un autre, certains jeux de données spécifiques existent, mais sont bien souvent restreints à certaines zones géographiques [1]. Il est bien connu que le résultat de toute méthode numérique est fonction des données en entrée, et l'utilisation de données hétérogènes liées aux multiples zones d'application ont pu favoriser le développement de mesures *ad hoc* qui exhibent un comportement désiré pour le jeu de données en entrée, tout en limitant l'application possible à cette seule région d'analyse.

De nombreux travaux de la littérature mettent en avant le problème de l'unité de surface modifiable ("modifiable areal unit problem" ou MAUP), qui se pose naturellement lorsqu'on agrège des données spatiales. Les sorties de la mesure dépendent du niveau d'agrégation et de l'unité de division géographique choisie. La sensibilité des résultats à ces choix constitue un gros inconvénient, indésirable dans toutes les applications, et se révèle être un vrai trouble lorsqu'il s'agit de comparer plusieurs métriques. Dans notre approche, l'utilisation de données géo-localisées, couplées à l'estimation d'une surface continue en utilisant une estimation de noyau de densité permet de contourner le problème du MAUP (Carlos et al. (2010)).

Le présent travail ne cherche pas a définir un nouvel indicateur qui serait meilleur que tous les indicateurs existants, mais propose une méthode et une plateforme pour en calculer un automatiquement, à une résolution spatiale quelconque, et à partir de données publiques. Nous évaluons notre outil en calculant une représentation SIG de l'indice d'entropie. Cela nous permet notamment de montrer qu'il est possible de dépasser la limitation séparant les indices intégraux et divisionnels (nous calculons la version géographiquement continue d'un indicateur divisionnel). Muni d'un tel outil, nous pourrons dans un second temps nous concentrer sur le calcul d'un indicateur optimal à l'aide de méthodes d'apprentissage supervisé, et comparer ce dernier aux indicateurs existants.

1. Bien que l'émergence des données nationales ouvertes (Open Data) soit en train de changer cela, tout au moins en Europe.

4 Données utilisées

OpenStreetMap [2] (OSM) est un projet collaboratif portant sur la création d'une carte éditable et ouverte du monde ; il s'agit d'un exemple de "volunteered geographic information" (VGI). OSM fournit, entre autres, des cartes routières générées par ses utilisateurs (Haklay et Weber (2008)). Des bénévoles du monde entier partagent des informations géographiques au sein d'OSM et sont parfois considérés comme des "capteurs intelligents" (Goodchild (2007)).

Depuis sa création, ce projet a été utilisé de manière de plus en plus intense pour une grande variété de tâches. Plusieurs métriques portant sur la qualité de ses données ont été proposées dans Forghani et Delavar (2014); Barron et al. (2014); Mooney et al. (2010); Fan et al. (2014), ainsi que des études sur la qualité par pays. Ainsi, OSM a été jugé comme "plutôt précis" pour l'Angleterre (Haklay (2010)) et comme supérieur aux données officielles pour le périmètre "Meridian 2" du Royaume-Uni. Ces études ont été étendues au cas de la France (Girres et Touya (2010)). En ce qui concerne l'Allemagne, le réseau routier d'OSM est considéré comme plus complet que pour des alternatives commerciales (Neis et al. (2011)). Par exemple, 99,8% du réseau routier de Hambourg sont couverts par OSM, selon le bureau de planification de la ville (Over et al. (2010)). Cette étude soulève également que "Outre le réseau routier, le réel avantage de cette base de données est la disponibilité de divers points d'intérêt". Ces derniers, contenant par exemple les arrêts de transports en communs ou les stations d'essence, permettent une meilleure compréhension des dynamiques urbaines, grâce à leur géo-localisation et aux diverses informations qui leur sont associées. En Chine, le volume des points d'intérêt a augmenté d'un facteur 9 au cours de la période 2007-2013 (Liu et Long (2016)).

Dans ce travail, nous utilisons des données OSM pour analyser la mixité des usages des sols au sein de villes. Utiliser cette base de données comporte plusieurs avantages. Premièrement, sa couverture est mondiale et elle permet déjà la comparaison de nombreuses villes sur la base d'une source de données homogène. Deuxièmement, OSM est mise à jour continuellement sur le principe de "crowd-sourcing", ce qui rend possible l'adaptation des informations aux changements rapides que connaissent les villes de nos jours. De plus, la communauté très active de ses utilisateurs améliore constamment la précision et la complétude des données. Troisièmement, il s'agit de données ouvertes. Ceci permet de corriger les erreurs de saisie de manière croisée, et surtout d'établir une cohérence entre différentes contributions en planification urbaine. Ainsi, les résultats de recherche deviennent reproductibles pour tout un chacun, ce qui est un facteur important pour l'évaluation et l'amélioration des outils proposés. Finalement, la présence de données géo-localisées permet une analyse spatiale plus fine qu'avec des données localisées sur des grilles, par exemple.

Actuellement, la principale faiblesse d'OSM réside dans la couverture encore imparfaite de quelques régions du monde. Cependant, surtout pour les grandes villes, ce problème est réduit grâce à la présence de nombreux contributeurs. En comparaison, des données de type "LULC" (Land Use Land Cover), typiquement utilisées dans ce genre d'études, ont plusieurs inconvénients. Tout d'abord, leur disponibilité est en général restreinte à des pays individuels, voire à certaines zones géographiques. Ensuite, même si leur résolution a été considérablement accrue, ces données ne contiennent pas d'informations sur l'intensité des usages des sols répertoriés (activités, résidentiel, etc.). De plus, la classification des usages des sols dans une seule classe d'usage par entité spatiale, ainsi que le fait que les catégories d'usages sont typique-

2. http://www.openstreetmap.org/

ment très agrégées, font que ces données sont insuffisantes pour analyser la mixité des usages des sols dans des villes (Christian et al. (2011)). Enfin, les bases de données LULC peuvent périmer rapidement à cause du rythme intense des dynamiques urbaines.

5 Approche proposée

Dans cette section, nous décrivons la chaîne de traitements qui nous permet de calculer une représentation spatialement continue de la mixité des usages des sols, à partir de données OSM brutes.

5.1 Extraction de données

Nous expliquons ici comment extraire les différents usages des sols qui seront ensuite analysés pour déterminer leur degré de mixité. Les données OSM d'un territoire donné, défini par ses contours, sont récupérées au format *shapefile* à travers le site *Mapzen Metro Extracts* [3]. Ces fichiers sont le résultat du processus *osm2pgsql* qui convertit des données OSM en bases de données PostgreSQL/PostGIS [4]. Des polygones et des points d'intérêt sont extraits de ces fichiers et utilisés comme entrées pour la suite des traitements.

Nous procédons d'abord à une classification des points d'intérêt (POI) et des polygones en usages de type "résidentiel" et "activité", en suivant le Wiki d'OSM [5]. Les points et polygones sont associés à des usages des sols à travers les informations qu'ils contiennent (clé, valeur), selon le Tableau 1.

Dans la base OSM, un grand nombre de polygones sont actuellement étiquetés comme "building", sans information supplémentaire. Dans ce cas, l'usage des sols d'un polygone est estimé selon la procédure détaillée dans le Tableau 2. D'abord, les polygones contenant une clé associée sans ambiguïté à un type d'usage des sols (résidentiel, activité ou autre, par exemple forêt, eau), sont classifiés en conséquence. Soient P_R, P_A et P_O les ensembles des polygones associés à ces trois types d'usage. Quant un polygone P possède "building" comme unique clé, l'usage est estimé comme suit :

$$LU(P) = LU(p) \quad if\ \exists p \text{ with } \min_{p \in P_R \cup P_A \cup P_O} A(p)\ ,\ P \subseteq p$$

où $A(\cdot)$ désigne l'aire d'un polygone et $LU(\cdot)$ son usage des sols (*land use*). Remarquons que des polygones associés à différents usages, peuvent se superposer. Nous utilisons alors l'information contenue dans le plus petit polygone qui englobe P et qui a un usage des sols connu.

Il arrive parfois que P ne soit englobé par aucun polygone dont l'usage des sols est connu. Dans ce cas, nous faisons l'hypothèse que P est de type "résidentiel". Ce choix *a priori* purement *ad hoc* est adopté pour contrebalancer le fait qu'OSM contient actuellement plus d'annotations de type "activité"" que de type "résidentiel". En effet, lorsque des utilisateurs définissent des polygones qui contiennent une activité, cette dernière est habituellement annotée, ce qui n'est pas forcément le cas avec les usages résidentiels. Néanmoins, une meilleure complétude

3. http://mapzen.com/data/metro-extracts/
4. http://postgis.net
5. http://wiki.openstreetmap.org/

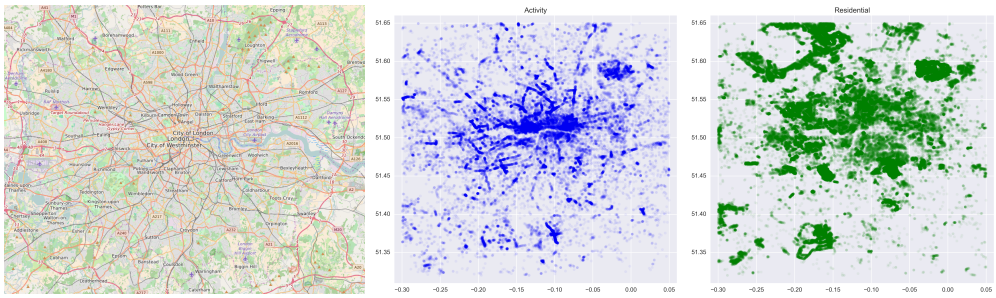
Clé	Valeurs
	(a) Activités
amenity	bar, pub, restaurant, biergarten, cafe, fast_food, food_court, ice_cream, pub, restaurant, college, kindergarten, library, public_bookcase, school, music_school, driving_school, language_school, university, fuel, bicycle_rental, bus_station, car_rental, taxi, car_wash, ferry_terminal, atm, bank, bureau_de_change, baby_hatch, clinic, dentist, doctors, hospital, nursing_home, pharmacy, social_facility, veterinary, arts_centre, brothel, casino, cinema, community_centre, fountain, gambling, nightclub, planetarium, social_centre, stripclub, studio, swingerclub, theatre, animal_boarding, animal_shelter, courthouse, coworking_space, crematorium, dive_centre, dojo, embassy, fire_station, gym, internet_cafe, marketplace, police, post_office, townhall
shop	*
building	commercial, office, industrial, retail, warehouse, cathedral, chapel, church, mosque, temple, synagogue, shrine, civic, hospital, school, stadium, train_station, transportation, university, public, kiosk, garage, garages, hangar, stable, cowshed, digester
leisure	adult_gaming_centre, amusement_arcade, beach_resort, dance, hackerspace, ice_rink, pitch, sports_centre, stadium, summer_camp, swimming_area, water_park
landuse	commercial, industrial, retail, port, quarry, salt_pond, construction, military, garages
	(b) Résidentiel
building	hotel, farm, apartment, apartments, dormitory, house, residential, retirement_home, terrace, houseboat, bungalow, static_caravan, detached
	(c) Autres
landuse	cemetery, landfill, railway, water, reservoir, basin, allotments, conservation, farmland, farmyard, forest, grass, greenfield, greenhouse_horticulture, meadow, orchard, pasture, peat_cutting, plant_nursery, recreation_ground, village_green, vineyard

TAB. 1: Classification des usages des sols en activité, résidentiel et autre, selon les données clé et valeur fournies par OSM.

des annotations résidentielles permettra de simplement abandonner cette hypothèse à l'avenir

Paires Clé=*Valeur*			Classification
leisure = *Activity*			Activité
amenity = *Activity*			Activité
shop = *Activity*			Activité
building = *Activity*			Activité
building = *Residential*			Résidentiel
building = *Yes*	*et*	landuse = *Activity*	Activité
	et	landuse = *Residential*	Résidentiel
	et	landuse = *Null*	à inférer
landuse = *Residential*	*et*	building = *Null*	Inférence : Résidentiel
landuse = *Activity*	*et*	building = *Null*	Inférence : Activité
landuse = *Other*	*et*	building = *Null*	Inférence : nul

TAB. 2: *Look up table* de la procédure d'assignation à une classe d'usage des sols à partir de données OSM.



(a) (b) (c)

FIG. 1: (a) OpenStreetMap de Londres, Angleterre. (b) et (c) montrent les lieux des usages des sols extraits pour, respectivement, les activités et le résidentiel.

si la qualité des annotations continue à augmenter.

Tous les polygones résidentiels avec une aire inférieure à $12m^2$ sont éliminés, comme il s'agit probablement de faux positifs plutôt que de bâtiments résidentiels.

Finalement, les polygones retenus sont convertis en points d'intérêt (POI) en utilisant comme position géographique le lieu de leurs centroïdes. L'objectif est d'obtenir une couverture de la ville par des points géo-référencés contenant des informations d'usage des sols.

Les résultats de cette procédure sont illustrés sur la figure 1 pour le cas de Londres, Angleterre, en utilisant de la transparence partielle pour permettre de distinguer clairement l'information dans les zones hautement concentrées.

5.2 Estimation de la densité par noyau

Nous décrivons dans cette section le processus d'estimation de la densité locale pour les deux usages d'habitation et d'activité. Nous utilisons pour cela le processus statistique d'estimation par noyau (KDE).

Comme montré dans Silverman (1986), le KDE est une technique reconnue pour l'analyse et la représentation de données complexes à des fins de représentations graphiques claires pour des non-mathématiciens.

Le KDE infère une fonction de densité f, sachant un échantillonnage indépendant et identiquement distribué $x_1, x_2, ..., x_n$ de la distribution de probabilité correspondante. L'estimateur du noyau de densité est défini par

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{h}\right), \tag{1}$$

$K(\bullet)$ étant le noyau, et h un paramètre de lissage dénommé bande passante. Dans notre cas, les x_i sont des vecteurs bi-dimentionnels contenant les coordonnées de longitude et latitude des points d'intérêt.

Étant donné un ensemble de points, le KDE interpole une surface continue en utilisant le noyau défini (e.g. une fonction Gaussienne/normale). Cette procédure présente les deux avantages principaux de lissage et d'interpolation spatiale, qui sont des caractéristiques pertinentes dans notre contexte où des données peuvent être manquantes. Ainsi, les centres d'activités et les centres résidentiels ponctuels servent à construire une surface continue permettant de remplacer de manière réaliste les données manquantes.

Comme décrit ci-dessus, le KDE est contrôlé par la fonction de noyau K et la bande passante h. Pour le noyau, nous pré-supposons dans ce travail que les usage résidentiels et d'activités ont tous une influence spatiale normalement distribuée. Ainsi, l'impact d'une activité est maximale à sa localisation d'origine et réduit avec la distance en suivant une distribution normale.

La bande passante est choisie pour représenter une "distance d'influence moyenne". Il faut mentionner ici que plusieurs méthodes de calcul automatique de bande passante existent dans la littérature, mais leur but est de réduire la complexité durant la phase d'estimation aux dépens de la précision. Dans notre cas, étant donné que nous cherchons à capturer l'influence des usages d'habitation et d'activité au sein de leur voisinage géographique, nous identifions la bande passante à la distance moyenne de déplacement piéton. Il est montré dans Atash (1994) que la distance moyenne à laquelle un Américain préférera prendre sa voiture plutôt que de marcher est de $400m$. Selon Aultman-Hall et al. (1997), cette distance est aussi considérée comme la plus grande distance moyenne qu'un voyageur en transit entre des stations de métro ou de bus aura à parcourir. De manière similaire (Western Australian Planning Commission (2007)) propose une distance comprise entre $400m$ et $500m$. Certaines études proposent cependant des distances moyennes de déplacement à pied plus grandes, comme par exemple dans McCormack et al. (2008), qui étudient l'association entre activité physique et le mélange de destinations comprises entre $400m$ et $1500m$ des maisons des résidents. Dans notre contexte, et dans le but de favoriser des voisinages avec une grande mixité d'usages, nous avons décidé d'adopter la valeur suggérée de $400m$, qui sera donc la bande passante de notre procédure. Il faut toutefois garder à l'esprit que cette valeur devrait si possible être adaptée aux pays ou villes étudiés si des travaux pré-existants permettent de la préciser plus avant.

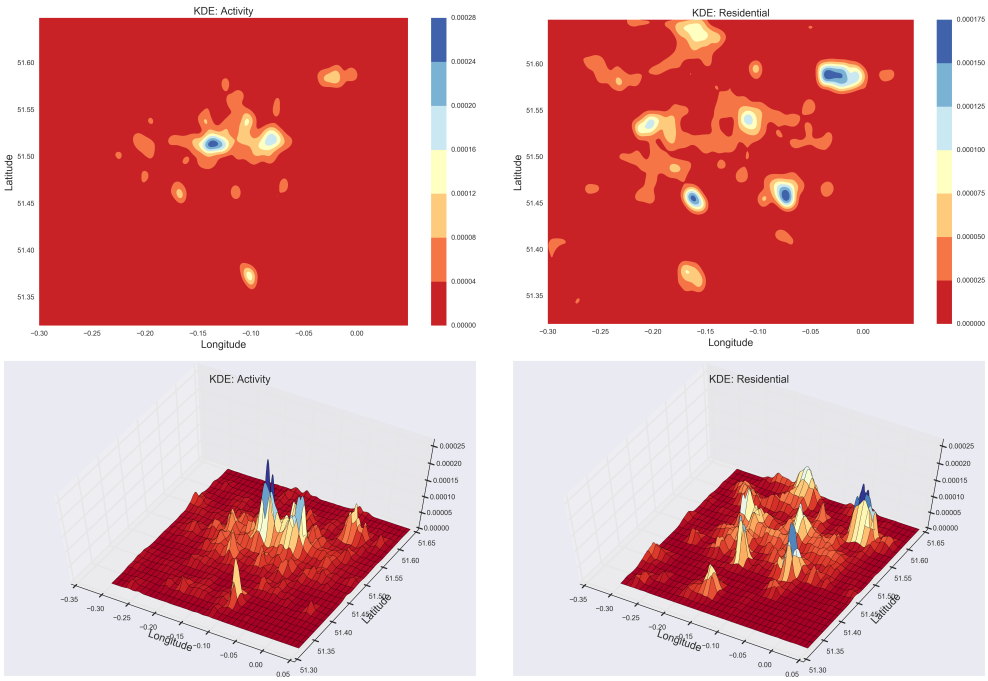


FIG. 2: KDE estimés pour Londres. A gauche, activités et à droite, résidentiel.

Comme mentionné ci-dessus, le KDE est un outil très efficace pour effectuer du lissage spatial et interpoler des données manquantes. Cependant, dans des cas limites tels que des frontières géographiques strictes (régions côtières, coupures franches, etc.) il impliquera un sur-lissage et l'estimation d'informations non existantes. Nous étudierons les limites de tels cas dans des travaux futurs.

Comme dans les travaux de Song et al. (2013), nous avons choisi de nous intéresser à deux types d'usages : les usages résidentiels et les usages d'activités (qui regroupent tous les usages non résidentiels tels que les magasins, les lieux publics, les lieux de travail, etc.). Par conséquent, tous les POI géo-localisés de la procédure ci-dessus sont utilisés pour calculer les représentations de ces deux catégories [6]. La fonction de densité de probabilité est ensuite évaluée le long d'une grille de points couvrant la région d'intérêt avec une résolution choisie par l'utilisateur. Dans les résultats présentés dans cet article, et étant donnée l'échelle des régions d'intérêt des villes analysées et le niveau de détail que nous recherchons dans les zones de centre ville, nous avons choisi une résolution de $100m$.

Enfin, pour terminer, les points calculés sont normalisés pour chaque KDE.

Les KDE résultants sont décrits dans la figure 2, appliqués à la ville Anglaise de Londres.

6. Il s'agit là d'une première étape de nos travaux, rien n'empêchant de détailler les usages dans des classes plus précises en fonction de la précision nécessaire pour le calcul de l'indice final.

5.3 Calcul de la mixité des usages des sols

Dans cette section, nous nous focalisons sur le calcul d'une mesure de la mixité des usages des sols, en partant des cartes de densité de type "activité"" et "résidentiel", estimées avec la méthode décrite ci-dessus. L'objectif est de décrire à quel point ces usages sont bien répartis dans une ville et de comprendre leur degré de co-localisation au sein de voisinages locaux. En règle générale, une bonne mixité locale est désirable, pour les raisons énoncées en début de l'article.

Selon Song et al. (2013), les mesures de mixité d'usages des sols doivent prendre en compte deux concepts sous-jacents : la *distance* et la *quantité*. La distance se réfère à la proximité entre des usages différents. Nous calculons cette proximité (nous parlons aussi de l'influence d'un usage sur un autre) en prenant en compte la distance usuelle de trajets pédestres (voir la section 5.2). En ce qui concerne la quantité, il s'agit pour nous de la sortie du calcul des densités tel que décrit dans la section 5.2.

En accord avec ces notions de distance et quantité, une mesure de mixité doit donc mettre en évidence des régions où différents usages sont présents en quantités comparables et à proximité les uns des autres.

Comme il a été mentionné plus haut, notre plateforme a vocation à être une aide pour des planificateurs urbains, et la possibilité de visualisation géo-spatiale de la mixité calculée est donc un point important. Dans ce contexte, il est également conçu pour mettre en évidence l'importance des différentes sous-régions au sein d'une ville : des grilles maillées calculées à partir des résultats d'estimation de densité par noyau permettent donc de déterminer l'importance de ces sous-régions en terme de l'intensité de leur usage des sols.

Nos premiers résultats sont visualisés à l'aide de "bubble plots", voir par exemple la figure 3. L'idée est de visualiser la LUM dans les régions d'une ville où l'intensité est élevée, soit au niveau activité, au niveau résidentiel, ou les deux. D'autres modes de visualisation ne permettent pas toujours de montrer cela.

Dans ce travail, nous utilisons l'indice d'entropie (EI – Entropy Index) comme mesure de mixité. Ce choix est essentiellement guidé par le fait qu'il s'agit d'une mesure fréquemment utilisée (voir par exemple Song et Knaap (2004)). Cependant, comme précisé section 3, notre but n'est pas ici d'évaluer la pertinence de cet indice ni d'en proposer un nouveau, mais de proposer une méthode qui permet de calculer directement une représentation spatiale continue de n'importe quel indice existant pour peu qu'il soit basé sur des données présentes dans la base OpenStreetMap.

Soit $P_j(i)$ la proportion d'usage de type j estimé pour le point i à l'aide des KDE de la section précédente. Soit K le nombre d'usages différents [7], alors :

$$EI(i) = \frac{-\sum_{j=1}^{K} P_j(i)\ ln(P_j(i))}{ln(K)} \tag{2}$$

L'EI varie entre 0 et 1, sa valeur croissant avec la mixité des usages.

La figure 4 présente l'usage des sols calculé pour la ville Anglaise de Londres. La figure 3 utilise un bubble plot pour présenter les intensités locales de mixité d'usage des sols. Une analyse de ces résultats est présentée section 6.

7. Actuellement, nous avons $K = 2$ (activités et résidentiel), mais dès qu'OSM devient plus complet et précis (annotations fines des usages, cf. tableau 1) un nombre plus grand de classes d'activités peut être traité.

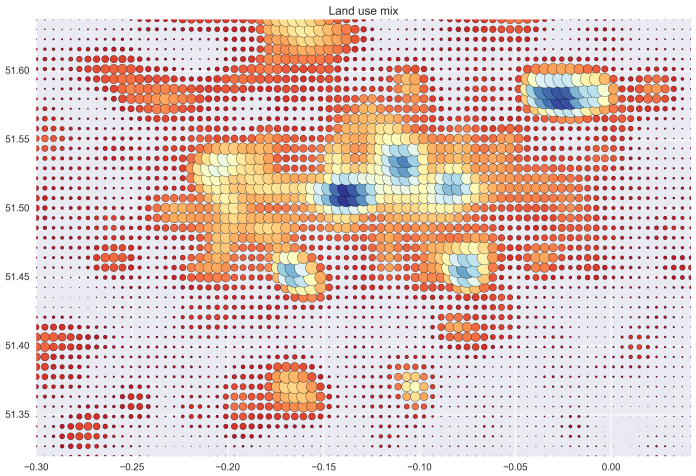


FIG. 3: “Bubble plot” de la LUM calculée pour Londres. La taille d’un point représente l’intensité locale de la mixité d’usage des sols.

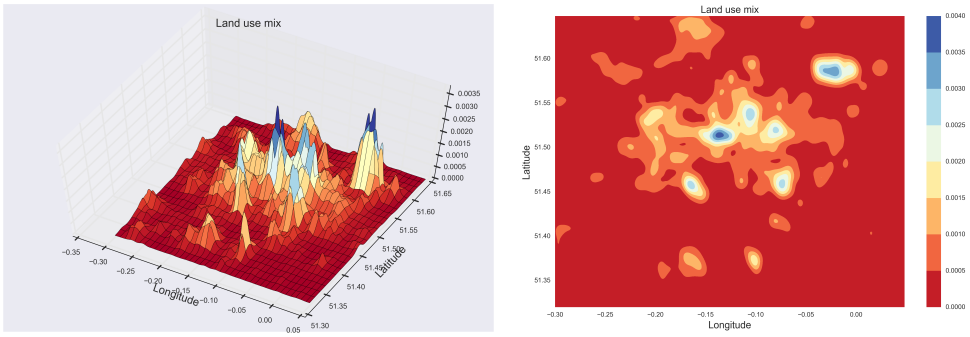


FIG. 4: LUM calculée pour Londres.

5.4 Implémentation

L’infrastructure logicielle de l’approche proposée, construite au-dessus d’une pile Python, est composée des modules suivants :

- `extract_uses.py` : étant donnés les points et polygones (*shapefiles*) *via* Mapzen Metro Extract, l’extraction des usages associés est effectuée selon la méthode décrite en section 5.1. La géométrie des polygones est traitée avec la librairie *Shapely* et la sortie finale est un *DataFrame Pandas* (McKinney (2010)) contenant les POIs géo-localisés et catégorisés, `pois_df`.
- `kde.py` : à partir des `pois_df`, ce module calcule les KDE pour les usages de type "activité" (`kde_act`) et "résidentiel" (`kde_res`) en utilisant la librairie *Statsmodels* (Seabold et Perktold (2010)).
- `loaders.py` : gère un store Hadoop distribué (HDFS, Hadoop Distributed File Store)

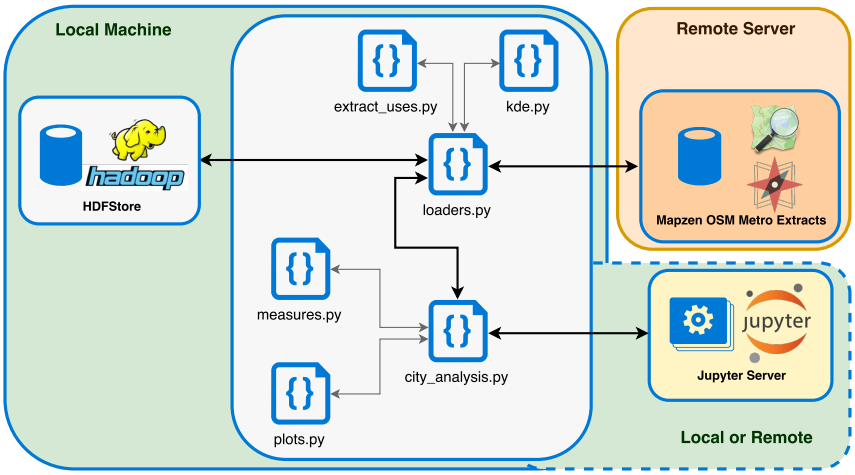


FIG. 5: Implémentation de la plateforme proposée.

à l'aide de la librairie décrite dans Alted et Fernández-Alonso (2003). Un tel HDFS stocke, pour chaque *ville* étudiée, les résultats des calculs potentiellement coûteux faits par les modules `extract_uses.py` et `kde.py` ci-dessus. Si, pour une *ville* donnée, ces informations ne sont pas stockées dans un HDFS, il est procédé comme suit :

1. Une requête pour les *shapefiles* de cette *ville* est envoyée à Mapzen [8]
2. Après réception des *shapefiles*, `extract_uses.py` et `kde.py` sont appelés afin de calculer `pois_df`, `kde_act` et `kde_res`
3. Les résultats sont stockés dans un HDFS local, sous une clé correspondant à la *ville*.

– `measures.py` : calcule les mesures de mixité qui sont implémentées (ici, l'indice d'entropie, cf. équation 2), à partir de `pois_df`, `kde_act` et `kde_res`, en utilisant des arrays *Numpy*, voir Van Der Walt et al. (2011).
– `plots.py` : ce module génère les sorties graphiques désirées à partir de `pois_df`, `kde_act` et `kde_res` (scatterplots, plots des KDE, etc.) à l'aide de la librairie *Matplotlib*, voir Hunter et al. (2007).
– `city_analysis.py` : c'est la classe Python principale, qui interagit avec le serveur Jupyter. Pour démarrer l'analyse pour une *ville*, elle demande les informations de région d'intérêt sur cette *ville* à `loaders.py`. Une fois celles-ci disponibles, la classe gère les appels aux méthodes `measures.py` et `plots.py` pour construire les sorties de manière interactive.

L'architecture de la plateforme est représentée sur la figure 5. Le code source est ouvert et disponible à l'URL `http://github.com/martibosch/landusemix` . Nous prévoyons d'implémenter une interface web permettant à des non-informaticiens d'utiliser notre outil en ligne.

8. `http://mapzen.com/data/metro-extracts/`

6 Applications

Cette section présente une première analyse de résultats sur deux villes Européennes aux structures assez différentes. Le but est d'abord de montrer que les données présentes dans la base OSM sont suffisamment précises pour reconstruire une carte continue de l'indice choisi à l'aide de la plateforme. Par ailleurs, cela permet aussi de constater que notre méthode permet de dépasser la limitation séparant les indices intégraux et divisionnels. En effet, nous calculons ici la version géographiquement continue d'un indicateur divisionnel, ce qui ouvre des perspectives sur la possibilité de comparer l'expressivité d'indices différents.

Cela dit, une validation plus poussée de la carte de LUM construite est nécessaire, et sera effectuée dans nos prochains travaux. Nous pourrons notamment comparer la carte LUM avec une vérité terrain pré-existante dans le cas de jeux de données suffisamment précis (et potentiellement propriétaires), reconstruire des indices intégraux connus sur des zones géographiques déjà analysées, et éliciter des avis d'experts du domaine sur les couches GIS calculées. Muni d'un tel outil, nous pourrons dans un second temps nous concentrer sur le calcul d'un indicateur optimal à l'aide de méthodes d'apprentissage supervisé, et comparer ce dernier aux indicateurs existants.

6.1 Grenoble, France

La ville Française de Grenoble est une ancienne cité Européenne de taille moyenne. Elle est localisée au pied des Alpes Françaises et possède une forme caractéristique de Y due aux massifs montagneux environnants de la Chartreuse, de Belledonne et du Vercors. Nous avons calculé la carte de LUM de cette agglomération de 665000 habitants. Un total de 33055 points et 118207 polygones ont été traités en $10min$; 11258 activités et 84445 points résidentiels ont été extraits. Les KDE de ces deux usages ont été calculés en $32min$, et la carte finale de mixité urbaine générée en $2s$.

Ces résultats sont présentés Fig. 6 et Fig. 7. On peut voir dans le KDE des activités une haute concentration autour du centre historique correspondant aux nombreuses activités commerciales qui y sont localisées. Ce motif est fréquent dans les villes Européennes dans lesquelles les centres contiennent souvent les plus grandes concentrations d'activités suite à une évolution régulière dans l'histoire.

La carte des usages résidentiels montre une forte concentration autour du centre historique, causée par le même phénomène. Les anciens habitats datant d'une époque antérieure à la voiture, la distance moyenne entre habitations est plus réduite, et leur concentration est donc plus élevée. Pour ce type de villes, les centres historiques présentent donc souvent de hauts niveaux de mixité urbaine.

Les résultats montrent de plus une caractéristique particulière de la ville de Grenoble. Les abords de la ville étant constitués de zones semi-montagneuses, plusieurs centres résidentiels peuvent être trouvés dans des zones suffisamment éloignées du centre pour fournir une bonne qualité de vie avec un accès rapide à la nature, tout en étant facilement accessible par transport public ou véhicule personnel. Les structures visibles aux coordonnées géographiques $5,58°E/45,13°N$ et $5,84°E/45,14°N$ représentent respectivement l'Est du plateau du Vercors et la vallée de Vaulnaveys. Il s'agit de deux zones résidentielles ayant beaucoup moins d'activités que le centre ville.

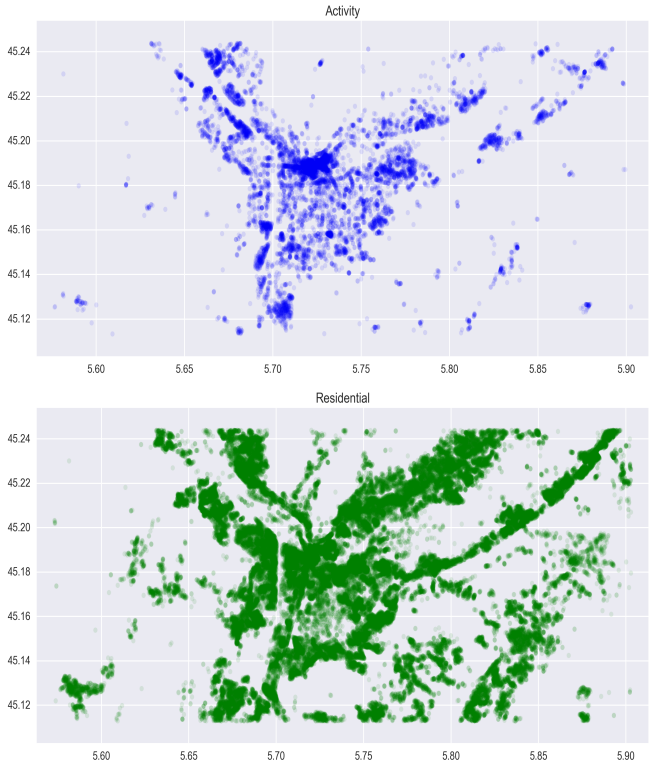


FIG. 6: Points d'intérêt extraits pour Grenoble, France (en haut les activités, en bas les zones résidentielles).

6.2 Londres, Angleterre

Nous avons évalué notre plateforme sur la cité de Londres. Un total de 80702 and 238626 points ont été respectivement extraits pour les usages de type "activité" et "résidentiel". Les résultats sont présentés Fig. 1 (usages), Fig. 2 (KDE) et Fig. 4 et 3 (LUM).

La cité Londonienne, qui a vécu de grands changements au cours de son histoire, possède actuellement $8,674$ millions d'habitants. On peut voir dans la figure 2 la City de Londres et le quartier de Soho comme ayant la plus grande densité d'activités. Du point de vue des usages résidentiels, East Dulwich, la colline de Higham, et la zone entre Wandsworth Common et Clapham Common sont les zones ayant le KDE le plus intense, ce qui correspond aux grandes concentrations d'habitation connues.

Croydon Vision 2020 est un programme de développement urbain qui vise à faire de Croydon un centre de vie, de commerce, de culture et d'activités dans le contexte du plan de développement local. Une importante concentration d'activités est donc présente dans cette zone, en conjonction avec une importante zone résidentielle un peu plus à l'Ouest. Nous devrions donc nous attendre à observer à cet endroit une forte mixité urbaine, d'autant plus que la volonté d'avoir un fort développement mixte est mise en avant par les concepteurs de ce pro-

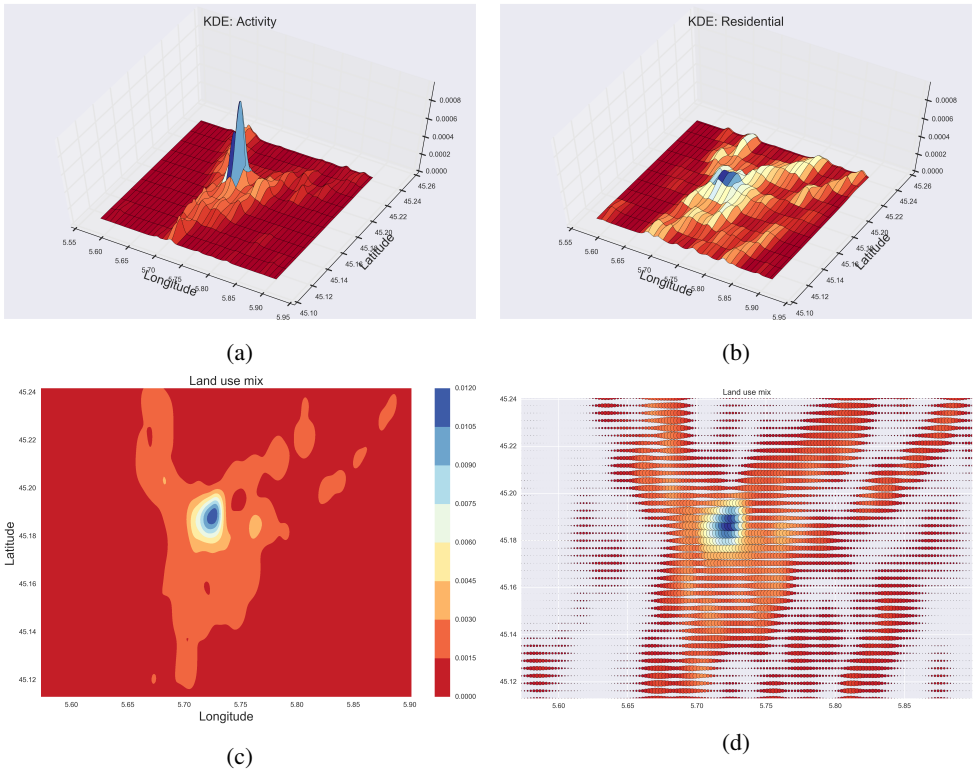


FIG. 7: Les résultats pour Grenoble, France. (a) et (b) montrent les KDE pour les activités et les zones résidentielles tandis que (c) et (d) montrent les LUM calculées selon deux modes différents.

gramme urbain. Et en effet, le résultat du calcul de l'indice de mixité montre qu'un haut niveau existe dans cette région, comme montré Fig.4, aux coordonnées géographiques -0.100594 et 51.376495. En surveillant l'évolution de cet indice dans cette zone, nous devrions par ailleurs constater une augmentation de l'intégration de différents usages dans le futur.

7 Conclusion et travaux futurs

Nous présentons dans cet article une plateforme pour capturer les informations de mixité d'usage des sols au sein des villes. L'outil logiciel présenté extrait des données collaboratives d'OpenStreetMap une représentation géo-localisée des usages. De par ses caractéristiques, cette source de données nous permet de traiter n'importe quelle zone géographique si suffisamment de données sont disponibles dans la base. Une estimation de densité par noyau permet de générer une distribution spatiale pour chaque usage. À partir de cette représentation, une mesure de mixité d'usage des sols est ensuite calculée. Dans le cas de ces travaux, nous utilisons l'indice d'entropie, mais n'importe quelle métrique peut être calculée avec cette

méthode. La couche SIG finale montre la zone géographique choisie enrichie par l'information de mixité urbaine dans le but d'aider les décideurs et les aménageurs.

Notre outil, disponible en ligne et dont l'implémentation est décrite, est ici appliqué aux villes de Londres et de Grenoble. Les estimations des usages résidentiel et d'activités sont pour l'instant validées en inspectant les voisinages locaux ayant une haute valeur d'usages, et en les re-contextualisant. Finalement, la distribution spatiale de la mixité urbaine est calculée, permettant une visualisation simple, et montrant les variations de la métrique sur l'ensemble de la zone d'intérêt.

Les travaux futurs sont multiples. Nous allons tout d'abord approfondir le processus de validation de la carte de LUM construite, en l'utilisant d'une part pour reconstruire des indices intégraux connus sur des zones géographiques déjà analysées, et d'autre part en élicitant des avis d'experts du domaine. Nous envisageons ensuite d'appliquer notre outil à l'analyse du phénomène de l'étalement urbain, en effectuant une interprétation numérique des valeurs de mixité "désirables" dans ce contexte. Nous planifions également d'enrichir la représentation actuelle en intégrant un modèle de transport qui permettra une estimation de la mixité urbaine plus précise en prenant en compte, en plus de leur localisation, l'accessibilité des différents usages, en partie conditionnée par l'infrastructure de transport. Enfin, nous analyserons plus en détail les différents types d'activités qui sont actuellement toutes regroupées dans une seule classe. En effet, la richesse des informations de la base OpenStreetMap permet de détailler les activités extraites en nombreuses classes (activité commerciale, industrielle, de loisir, etc.). Notre modèle permet de créer un KDE pour chaque classe et, en faisant varier leur importance relative, pourrait converger vers une représentation plus réaliste.

Références

Alted, F. et M. Fernández-Alonso (2003). Pytables : processing and analyzing extremely large amounts of data in python. In *PyCon*, pp. 1–9.

Atash, F. (1994). Redesigning suburbia for walking and transit : Emerging concepts. *Journal of Urban Planning and Development 120*(1), 48–57.

Aultman-Hall, L., M. Roorda, et B. W. Baetz (1997). Using gis for evaluation of neighborhood pedestrian accessibility. *Journal of Urban Planning and Development 123*(1), 10–17.

Barron, C., P. Neis, et A. Zipf (2014). A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS 18*(6), 877–895.

Breheny, M. (1995). The compact city and transport energy consumption. *Transactions of the Institute of British Geographers 20*(1), 81–101.

Burgess, R. (2000). The compact city debate : A global perspective. In R. Burgess et M. Jenks (Eds.), *Compact cities : Sustainable urban forms for developing countries*, pp. 9–24. Spon Press : London, UK.

Carlos, H. A., X. Shi, J. Sargent, S. Tanski, et E. M. Berke (2010). Density estimation and adaptive bandwidths : a primer for public health practitioners. *International Journal of Health Geographics 9*(1), 1.

Cervero, R. (1996). Mixed land-uses and commuting : evidence from the american housing survey. *Transportation Research Part A : Policy and Practice 30*(5), 361–377.

Cervero, R., C. Ferrell, et S. Murphy (2002). Transit-oriented development and joint development in the united states : A literature review. Technical Report 52, Transportation Research Board.

Christian, H. E., F. C. Bull, N. J. Middleton, M. W. Knuiman, M. L. Divitini, P. Hooper, A. Amarasinghe, et B. Giles-Corti (2011). How important is the land use mix measure in understanding walking behaviour ? results from the reside study. *International Journal of Behavioral Nutrition and Physical Activity 8*(1), 1.

Ewing, R. H. (1995). Characteristics, causes, and effects of sprawl : A literature review. In *Urban Ecology*, pp. 519–535. Springer.

Fan, H., A. Zipf, Q. Fu, et P. Neis (2014). Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science 28*(4), 700–719.

Forghani, M. et M. R. Delavar (2014). A quality study of the openstreetmap dataset for tehran. *ISPRS International Journal of Geo-Information 3*(2), 750–763.

Frank, L. D., M. A. Andresen, et T. L. Schmid (2004). Obesity relationships with community design, physical activity, and time spent in cars. *American journal of preventive medicine 27*(2), 87–96.

Girres, J.-F. et G. Touya (2010). Quality assessment of the french openstreetmap dataset. *Transactions in GIS 14*(4), 435–459.

Goodchild, M. F. (2007). Citizens as sensors : web 2.0 and the volunteering of geographic information. *GeoFocus 7*, 8–10.

Gordon, P. et H. W. Richardson (1997). Are compact cities a desirable planning goal ? *Journal of the American Planning Association 63*(1), 95–106.

Haklay, M. (2010). How good is volunteered geographical information ? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B : Planning and design 37*(4), 682–703.

Haklay, M. et P. Weber (2008). Openstreetmap : User-generated street maps. *Pervasive Computing, IEEE 7*(4), 12–18.

Hunter, J. D. et al. (2007). Matplotlib : A 2D graphics environment. *Computing in science and engineering 9*(3), 90–95.

Johnson, M. P. (2001). Environmental impacts of urban sprawl : a survey of the literature and proposed research agenda. *Environment and planning A 33*(4), 717–735.

Kockelman, K. (1997). Travel behavior as function of accessibility, land use mixing, and land use balance : evidence from san francisco bay area. *Transportation Research Record : Journal of the Transportation Research Board 1607*, 116–125.

Leccese, M. et K. McCormick (2000). *Charter of the new urbanism*. McGraw-Hill Professional.

Liu, X. et Y. Long (2016). Automated identification and characterization of parcels with openstreetmap and points of interest. *Environment and Planning B : Planning and Design 43*(2), 341–360.

McCormack, G. R., B. Giles-Corti, et M. Bulsara (2008). The relationship between destination proximity, destination mix and physical activity behaviors. *Preventive Medicine 46*(1), 33–

40.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Volume 445, pp. 51–56.

Mooney, P., P. Corcoran, et A. C. Winstanley (2010). Towards quality metrics for openstreetmap. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 514–517. ACM.

Neis, P., D. Zielstra, et A. Zipf (2011). The street network evolution of crowdsourced maps : Openstreetmap in germany 2007–2011. *Future Internet 4*(1), 1–21.

Nelson, A. et J. Duncan (1995). *Growth management principles and practices*. Planners Press, American Planning Association.

Neuman, M. (2005). The compact city fallacy. *Journal of Planning Education and Research 25*(1), 11–26.

Newman, P. et J. Kenworthy (2006). Urban design to reduce automobile dependence. *Opolis 2*(1), 35–52.

Over, M., A. Schilling, S. Neubauer, et A. Zipf (2010). Generating web-based 3d city models from openstreetmap : The current situation in germany. *Computers, Environment and Urban Systems 34*(6), 496–507.

Raines, G. L. (2002). Description and comparison of geologic maps with fragstats—a spatial statistics program. *Computers & Geosciences 28*(2), 169–177.

Seabold, S. et J. Perktold (2010). Statsmodels : Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pp. 57–61.

Sierra Club (1998). Sprawl : The dark side of the American dream. Technical report.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Smart Growth Network (2006). This is smart growth. Technical report.

Song, Y. et G.-J. Knaap (2004). Measuring the effects of mixed land uses on housing values. *Regional Science and Urban Economics 34*(6), 663–680.

Song, Y., L. Merlin, et D. Rodriguez (2013). Comparing measures of urban land use mix. *Computers, Environment and Urban Systems 42*, 1–13.

Squires, G. D. (2002). *Urban sprawl : Causes, consequences, & policy responses*. The Urban Insitute.

United Nations (2014). World urbanization prospects : The 2014 revision. Technical Report ST/ESA/SER.A/352.

United States Department of Housing and Urban Development (1999). State of the cities – 1999. Technical report.

Van Der Walt, S., S. C. Colbert, et G. Varoquaux (2011). The numpy array : a structure for efficient numerical computation. *Computing in Science & Engineering 13*(2), 22–30.

Western Australian Planning Commission (2007). Liveable neighbourhoods : a Western Australian Government sustainable cities initiative.

Summary

Population in urban areas has been increasing at an alarming rate in the last decades. This evidence, together with the rising availability of massive data from cities, has motivated research on sustainable urban development.

In this paper we present a GIS-based land use mix analysis framework to help urban planners to compute indices for mixed uses development, which may be helpful towards developing sustainable cities. Residential and activities land uses are extracted using OpenStreetMap crowd-sourcing data. They are used to compute a continuous representation of the entropy index, but can be generalized to the building of any other index. The framework is applied to several cities and the first results are presented.

Profilage sémantique et probabiliste de zones géographiques

Jérémy Lhez, Olivier Curé

LIGM (UMR 8049), CNRS
ENPC, ESIEE, UPEM, Marne-la-Vallée, France
{prenom.nom}@u-pem.fr

Résumé. La détection d'anomalies dans des réseaux de capteurs est une problématique prépondérante dans de nombreux domaines où les ressources sont précieuses. Dans le projet FUI (Fonds Unique Interministériel) WAVES, nous analysons et traitons des flux de données relatifs à la distribution d'eau potable. Ces traitements imposent différentes formes de raisonnement. Une partie de ces inférences se base sur le profilage de zones géographiques englobant ces capteurs. Dans cet article, nous présentons deux méthodes de profilage complémentaires qui exploitent des mesures provenant des capteurs ainsi que la topologie du réseau, des connaissances externes, p.ex., les points d'intérêts, et un raisonnement probabiliste. La complémentarité de nos approches est validée par des experts du domaine au travers d'expérimentations menées sur de données réelles de terrain.

1 Introduction

L'implication des technologies du Web Sémantique, en particulier la capacité à raisonner, dans le domaine de l'internet des objets (IoT) présente un fort potentiel pour la détection d'anomalies. Ces détections peuvent améliorer la gestion de ressources précieuses, p.ex., énergie, et ainsi avoir un impact important sur des aspects économiques et écologiques. Ce constat concerne totalement notre cas d'usage qui porte sur le transport et la distribution d'eau potable.

De nos jours, les fuites d'eau potable dans le monde représentent une perte de 32 milliards de dollars chaque année. Outre les pertes financières, elles posent également des problèmes de maintenance, puisque 90% d'entre elles sont invisibles, et sont donc très difficiles à réparer. Pour pallier ces difficultés, de plus en plus de capteurs de mesure, p.ex., pression, débit, chlore, turbidité, sont installés dans ces réseaux. Ils fournissent des mesures en temps plus ou moins réel sur l'état du réseau. Par l'analyse de ces données, il devient alors possible de détecter automatiquement des anomalies potentielles. Avant de valider l'origine de ces anomalies, certains traitements intégrant le contexte du capteur sont nécessaires. Dans notre cas sur l'eau potable, il est particulièrement important de caractériser les zones géographiques où se trouvent les capteurs. Étant donné l'expansion du nombre de capteurs sur de tels réseaux, il est indispensable de développer une méthode de profilage automatique. Dans cet article, nous présentons deux méthodes de profilage complémentaires. Celles-ci se basent sur des connaissances externes obtenues sur le Web, et se combinent via les mesures fournies par notre partenaire sur le projet, un acteur international majeur de la distribution d'eau.

Cet article est organisé de la manière suivante : nous exposons notre cas d'utilisation spécifique dans la seconde section, puis nous détaillons les données en entrée utilisées par le projet. La quatrième section présente nos méthodes de profilage et leur complémentarité, et nous en évaluons les performances en section 5. La section 6 présente un état de l'art de divers systèmes de profilage existant pour comparaison ; nous terminons notre article avec une conclusion en section 7.

2 Présentation du projet WAVES et de son cas d'usage

Le projet WAVES [1] est une plateforme évolutive et distribuée pour traiter des flux sémantiques sur la base de plateformes open source. Son objectif principal est d'identifier les anomalies présentes dans les données des flux via différentes méthodes, et d'en déterminer les causes en utilisant des méthodes de raisonnement exploitant des sources d'information externes.

Les données reçues subissent un filtrage apportant certaines corrections aux mesures manquantes ou incohérentes. Puis les données sont sémantisées et converties dans un format unique correspondant à une sérialisation compressée de RDF (Resource Description Framework) (rdf (2014), afin d'être traitées plus rapidement. Par la suite, chaque flux est requêté et analysé afin d'identifier des anomalies potentielles, suivant des règles définies par le client. Ces anomalies, ainsi que les données problématiques détectées lors du filtrage, sont confiées à un module de raisonnement qui se charge de trouver leur origine ; nous utilisons pour cela des sources de données spécifiques au réseau d'eau potable (diverses mesures, topologie) ainsi que des bases de connaissances externes (données géographiques, événements permettant de justifier les anomalies). Les origines identifiées sont finalement classées par probabilité. Les erreurs détectées et leurs sources proposées sont ensuite mises en forme par un module de visualisation, qui permet à l'opérateur final d'agir en conséquence : en validant/refusant les déductions du module de raisonnement.

Notre cas d'usage précis concerne la détection de fuites et autres anomalies sur un réseau de distribution d'eau potable. Il est possible d'appliquer diverses méthodes de détection sur les valeurs reçues pour identifier plusieurs sortes de problèmes : dépassement de seuil règlementaire, variation de la moyenne journalière, profil de consommation inhabituel, etc.

Pour identifier les origines possibles de ces anomalies, le module de raisonnement doit se baser sur l'analyse de données externes. En effet, si les données incohérentes ou manquantes peuvent s'expliquer relativement facilement (capteur défectueux, obstacle empêchant l'envoi de données...), il devient très difficile de déterminer de manière sûre l'origine de problèmes plus spécifiques (diminution de pression ponctuelle, augmentation du taux de chlore...). Il convient donc d'étudier d'autres sources de données pour trouver des causes valables ; il s'agit majoritairement de données événementielles, qui bouleversent la moyenne habituelle des données. On peut citer les accidents, tels que les incendies ou les inondations, mais aussi les manifestations culturelles, comme un concert, une compétition sportive ou encore une simple consommation plus importante de la part d'un gros consommateur sur le réseau, pour le remplissage d'une piscine ou l'arosage d'un golf, etc.

Il est donc difficile d'identifier l'origine la plus probable au premier abord. On peut se référer à des paramètres comme la distance entre le capteur et les sources d'anomalie ou le

1. http ://www.waves-rsp.org/

temps entre la détection de l'erreur et la date des événements pour établir un classement par pertinence exploitant une approche probabiliste. Il faut aussi tenir compte du lien entre l'événement et l'anomalie (un remplissage de piscine ne peut pas causer d'augmentation de température). Notre client éprouve un intérêt particulier pour le module de raisonnement, qui est la principale fonctionnalité attendue du projet, et nous avons longuement discuté des méthodes à employer pour le rendre le plus précis possible. Nous avons débattu au sujet des sources de données externes pertinentes et nécessaires pour la découverte des origines d'erreurs, et nous avons convenu qu'établir un profilage des secteurs de consommation pouvait être extrêmement bénéfique pour orienter ces recherches.

Dans notre cas d'usage, chaque capteur dépend d'un secteur de consommation spécifique, et une anomalie qu'il détecte sera donc imputée au secteur plutôt qu'à un simple point. En utilisant cette méthode, il devient possible d'identifier la nature de la zone où se trouve le capteur (s'agit-il d'une zone résidentielle ? d'un secteur industriel ?), et d'ajuster ainsi les probabilités sur les sources potentielles identifiées. Mieux, on peut avant cela donner la priorité à certains événements jugés plus probables : si le capteur est situé dans une zone touristique, on donnera une probabilité plus forte, disons à une surconsommation du parc d'attraction, plutôt qu'à une exposition d'art située plus proche du capteur.

3 Données et connaissances nécessaires au profilage

Pour mener à bien nos travaux de profilage, nous disposons d'un grand nombre de données de mesure portant sur plusieurs zones géographiques nationales et internationales, p.ex., des villes, des régions. Nous pouvons y trouver les données reçues, à raison d'un fichier par capteur, regroupées dans des dossiers en fonction du type de mesure. Chaque donnée relevée se présente sous forme d'un couple date-valeur, il est donc aisé de modéliser un flux de données pour chaque capteur pour effectuer des tests.

En plus de ces données de capteurs, le répertoire contient des fichiers de configuration, au format XML et CSV, contenant de nombreuses informations sur la région concernée. L'unité des mesures y est indiquée, ainsi que des détails sur la répartition de la région en secteurs de consommation, la position des capteurs sur ces secteurs, et bien d'autres.

Outre ces données, nous avons également reçu un fichier shapefile (fichier de formes, format ESRI ; voir Esr (1998)) et un fichier KML représentant la région et ses secteurs de consommation, permettant une visualisation précise des données qui nous serviront essentiellement pour mener des vérifications sur nos résultats.

Nos méthodes de profilage ne se limitent pas à l'analyse de mesures du réseau. Après de nombreux échanges avec les experts métiers, nous nous sommes rapidement rendu compte que des données et connaissances externes seraient nécessaires. Dans cette optique, nous avons étudié plusieurs grandes bases de connaisances du Linked Open Data (LOD), p.ex. DBPedia, Wikidata et Freebase. Nous avons évalué celles-ci en donnant une importance particulière aux données géographiques, p.ex., densité, superficie mais également les points d'intérêt (POI). A la suite de cette étude, nous nous sommes rapidement concentrés sur des bases de connaissances spécialisées sur l'information géographique. La source la plus intéressante s'est avérée être OpenStreetMap.

Il s'agit d'un projet international ayant pour but de cartographier le monde entier, en récoltant une majorité de détails. OpenStreetMap regroupe une grande variété de POI et de détails,

qui peuvent être visualisés via une interface en ligne [2] sous forme de carte, interrogés par le biais d'un système de requêtes pour utilisations sur diverses plateformes, ou récupérés depuis un navigateur pour enregistrement ou traitement.

On peut recenser d'autres bases de données cartographiques disponibles en ligne, offrant divers niveaux d'information.

HERE est un logiciel de cartographie fournissant globalement le même genre d'informations qu'OpenStreetMap : POI, accès depuis différentes plateformes, etc. Un accès en ligne est également fourni [3], mais l'utilisation des données est payante.

Enfin, GeoNames [4] est une base de données géographiques gratuite recensant de nombreuses informations concernant divers points d'intérêt, regroupés en différentes catégories. Certains emplacements ne sont que des points, d'autres sont sous forme de surfaces (de polygones), malgré tout, toutes ces informations se retrouvent plus détaillées dans OpenStreetMap.

4 Méthodes de profilage

Nous avons implémenté deux méthodes de profilage dans le cadre de notre projet, utilisant les données fournies par OpenStreetMap. Leurs résultats sont souvent complémentaires, et peuvent être combinés pour plus de précision avec une troisième méthode, évaluant la consommation en fonction de la taille du réseau. Nous verrons lors de l'évaluation que les résultats finaux issus de cette combinaison présentent une qualité très satisfaisante qui traduisent bien la réalité du terrain.

4.1 Profilage par zones - Points d'intérêt

La première méthode de profilage exploite les POI et tend à caractériser des zones géographiques. Elle nous permet de déterminer la composition des secteurs de consommation en termes de type de surface (industrialisation, agriculture, etc.). Établir ce type de profil permet d'évaluer l'utilisation des sols de chaque secteur de consommation, et donc de savoir quelle surface y est majoritairement représentée, afin d'orienter les recherches des sources d'anomalies. Pour cela, nous avons utilisé les données géographiques d'OpenStreetMap, qui recense un très grand nombre de POI, répartis par type (restauration, monument, etc.).

Dans un premier temps, nous avons déterminé les grandes catégories de surface à identifier pour notre cas d'utilisation. À partir des souhaits des experts du domaine de l'eau, nous en avons retenues cinq : naturelle, industrielle, résidentielle, touristique et agricole ; mais la liste de ces catégories peut être enrichie, en définissant des sous-classes à ces quatre super classes, p.ex., la zone naturelle peut être décomposée pour différencier les secteurs agricoles de ceux purement naturels. Cependant, un tel enrichissement ne sera pertinent que si la qualité et la quantité des POI des zones traitées sont suffisantes.

Ensuite, nous avons établi un fichier de notation, recensant les POI représentatifs de ces surfaces, auxquels nous avons attribué un score. OpenStreetMap a créé différentes catégories recensant les POI, établissant une hiérarchie de concepts (un bar est caractérisé comme un lieu de restauration et est classé dans la catégorie des aménités). Nous avons repris cette hiérarchie

2. https ://www.openstreetmap.org
3. https ://wego.here.com
4. http ://www.geonames.org/

en y intégrant les catégories de surface mentionnées précédemment, pour indiquer les POI qui en sont représentatifs. La note attribuée aux POI indique leur pertinence (par exemple, une forêt sera plus représentative d'une zone naturelle qu'une aire de jeu ; sa note sera donc plus élevée) ; les fichiers de notes ont été soumis à notre client pour validation, nous avons tenu compte de leur retour pour corriger, puis finalement valider notre approche. Ces fichiers de notation peuvent néanmoins être ajustés en fonction des résultats obtenus si les vérifications ne s'avèrent pas convenables.

Notre programme se base sur ces informations pour établir le profilage. Dans un premier temps, il parcourt les fichiers de configuration des secteurs de consommation, afin de les modéliser et d'en extraire les limites (coordonnées des points inférieur gauche et supérieur droit). Ces coordonnées permettent d'envoyer des requêtes au service d'OpenStreetMap pour récupérer les POI présents dans le secteur concerné (il faut refaire un test d'inclusion : la zone extraite couvre plus de surface que le secteur de consommation). Ensuite, pour chacun des POI du secteur, nous nous basons sur le fichier de notation pour savoir de quelle type de surface il est représentatif : en faisant de simples additions, on attribue ainsi un score à chaque surface, déterminant sa prépondérance dans le secteur de consommation. Dans l'objectif d'une comparaison efficace, nous normalisons les données et présentons des pourcentages d'appartenance à une de nos catégories :

$score\_zone = \sum note_{poi}$ ou $score\_zone$ caractérise une zone géographique et $note_{poi}$ représente un score attribué à un POI se trouvant dans cette même zone géographique.

$proportion_{zone} = \frac{score\_zone}{\sum_{i=1}^{n} score\_zone_i} \times 100$ où n est le nombre de zones géographiques à considérer.

Les résultats sont calculés en fonction des aires des zones pertinentes relevées, et non en fonction de la surface du secteur de consommation. Nous avons réalisé des vérifications de nos résultats sur l'intégralité des données à notre disposition, en modélisant les différents secteurs de consommation sous forme de polygones, et en les superposant aux données d'OpenStreetMap via Google Earth. Les secteurs étant relativement peu étendus, il est aisé de repérer les POI les plus nombreux ; on peut aussi se baser sur le code couleur des différentes surfaces (vert pour les zones naturelles, etc.). Les résultats ont été éxaminés pour l'ensemble des zones, et nous avons conclu que seule une partie des résutats est satisfaisante. En effet, certains types de surface possèdent peu de POI représentatifs, et la proportion finale ne correspond pas à la réalité (par exemple, une zone naturelle contient très peu de POI). Autre problématique : certains secteurs parmi les plus petits n'ont que peu de POI, et leur profil obtenu a donc une précision diminuée. Puisque les POI ne convenaient pas dans certains cas, nous avons décidé d'utiliser les autres types de données qu'OpenStreetMap mettait à disposition.

4.2 Profilage par zones - Régions

La seconde méthode développée est un autre profilage par région, variante de la méthode utilisant les POI d'OpenStreetMap. L'objectif est donc le même que précédemment : obtenir la proportion des différents types de surface au sein des secteurs de consommation. Toutefois, l'utilisation de surfaces réelles à la place de notes arbitraires permet d'avoir un meilleur aperçu des proportions calculées une fois le profilage effectué.

Pour ce profilage, le système extrait d'OpenStreetMap non seulement les POI, mais aussi les linéaires (routes/barrières/cours d'eau) ; une fois ces derniers récupérés, il faut identifier

ceux qui sont fermés, et qui forment donc des polygones. Nous réutilisons le même fichier de notes que précédemment pour identifier les types de régions représentatives de surfaces géographiques, de la même manière que précédemment ; les notes ne sont toutefois pas utilisées. Il faut ensuite représenter ces régions, afin de pouvoir calculer leur aire. Cette partie s'est avérée délicate : OpenStreetMap représente les linéaires en listant les identifiants des POI les composant ; il peut donc y avoir des cas où les POI, étant en dehors de la *bounding box*, ne sont pas présents parmi les données récupérées. Nous avons décidé de ne tenir compte que des points présents dans le secteur de consommation, et de fermer le polygone prématurément si certains venaient à manquer : récupérer davantage de données s'avèrerait plutôt complexe et nuirait aux performances. Plutôt que d'additionner des notes pour chaque surface géographique, cette méthode de profilage utilise les aires des régions, calculées à partir des coordonnées de leurs polygones. En utilisant les surfaces de chaque secteur de consommation, on peut ainsi calculer un pourcentage de la même manière qu'avec les notes des POI.

$surface_{zone} = \sum surface_{polygone}$ pour chaque $polygone$ représentatif de la $zone$ courante, et

$proportion_{zone} = \frac{surface_{zone}}{\sum_{i=1}^{n} surface_{zone_i}} \times 100$ pour les n types de zone définis.

Ici aussi on peut ajuster les résultats en modifiant le fichier de notes (nous parlons ici de l'arborescence uniquement, puisque les notes ne sont pas prises en considération) ; on peut aussi changer la méthode de calcul d'aire pour obtenir d'autres résultats. Le tableau 1 présente les résultats obtenus par chaque méthode de profilage, pour chaque secteur de consommation (indiqués sur la figure 1). On peut noter que les résultats des deux méthodes sont différents : certaines proportions correspondent, mais d'autres ont une marge de différence significative. On retrouve la problématique précédente de représentation : tous les types de surfaces ne sont pas aisément identifiables avec des polygones (dans notre cas d'usage, les zones touristiques).

zone	type de surface									
	industrielle		naturelle		résidentielle		touristique		agricole	
	POI	région	POI	région	POI	région	POI	région	POI	région
P Laval	15.22	15.81	3.26	80.93	73.37	3.26	8.15	0.00	0.00	0.00
V Nouvelle	13.41	17.87	1.45	54.76	56.06	19.68	29.08	0.22	0.00	7.47
Hubies D.	4.55	0.00	0.00	3.45	29.55	55.72	65.90	0.02	0.00	40.81
Brezin	0.00	0.00	0.00	80.99	100.00	19.01	0.00	0.00	0.00	0.00
Guyancourt D.	0.00	7.01	0.00	90.16	96.87	0.60	3.13	2.23	0.00	0.00
Louveciennes	13.31	1.88	2.63	64.29	45.96	11.85	38.10	2.61	0.00	19.37
Hubies H.	1.17	0.02	1.31	94.02	35.33	5.48	62.19	0.48	0.00	0.00
Haut-Clagny	20.95	0.45	9.52	69.49	60.95	29.79	8.58	0.27	0.00	0.00
Garches	7.20	0.00	5.95	70.08	27.57	29.76	59.28	0.16	0.00	0.00
Gobert	15.94	0.00	2.72	47.71	24.46	51.62	56.88	0.67	0.00	0.00
Satory	6.81	59.41	0.00	12.49	71.91	13.27	21.28	0.06	0.00	14.77

TAB. 1 – *Résultats des deux méthodes de profilage, en pourcentage)*

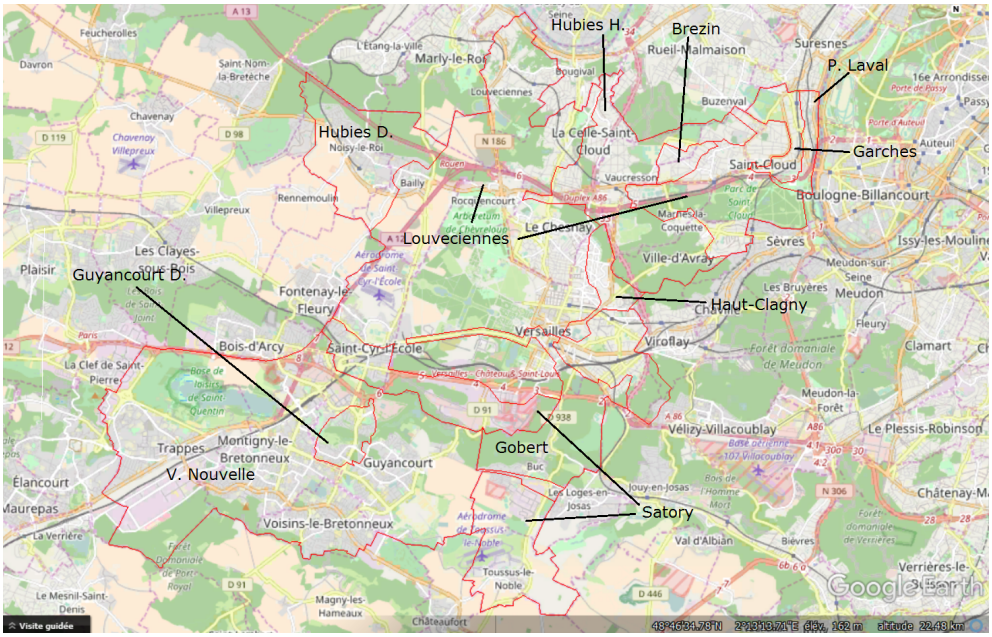


FIG. 1 – *Secteurs de consommations de Versailles, superposés aux données d'OpenStreetMap*

4.3 Rapport de consommation

Comme mentionné précédemment, nous avons trouvé un moyen de déterminer la méthode de profilage la plus adaptée, ou de les combiner le cas échéant, en établissant le rapport de consommation pour chaque secteur de consommation. Nous nous basons sur les anciennes relèves des capteurs afin d'établir une moyenne de consommation journalière. Les archives de mesurez dont nous disposons sont divisées en répertoires pour chaque type de mesure (seul le débit nous intéresse), contenant des fichiers portant chacun le nom d'un capteur. Les mesures y sont enregistrées dans deux colonnes, la première contenant la date et l'heure de la relève, la deconde la valeur mesurée. La moyenne se calcule donc de manière simple ; on peut par ailleurs l'ajuster en sélectionnant une période plus ou moins longue (sur une semaine, un mois, une saison...). Il faut tenir compte des flux d'entrée et de sortie : certaines valeurs peuvent être négatives. Nous divisons ensuite cette moyenne par la longeur totale du réseau dans le secteur de consommation concerné pour obtenir le rapport. L'étendue du réseau est extraite du fichier de configuration. Nous avons choisi cette valeur plutôt que la surface de la zone, car certains secteurs sont de petite taille mais avec un réseau très dense ; ce n'est pas gênant pour les grandes zones vides, car le réseau est quand même présent, même sans consommateur. Notons que le résultat du rapport peut être négatif : cela correspond aux cas ou il y a plus de débit en sortie du secteur qu'en entrée, potentiellement du fait de la présence de réservoirs dans la zone.

4.4 Complémentarité des méthodes de profilage

Comme nous l'avons mentionné, chacune de nos méthodes de profilage prises seule ne permet pas toujours d'établir un profilage précis. Certaines surfaces géographiques n'ont pas de POI représentatifs, et à l'inverse certains points n'ont pas de représentation en polygones, par exemple un champ. Il nous fallait trouver une méthode permettant de définir l'approche la plus pertinente, ou de les combiner si les deux résultats sont jugés corrects.

Le rapport de consommation établi précédemment permet de créer ce lien. Il représente la consommation moyenne du secteur en fonction de la longueur du réseau, ce qui correspond à ce que l'on peut appeler le taux de consommation. Le taux de consommation est bas lorsqu'il y a une faible concentration de consommateurs dans la zone : le secteur sera donc plutôt composé de polygones, représentatifs de zones plus larges. A l'inverse, si le taux est élevé, il y a une haute probabilité que la quantité de points d'intérêt dans le secteur soit également importante (beaucoup de consommateurs). Dans le cas d'un taux moyen, on peut se contenter de faire la moyenne entre les deux profilages, afin de mieux préciser le résultat (dans ces cas, l'écart entre les deux approches est plus faible) ; cela représente des zones mitigées, avec par exemple des quartiers d'habitation bordés de parcelles agricoles. Ces calculs peuvent être résumés par les formules suivantes.

$$rapport_{secteur} = \frac{debit\_moyen_{secteur}}{longueur\_reseau_{secteur}}$$

$rapport_{secteur} < borne_{inf}$:

$$proportion_{surface} = proportion\_surface_{region}$$

$rapport_{secteur} > bone_{sup}$:

$$proportion_{surface} = proportion\_surface_{POI}$$

$borne_{inf} < rapport_{secteur} < borne_{sup}$:

$$proportion_{surface} = \frac{proportion\_surface_{region} + proportion\_surface_{POI}}{2}$$

zone	catégorisation	rapport de débit	évaluation
P. Laval	résidentielle	élevé	correct
V. Nouvelle	rés. - nat.	moyen	correct
Hubies D.	rés. - agr.	faible	nuancé
Brezin	rés. - nat.	moyen	nuancé
Guyancourt D.	rés. - nat.	moyen	correct
Louveciennes	rés - tour.	élevé	correct
Hubies H.	naturelle	faible	correct
Haut-Clagny	résidentielle	élevé	correct
Garches	touristique	élevé	correct
Gobert	naturelle	faible	correct
Satory	industrielle	faible	nuancé

TAB. 2 – *Evaluation des résultats finaux*

5 Évaluation

Les mesures suivantes ont été effectuées sur un ordinateur portable avec le système d'exploitation Windows 8, équipé d'un processeur Intel Core i7 (2.90 GHz), avec 16GB de RAM, utilisant le JDK/JRE 1.8.

Diverses données ont été mises à notre disposition, au niveau national ou concernant des pays étrangers. Nous disposons de près de 200 secteurs de consommation, souvent concentrés dans certaines régions, parfois dispersés sur une large zone. Les vérifications ont été effectuées sur l'ensemble des jeux d'essai fournis par notre partenaire ; nous avons également généré des secteurs de consommation uniformes pour tester nos méthodes, p. ex. des zones industrielles ou des parcs d'attraction. Par soucis de briéveté, et en raison de la confidentialité des données à notre disposition, nous avons choisi de ne présenter qu'un seul cas d'utilisation précis : les tests exposés concernent la ville de Versailles et de ses alentours. Ses onze secteurs de consommation sont représentés (pour un total d'environ 350.000 habitants), comprenant un total de 67 capteurs de débit (les autres types de mesure ne sont pas pertinents pour les approches de profilage présentées dans cet article). La figure 1 donne un aperçu de la répartition des différents secteurs, représentés sur la couche de données d'OpenStreetMap. Le chateau de Versailles, monument national très reconnu, a été indiqué comme point de repère. Les données sont donc suffisament réparties et nombreuses pour permettre d'évaluer correctement notre algorithme.

L'algorithme commence par l'initialisation : extraction des données des secteurs de consommation du fichier de configuration, et récupération des données d'OpenStreetMap. Le volume de données à récupérer est de l'ordre du méga-octet, nous avons donc préféré stocker ces informations dans des fichiers locaux afin d'optimiser les performances de l'algorithme, compte tenu du fait que ce genre de données géographique est relativement statique.

Le tableau 3 donne le temps d'exécution des méthodes de profilage, du calcul du rapport de consommation, ainsi que la quantité de données disponibles, pour chaque secteur de consommation. Le temps d'exécution pour le calcul du rapport dépend essentiellement du nombre de capteurs sur le secteur : cela détermine le nombre de fichier auxquels il faut accéder pour extraire les mesures et calculer la moyenne. À l'inverse, le temps d'éxécution du profilage par POI est lié à la quantité de données extraites d'OpenStreetMap, et dépendra donc dans notre cas de la taille du fichier local enregistré. C'est également le cas pour le profilage par région, mais le temps d'éxécution sera plus important du fait de l'utilisation à la fois des POI et des polygones (c'est cette valeur qui est indiquée dans la dernière colonne du tableau). Malgré l'utilisation de données parfois importantes (plus de 120Mo dans le cas du plus grand secteur), le temps d'exécution de l'ensemble des méthodes est satisfaisant ; d'autant que le profilage ne s'effectue pas directement sur les flux de données. La combinaison des méthodes de profilage, détaillée en section 4.4 (page 8), combine ces deux méthodes par le biais d'une vérification simple, et effectue la moyenne pour les zones mixtes.

Afin de nous assurer de la pertinence de nos résultats finaux, nous avons consulté plusieurs opérateurs de notre client sur ce projet : étant très familier avec les zones exploitées en exemple, ils ont pu juger de nos résultats avec précision. Le profilage a été examiné par trois experts différentes, et jugé correct pour les secteurs fournis ; leurs observations sont consignées dans le tableau 2. La validation des méthodes basées sur les données d'OpenStreetMap peut être réalisée par superposition de couches géographiqes : l'utilisation de fichiers .kml permet de superposer les polygones représentant les secteurs de consommation avec la vue satellite de

zone	nombre de capteurs	données OSM disponibles	profilage		
			ratio de consommation	POI	région
P. Laval	2	5.4Mo	270ms	25ms	605ms
V. Nouvelle	16	53.8Mo	620ms	282ms	630ms
Hubies D.	1	5.8Mo	190ms	13ms	30ms
Brezin	1	3.1Mo	150ms	8ms	72ms
Guyancourt D.	2	4.2Mo	161ms	13ms	50ms
Louveciennes	19	123.2Mo	850ms	1118ms	1290ms
Hubies H.	13	37.15Mo	740ms	163ms	180ms
Haut-Clagny	4	8.6Mo	260ms	21ms	32ms
Garches	3	7.0Mo	220ms	205ms	46ms
Gobert	3	15.4Mo	201ms	36ms	105ms
Satory	5	32.5Mo	292ms	103ms	215ms

TAB. 3 – *Tableau de performance des différents algorithmes*

Google Earth. Nous pouvons ajouter les couches de données d'OpenStreetMap ou d'autres sources géographiques pour une meilleure comparaison.

6 État de l'art

De nombreuses méthodes de profilage existent à l'heure actuelle. Elles sont employées dans de nombreuses applications et sur des domaines variés. Toutefois, bien que certaines se basent sur des données géographiques, elles n'utilisent pas les mêmes méthodes que notre système. Nous avons étudié divers travaux ayant des considérations communes à notre méthode : nous avons cherché des projets avec des composantes spatiale, temporelle et sémantique, afin de pouvoir les comparer à notre approche.

La première méthode retenue, Węcel (2015), récupère les données géographiques relatives aux relais de communication mobile, ainsi que toutes sortes d'informations géolocalisées dans le voisinage de ces relais. Avec ces données, un profilage des relais de communication est établi ; en ce servant de ce profil, le système arrive à généraliser le comportement des personnes appelant grâce aux statistiques d'appel. Ce système de profilage a donc bien une composante géographique, et une composante temporelle (puisque les données sur les appels reçus concernent une période précise). Il repose sur plusieurs sources de données libres, notamment des ontologies géographiques dérivées d'OpenStreetMap. Dans le cadre du raisonnement, Węcel (2015) utilise les différentes catégories de POI définies par OpenStreetMap dans une ontologie ; chaque concept permettant de caractériser des lieux est accompagné d'un poids reflétant son importance dans l'analyse. Toutes ces techniques sont également employées dans notre approche, bien que notre cas d'utilisation ne soit pas le même. La formule finale que nous utilisons est très différente des calculs de cette méthode, car nous combinons les résultats de plusieurs méthodes de profilage, et pas directement des données géographiques.

Noulas et al. (2011) a pour but la modélisation de l'activité humaine en utilisant des catégories de lieux ; l'algorithme est basé sur une méthode de clustering, appliquée sur les zones géographiques et les données du réseau social Foursquare. Le profilage est réalisé en se ba-

sant sur les activités recensées par l'application (coordonnées GPS des restaurants, boutiques, etc...) ; ce principe a donc également une composante temporelle et spatiale. Ici encore, un système de classement basé sur des scores attribués sur les catégories de données est utilisé dans la formule de clustering. Le résultat donne les concentrations de POI de même nature, mais ne permet pas de caractériser de grandes zones : l'objectif ici est d'identifier des groupements d'activité de même nature.

Dans le domaine de la criminalité, on peut mentionner les travaux de Snook et al. (2002), qui traite du profilage géographique des criminels en série, et Rossmo (2014), qui sont basés sur les actions des gangs. Il s'agit dans chaque cas d'un profilage géographique, avec une composante temporelle, et pour lesquels l'usage de données extérieures est requis.

Dans chacun des cas, le profilage est réalisé à partir de POI extraits de données en ligne depuis différentes sources d'information. Si notre méthode utilise ce genre d'approche, elle ne s'y limite pas : en effet, nous utilisons deux calculs différents permettant de caractériser les secteurs de consommations, et les combinons pour obtenir un résultat plus précis. Les travaux mentionnés ne font que combiner des sources de données en ligne pour fournir un résultat, sans le réexploiter elles-mêmes. De plus, notre algorithme est le seul à utiliser des données de surfaces en plus des POI ; comme nous l'avons exposé précédemment, cela permet de combler le manque de POI représentatifs de certaines zones géographiques.

7 Conclusion

Dans cet article, nous avons présenté l'association de deux méthodes de profilage géographique différentes mais complémentaires, permettant d'évaluer les proportions de différentes catégories de surfaces au sein d'un groupe de secteurs. Les méthodes décrites ont été qualifiées d'ajout important et bénéfique pour l'un des modules du projet, et les résultats ont été évalués par des experts terrain dans le domaine de notre cas d'utilisation et jugés conforme dans notre cadre d'étude.

Nous prévoyons l'intégration de ce système de profilage au sein du projet de recherche FUI WAVES. Il jouera une place prépondérante dans notre module de raisonnement puisque certains comportements du réseau pourront être justifiés par les catégories associées à une zone donnée. Le système qui devrait être mis en production dans les prochains mois chez notre partenaire s'exécutera sur une large partie du territoire national ainsi que des grandes villes à l'international.

Notre système offre la possibilité d'ajuster les zones géographiques à définir, aussi nous pouvons modifier les types de surfaces utilisées pour l'établissement du profilage an fonction du cas d'usage. Nous travaillons actuellement à enrichir les données en entrée avec plus de données de contour et de délimitation, afin de préciser les résultats des méthodes de profilage. En améliorant nos catégories de profilage, nous estimons pouvoir améliorer la finesse globale de notre profilage ; augmenter les nombre de catégories définies peut également contribuer à ce but.

Références

(2014). RDF 1.1 concepts and abstract syntax. Technical report, World Wide Web Consortium.

Esr, I. (1998). *ESRI Shapefile Technical Description*. Environmental Systems Research Institute, Inc.

Noulas, A., S. Scellato, C. Mascolo, et M. Pontil (2011). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Rossmo, D. K. (2014). *Geographic Profiling*, pp. 1934–1942. New York, NY : Springer New York.

Snook, B., D. Canter, et C. Bennell (2002). Predicting the home location of serial offenders : a preliminary comparison of the accuracy of human judges with a geographic profiling system. *Behavioral Sciences & the Law 20*(1-2), 109–118.

Węcel, K. (2015). Linked geodata for profiling of telco users. *Studia Ekonomiczne 234*, 199–213.

Summary

Enabling anomaly detection in sensor networks is becoming a dominating issue for many application domains where the measured resources are precious. In the context of the WAVES FUI (Fonds Unique Interministériel) project, we are handling and analyzing data streams concerning the distribution of potable water. The associated processing imply different forms of reasoning. Some of them are based on the profiles of geographical sectors where the sensors are capturing information. In this paper, we present two complementary profiling methods that are using sensor measures as well as their network topology, external knowledge, f.ex. points of interest, and probabilistic inferences. The result of the combination is validated by domain experts through evaluations conducted on real-world data sets.

Une approche semi-automatique pour l'extraction d'informations liées aux itinéraires culturels à partir des réseaux sociaux : cas de la Via Francigena

Nathalie Valmond-Leblanc*, Eric Kergosien**
Natalia Grabar* Marta Severo**

*STL, Université de Lille
Domaine Universitaire Point de bois, BP 60149, 59653, Villeneuve d'Ascq, France
prenom.nom@univ-lille3.fr,
**GERiiCO, Université de Lille
Domaine Universitaire Point de bois, BP 60149, 59653, Villeneuve d'Ascq, France
prenom.nom@univ-lille3.fr

Résumé. Dans le cadre de l'analyse des itinéraires culturels, objet complexe du point de vue territorial, symbolique et social, le projet Itinéraire Numérique s'attache à analyser les contenus numériques disponibles en ligne pour la valorisation de l'itinéraire Via Francigena. Dans cet article, nous proposons une méthodologie semi-automatique pour l'extraction d'information spatiale liées à la Via Francigena à partir des données Instagram.

1 Introduction

Le fait que le patrimoine culturel soit disponible au format numérique modifie la manière de le comprendre et de le représenter. En particulier, cette abondance des données numériques a produit une nouvelle effervescence vers les possibilités de représentation du patrimoine. Des projets comme la plateforme de l'Institut Culturel de Google, le projet ORBIS de Stanford (Stanford Geospatial Network Model of the Roman World), ou les différents projets Europeana constituent de bons exemples d'une telle effervescence. L'objectif du projet « Itinéraire Numérique » [1] est de fédérer une approche interdisciplinaire pour étudier les modalités de représentation numérique du patrimoine culturel lié à l'itinéraire Via Francigena. Dans cet article nous décrivons une méthode semi-automatique pour extraire les entités spatiales (ES) à partir de messages Instagram. L'objectif à terme est de pouvoir proposer au grand public une interface dynamique permettant de découvrir la Via Francigena à travers les informations diffusées par des utilisateurs d'instagram (commentaires, photos, etc.). Nous traitons dans ces travaux des ES dites simples, aussi nommées entités spatiales absolues (ESA), à la différence des entités spatiales plus complexes intégrant des relations spatiales appelées entités spatiales relatives (ESR) Lesbegueries et al. (2006). La méthode combine une approche morpho-syntaxique à une approche lexicale s'appuyant sur la liste des étapes de l'itinéraire et sur Wikipedia pour identifier les ES dans les messages Instagram. En section 2, nous présentons un rapide état des

1. http ://www.itinerairesculturels.fr

travaux pour l'extraction des ES à partir de données de réseaux sociaux. En section 3, nous présentons notre méthodologie et les premiers résultats obtenus. Nous concluons en section 4 en précisant les perspectives.

2 Travaux connexes

Le travail de Pospescu et al. Pospecu et al. (2009) a pour objectif l'extraction automatique d'informations touristiques à partir de photos de voyages Flikr. Ils s'intéressent plus particulièrement à ce que les gens visitent, la durée de leur visite et les photos panoramiques. L'extraction des ES contenues dans les titres et les tags s'effectue selon une approche lexicale qui se base sur la construction d'un gazetier à partir de Wikipédia et qui associe à chaque nom de lieu ses coordonnées géographiques et son type. Mohamed et Oussalah (2014) présente une recherche similaire en se focalisant surtout sur la catégorisation des entités nommées (personnes, lieux et organisations) en se basant sur les Infobox de Wikipédia. D'autres méthodes permettant l'extraction d'ES telles que la méthode TEXT2GEO Tahrat et al. (2013) qui présente une approche hybride combinant des patterns linguistiques avec une approche d'apprentissage supervisé. La combinaison des deux méthodes améliore significativement la précision et le rappel des résultats. Cependant, le principal inconvénient de la méthode d'apprentissage supervisée réside dans l'exigence de données annotées manuellement, ce qui est plutôt une tâche coûteuse et complexe.

3 Travaux réalisés

Le corpus de départ est un ensemble de 7034 posts multilingues (italien, anglais et français), publiés entre septembre 2011 et janvier 2016, extraits du réseau social Instagram car mentionnant le tag « Via Francigena ». La méthodologie que nous appliquons dans nos travaux se décompose en 2 étapes : Extraction des étapes de la Via Francigena et des ES à partir de la ressource Wikipedia ; Désambiguisation des ES.

Pour **l'extraction des ES**, nous avons combiné une méthode morpho-syntaxique et une approche lexicale s'appuyant sur la liste des étapes de la Via Francigena et la liste d'ES générée à partir d'articles Wikipédia. Pour cela, nous avons sélectionné dans Wikipedia les articles dont les lieux sont géocalisés dans les régions que traverse la Via Francigena. Un article contenant différentes métadonnées (commune, région, province, coordonnées, description, lien de la page) est attaché à chaque ES de la liste obtenue. Le marquage du corpus Instagram avec la liste des étapes et les données Wikipédia permet d'obtenir une liste d'ES candidates à décrire la Via Francigena. Pour chaque ES candidate, une (voir deux) étiquette est associée pour préciser son type en nous appuyant sur la description et sur le nom de l'article provenant de Wikipedia. Nous avons défini pour cela plusieurs motifs s'appuyant sur des lexiques, afin d'identifier cinq types distincts (étapes, monuments, lieux religieux, lieux naturels, autres lieux).

La **désambiguïsation** s'applique aux ES identifiées à partir de la liste Wikipédia qui ne correspondent pas à une étape de l'itinéraire. Nous retenons une ES si elle est localisée dans une région, province ou commune par laquelle passe la Via Francigena. Les coordonnées géographiques extraites des données Wikipédia nous permettent de calculer la distance entre un lieu et une étape. Au regard des expérimentations menées sur le corpus visant à identifier la

distance adéquate entre les étapes de la Via Francigena et les ES candidates (5, 10, 15, 20, 30, 50, 100km), nous retenons les ES qui se trouvent à moins de 15 km d'une étape. Par exemple, l'ES « passo della cisa », identifiée dans 3 posts et caractérisée dans Wikipedia comme appartenant à une région par laquelle passe la Via Francigena, est identifiée comme pertinente car localisée à 6km de l'étape Berceto.

Des **analyses statistiques** résultent de ces premières expérimentations. Tout d'abord, à partir des 7034 posts de départ, 2598 contiennent une ou plusieurs ES pertinente(s). Aussi, 6186 occurrences d'ES sont identifiées, à savoir 3874 pour les étapes et 2312 pour les autres ES pertinentes. Les figures 1 et 2 mettent en avant les ES les plus fréquentes, respectivement les étapes et les ES provenant de Wikipedia.

Un **démonstrateur Web**, développé dans le cadre du projet [2], permet de (1) visualiser l'itinéraire culturel Via Francigena de façon dynamique ; et (2) afficher sur la carte les informations pertinentes relatives aux ES présentes dans Wikipedia. Les liens vers la description plus complète des lieux provenant de Wikipedia sont en cours d'intégration.

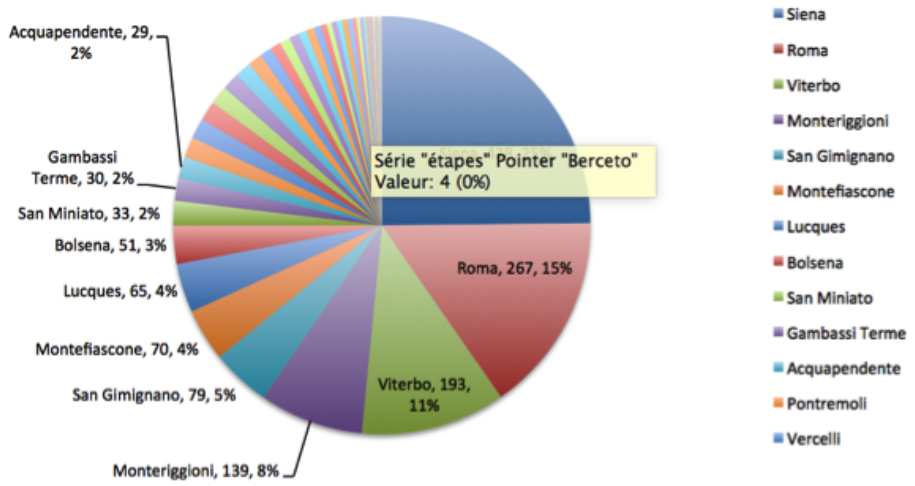


FIG. 1 – *Répartition nb occurrences étapes ViaFrancigena dans les posts Instagram.*

4 Conclusion et perspectives

Nous avons présentés les premiers résultats des travaux menés dans le cadre du projet Itinéraire Numérique visant à valoriser l'itinéraire Via Francigena. Nous remarquons, en analysant les informations relatives à l'itinéraire présentes dans les réseaux sociaux et notamment dans Instagram, que les internautes s'expriment sur le sujet (posts et photos associées). Le démonstrateur est en cours d'amélioration pour ajouter l'accès aux articles Wikipedia, et pour le rendre plus attractif. En perspectives, nous souhaitons étendre l'analyse à Tweeter puis Facebook.

2. Prochainement disponible à l'adresse suivante : http ://geriico-demo.univ-lille3.fr/ItineraireNumerique

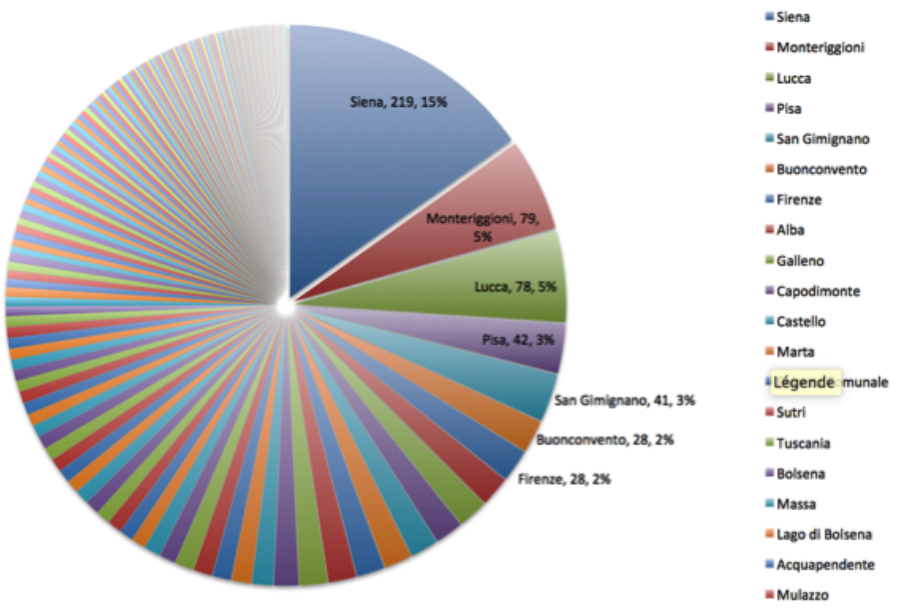


FIG. 2 – *Répartition nombre d'occurrences ES Wikipedia dans les posts Instagram.*

Références

Lesbegueries, J., C. Sallaberry, et M. Gaio (2006). Associating spatial patterns to text-units for summarizing geographic information. *ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop, LIUPPA*, 40–43.

Mohamed, A. et M. Oussalah (2014). Identifying and extract named entities from wikipedia database using entity infoboxes. *International Journal of Advanced Computer Science and Applications 5, n7*, 1713–1716.

Pospecu, M., G. Grefenstette, et P.-A. Moellic (2009). Mining tourist information from user-supplied collections. *CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management*, 1713–1716.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teiseire (2013). Text2geo : from textual data to geospatial information. *WIMS: International Conference on Web Intelligence, Mining and Semantics*, 4.

Summary

As part of the analysis of cultural routes, complex object of the symbolic, social and territorial point of view, the project Digital Route focuses on the analysis of digital content available online for the development of the Via Francigena route. In this article, we propose a semi-

automatic methodology for the extraction of spatial information related to Via Francigena from Instagram data.

Modèles de validation syntaxique et sémantique pour un stockage cohérent de données géo-spatiales imprécises dans les systèmes NoSQL document.

Besma KHALFI[*,***] Cyril De RUNZ[*,**] Sami FAIZ[***], Herman AKDAG[*]

[*]LIASD, université de Paris 8, Paris, France
khalfi@ai.univ-paris8.fr, akdag@ai.univ-paris8.fr
[**]CReSTIC, université de Reims Champagne-Ardenne, Reims, France
cyril.de-runz@univ-reims.fr
[***]LTSIRS, université de Tunis el Manar, Tunis, Tunisie
sami.faiz@insat.rnu.tn

Résumé. Avec l'explosion quantitative de données numériques et l'apparition du concept Big Data, le stockage de grandes masses de données géographiques trouvent les solutions NoSQL beaucoup plus efficaces en termes de performance et de temps de traitement. Cependant ces dernières n'offrent pas une cohérence des données stricte. En outre, si les données sont géo-spatiales et de nature imprécise, des complexités supplémentaires sont imposées à cause des caractéristiques structurelles et sémantiques plus complexes des données spatiales imprécises. Dans cet article, nous proposons une nouvelle méthodologie basée sur une validation syntaxique et sémantique des entités géo-spatiales floues avant d'être stockées dans des bases de données NoSQL orientées document.

1 Introduction

La croissance importante de données géo-spatiales soulève de gros problèmes de stockage et d'analyse de données et pose des défis supplémentaires dans la vérification de leur qualité (Li et al., 2016). Les données géo-spatiales étant volumineuses et hétérogènes, les imperfections les affectent encore plus quantitativement et qualitativement. Afin d'assurer une meilleure qualité de données géographiques, il est important d'étudier leur caractère imparfait et, par conséquent, de l'intégrer dans le processus d'analyse (Li et al., 2016).

Parmi les sources d'imperfection de données géo-spatiales, nous nous intéressons aux données imprécises pouvant être issues d'approximations, ou de sources subjectives (Desjardin et al., 2015). Basées sur la théorie des ensembles flous Zadeh (1965), de nombreuses approches ont été suggérées afin de modéliser l'imprécision dans les bases de données (Schneider, 2008).

Les systèmes NoSQL sont des systèmes non relationnels ; leurs capacités en termes de stockage, de passage à l'échelle et d'hétérogénéité des données dépassent celles des systèmes relationnels (Borkar et al., 2012). Cependant, les bases de données NoSQL ne fournissent pas une cohérence stricte des données et ne sont connues que pour les propriétés BASE [1]. Même si

1. Les propriétés BASE (Basically Availabe, Soft State, Eventuallty Consistent) des systèmes NoSQL.

l'absence du schéma est un atout, tout système doit faire quelques suppositions sur la structure de données et ajouter plus de code pour leur exploitation, ce qui implique un fort impact sur la qualité du traitement de données.

L'objectif de notre recherche est de fournir des modèles capables d'assurer la cohérence syntaxique et sémantique de données géo-spatiales imprécises en vue de pouvoir exploiter des données de qualité dans un environnement de Big Data.

Dans le reste de l'article, nous présenterons, dans un premier temps, les caractéristiques des données géo-spatiales floues en abordant les problèmes de cohérence dans les systèmes NoSQL. Ensuite, nous décrirons les solutions mises en place permettant de valider leurs cohérences structurelles et sémantiques. La troisième partie sera consacrée aux résultats d'expérimentations. Enfin, nous tirerons les conclusions et présenterons nos perspectives.

2 Contexte

2.1 Les données géo-spatiales imprécises

Les observations de l'espace révèlent la complexité et la variation de la réalité géographique. Contrairement aux espaces géographiques bien définis ayant des limites précises, nettes et linéaires, beaucoup d'autres situations dégagent des objets géographiques avec des limites plus ou moins nettes et continues. Chacun de ces objets présente un cœur où tous les éléments en son sein appartiennent pleinement au dît objet et des bordures qui lui appartiennent plus ou moins.

Par conséquent, la question « Comment caractériser un territoire imprécis ? » est une question liée aux outils de représentation en termes de modèles mathématiques et de modèles informatiques.

La théorie des ensembles flous est une des solutions mathématiques définies pour représenter l'imprécision. Grâce à elle, il est possible d'exprimer une appartenance partielle d'une valeur à un ensemble. Alors, si E est un ensemble flou et e est un élément de E, la proposition « e est un membre de E » n'est pas nécessairement soit vraie soit fausse. Elle peut être vraie que dans une certaine mesure. E est caractérisée par une fonction d'appartenance μ_E prenant ses valeurs dans [0, 1].

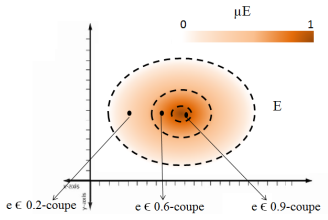


FIG. 1 – *Un ensemble flou E avec trois α-coupes.*

Pour stocker un ensemble flou, il est nécessaire de le discrétiser de sorte qu'un nombre fini de coupes doit être pris en considération afin de représenter l'ensemble flou (voir FIG. 1). Chaque coupe, que l'on appelle α-coupe, est l'ensemble des valeurs du domaine (l'ensemble

des e) ayant un degré d'appartenance supérieur ou égal à α ($\mu_E(e) \geq \alpha$). Une α-coupe de E est notée E_α. E_1 est appelé noyau ou cœur de E. Par convention,

— E_0, appelé support de E, est l'ensemble des éléments ayant un degré d'appartenance strictement supérieur à 0.
— un objet flou est un objet imprécis représenté par la théorie des ensembles flous.

2.2 La cohérence de données dans les systèmes NoSQL

Le maintien de la cohérence de données est une caractéristique essentielle des systèmes relationnels où le SGBD doit assurer une cohérence par rapport aux schémas à travers un contrôle de types et de structure et aussi une cohérence de données entre elles à travers un contrôle sémantique. Pour son maintien, le concept de règles d'intégrité (structurelles et non structurelles) est fondamental. Actuellement, les bases de données NoSQL sont le plus utilisés pour stocker de grandes masses de données grâce à l'ensemble d'outils et de langages de programmation fournis pour développer des applications sur ces nouveaux systèmes. Le problème d'incohérence de données constitue un défi pour ces systèmes car la conception sans avoir un modèle de données explicite et sans contraintes d'intégrité rend la cohérence de données très douteuse et le contrôle d'accès très complexe.

Compte tenu de la diversité significative de données dans les systèmes NoSQL, certains travaux ont mis en place des solutions de migration de systèmes relationnels vers les systèmes NoSQL mais qui ne semblent pas être une option réalisable à cause du coût d'implémentation élevé (Vitolo et al., 2015). Certaines autres méthodologies (Lotfy et al., 2016; Baker et al., 2011; Kanwar et al., 2013), ont proposé des améliorations pour conserver la cohérence et les propriétés ACID [2] des systèmes relationnels dans un environnement NoSQL. Ces solutions visent à fournir les propriétés ACID, mais ils n'abordent pas le problème de la cohérence de la structure de données lors du stockage.

En fait, les bases de données NoSQL sont utilisées aujourd'hui pour traiter les grandes masses de données géo-spatiales. Malheureusement, comme elles ne prennent pas en charge les transactions, l'intégrité de données et la durabilité, elles ne garantissent pas la qualité de données. La plupart d'entre elles n'exigent pas de contraintes structurelles sur les données et n'ont pas de schéma. Sans un schéma explicite, il est très difficile de s'assurer de la cohérence de données floues. De plus, l'accès et l'analyse de ces données sans avoir une certaine idée de leur schéma ou de leurs contraintes structurelles ne donnnent pas lieu à une analyse significative et intéressante.

Comprendre cette problématique implique de relever des défis pour le stockage et l'analyse des données. Dans la section suivante, nous présenterons notre solution pour gérer la cohérence des données géo-spatiales floues dans des systèmes non relationnels de type document.

3 Méthodologie de validation de données géo-spatiales floues

Nous travaillons avec le système NoSQL orienté document (Laxmaiah et Govardhan, 2013). Le modèle document offre un meilleur moyen que les modèles clé-valeur ou colonne ou graphe pour représenter des données complexes telles que les données géo-spatiales floues. Basés sur

2. Les propriétes ACID (Atomicity, Consistency, Isolation, Durability) des systèmes relationnels.

une approche d'agrégation, les systèmes document permettent d'améliorer les performances d'accès aux données et réduire le temps de latence.

Une base de données document est une collection de documents au format JSON. Ce format contient des informations structurelles implicites par lesquelles chaque document présente une structure d'un seul objet ou d'une collection d'objets. Les propriétés des objets sont des paires nom/valeur qui incluent des types de données de base, des valeurs multiples ou encore des types plus complexes (Crockford, 2006). GeoJSON est une extension de JSON pour l'encodage des données spatiales. Par rapport à d'autres vocabulaires connus tels que GeoSPARQL et NeoGeo, GeoJson est un vocabulaire compact disposant d'un modèle de données de base et prend en charge plusieurs types de géométrie. C'est un encodage léger, et moins verbeux que les encodages GML ou WKT utilisé par exemple par GeoSPARQL. En plus qu'il peut couvrir un grand nombre de cas d'utilisation.

Pour coder correctement la structure de données géo-spatiales floues, nous proposons premièrement le schéma *Fuzzy Geo-JSON*. Il est basé sur le format *GeoJSON* pour les données géo-spatiales (Butler et al., 2008). Fuzzy Geo-JSON étend GeoJSON afin de donner la définition de structures requises pour les différents types d'objets géo-spatiaux flous. Les caractéristiques sémantiques d'objets géo-spatiaux flous ne peuvent pas être contrôlées par le schéma Geo-JSON flou. Par conséquent, nous définissons dans un second lieu des modèles pour valider sémantiquement ces caractéristiques.

3.1 Le schéma Fuzzy GeoJSON

Le schéma $Fuzzy\ GeoJSON$ est basé sur les spécifications *json-schema-core* et *json-schema-validation*. Il spécifie les termes utilisés pour identifier les objets, les types de données et la structure de données définis dans l'IETF RFC 4627 (Crockford, 2006). FIG. 2-(a) montre la structure générale du schéma, certains détails ont été remplacés par le bloc $\{\cdots\}$. Fuzzy Geo-JSON garde la partie ***géométries*** comme elle est définie avec Geo-JSON. Cette partie donne les définitions de structure JSON des géométries de base à savoir le point, la ligne et le polygone. En se basant sur ces structures, la partie ***définitions*** donne le schéma des structures des objets géo-spatiaux flous.

FIG. 2 – *Le schéma Fuzzy GeoJSON pour les données géo-spatiales floues.*

Deux classes génériques sont fournies : ***fuzzyFeature*** et ***fuzzyFeatureCollection***. Le *fuzzyFeatureCollection* (FIG. 2-(b)) présente une collection de *fuzzyfeature*. Le *fuzzyFeature* (FIG.

2-(c)) définit tout type d'objet géo-spatial flou. Les objets géo-spatiaux flous peuvent être de structure géometrique simple :

— point flou (*fuzzyPoint*),
— polygone flou (*fuzzyPolygon*)
— ou ligne floue (*fuzzyLineString*).

Les objets géo-spatiaux flous peuvent être aussi de géométrie composite. Le schéma définit deux sous-types :

— les géometries composites homogènes : collection de points flous (*multiFuzzyPoint*), collection de polygones flous (*multiFuzzyPolygon*) et collection de lignes floues (*multiFuzzyLineString*)
— les géométries composites hétérogènes telles que les géométries alternatives, facultatives et multiples (*FuzzyAlternativeGeometry*, *FuzzyFacultativeGeometry* et *FuzzyMultipleGeometry*).

3.2 Les contraintes sémantiques

Étant donné un ensemble d'objets géo-spatiaux flous $O^E = \{O_1, \ldots, O_n\}$ où chaque objet flou O possède une géometrie $G_O \epsilon \{Point, Line, Polygon\}$ et est composé d'un ensemble fini de coupes $C^O = \{C_1, \ldots, C_m\}$. Chaque coupe C est définie par le couple (G_C, α_C) où $G_C \epsilon \{Point, Line, Polygon\}$ est la géometrie de la coupe et $\alpha_C \epsilon [0, 1]$ est le degré d'appartenance de la coupe C à l'objet flou O.

Une première contrainte sémantique est liée à la cohérence de chaque coupe C en fonction de son degré d'appartenance et de la géométrie de l'objet flou. Une coupe cohérente $C \epsilon C^O$, est définie par (O, $consistentCut^C$) où O est l'objet flou auquel appartient la coupe et $consistentCut^C$ la fonction qui vérifie que (1) pour un degré d'appartenance $\alpha_C = 1$, la géometrie de la coupe (G_C) doit être la même que celle de l'objet O ou (2) pour un degré d'appartenance $\alpha_C \neq 1$, la géometrie de la coupe G_C doit être de type $Polygon$ (voir algorithme 1).

Algorithme 1 Vérifier la cohérence des coupes

Requis : Initialiser avec O
  $G_O \leftarrow$ la géometrie de chaque objet O
  pour chaque C_i dans C^O **faire**
    $G_{C_i} \leftarrow$ la géometrie de C_i
    $\alpha_{C_i} \leftarrow$ le degré d'appartenance de C_i
    si $consistentCut(G_O, G_{C_i}, \alpha_{C_i})$ est FAUX **alors**
      retourner message d'erreur et SORTIR
    fin si
  fin pour
  retourner VRAI

La deuxieme contrainte sémantique est liée à la cohérence de l'ensemble de coupes définissant l'objet flou. Un objet flou cohérent doit avoir un ensemble de coupes connectées et normalisées. Un objet flou connecté et normalisé $O \epsilon O^E$, est défini par (C^O, $Norm^O$, $Inclus^O$) où C^O = $\{C_1, \ldots, C_m\}$ présente l'ensemble de coupes cohérentes de l'objet flou O. La fonction

$Norm^O$ vérifie que $\forall\ C\ \epsilon\ C^O$ avec le degré d'appartenance α_c, $Max(\alpha_c) = 1$. L'ensemble des coupes est donc normalisé. Pour chaque couple $\{C_i, C_k\}$ avec respectivement leurs types de géométrie et leurs degrés d'appartenance $\{(G_{Ci}, G_{Ck}), (\alpha_{Ci}, \alpha_{Ck})\}$, la fonction $Inclus^O$ vérifie la relation d'inclusion entre les coupes et vérifie que si $(\alpha_{Ci} > \alpha_{Ck})$ alors $(G_{Ci} \subset G_{Ck})$. Si l'ensemble des coupes n'est pas normalisé ou si la contrainte d'inclusion est violée, un message d'erreur est signalé à l'utilisateur (voir algorithme 2).

Algorithme 2 vérifier la cohérence de chaque objet

Requis : Initialiser avec O
  si $norm(O) \neq 1$ **alors**
    retourner message d'erreur et SORTIR
  sinon
    pour chaque C_i dans C^O **faire**
      $G_{C_i} \leftarrow$ la géometrie de la coupe
      $\alpha_{C_i} \leftarrow$ le degré d'appartenance de la coupe
      $RC \leftarrow$ le reste des coupes de l'objet C^O
      pour chaque C_k dans RC **faire**
        $G_{C_k} \leftarrow$ la géometrie de la coupe
        $\alpha_{C_k} \leftarrow$ le degré d'appartenance de la coupe
        si $Inclus((G_{C_i}, \alpha_{C_i}), (G_{C_k}, \alpha_{C_k}))$ est FAUX **alors**
          retourner message d'erreur et SORTIR
        fin si
      fin pour
    fin pour
  fin si
  retourner VRAI

4 Expérimentations

Les principes de la méthodologie de validation sont implémentés à travers un prototype développé en Java où l'utilisateur vérifie l'ensemble de données floues syntaxiquement contre le schéma Fuzzy GeoJSON et sémantiquement contre un ensemble de fonctions développées vérifiant les contraintes sémantiques. Dans la suite, quelques résultats d'évaluation sont donnés.

Un premier exemple de validation syntaxique génère un message d'erreur dans le fichier de journalisation est illustré par FIG. 3. Il affiche une erreur liée à une incohérence de structure d'une coupe. On suppose qu'un polygone flou ne doit pas avoir que des coupes avec des géométries de type Polygone. Dans l'ensemble de coupes, une est détectée avec un type de géométrie point.

Un deuxième exemple de validation sémantqiue est présenté par FIG. 4. Un message d'erreur lié à des coupes non connexes est affiché. Deux objets flous ne forment pas des ensembles flous connexes. Les coupes pour chaque objet doivent être incluses entre-elles selon la valeur du degré d'appartenance.

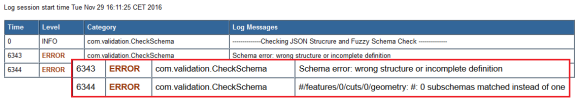


FIG. 3 – *Detection d'une erreur liée à la géometrie d'une coupe.*

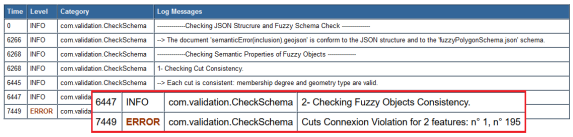


FIG. 4 – *Detection d'erreur de connection.*

5 Conclusion

La simplification de modèles géo-spatiaux flous ou leur absence entraîne la propagation des erreurs, ou même leur augmentation. Nous avons proposé une méthodologie basée sur un schéma spécifique appelé *Fuzzy GeoJSON* augmenté d'un ensemble de contraintes sémantiques que les données doivent valider pour être stockées correctement dans les systèmes NoSQL document.

Cette méthodologie de validation est une première étape dans l'objectif de fournir un cadre complet de gestion du Big Geo Data flou en offrant d'autres possibilités d'interrogation, d'exploration et de visualisation. Notre prochaine direction de recherche consiste à interroger les grandes masses de données geo-spatiales floues sur les relations topologiques.

Références

Baker, J., C. Bond, J. C. Corbett, J. Furman, A. Khorlin, J. Larson, J.-M. Leon, Y. Li, A. Lloyd, et V. Yushprakh (2011). Megastore : providing scalable, highly available storage for interactive services. In *Conference on Innovative Database Research (CIDR)*, Volume 11, pp. 223–234.

Borkar, V. R., M. J. Carey, et C. Li (2012). Big data platforms : what's next ? *ACM 19*(1), 44–49.

Butler, H., M. Daly, A. Doyle, S. Gillies, T. Schaub, et C. Schmidt (2008). The GeoJSON format specification. Technical report.

Crockford, D. (2006). The application/json media type for JavaScript Object Notation (JSON). Technical report.

Desjardin, E., B. Lefebvre, et C. D. Runz (2015). Intégration de l'imperfection de l'information dans les dynamiques spatiales. définitions, outils et exemples. *Revue Internationale de Géomatique 25*(3), 437–463.

Kanwar, R., P. Trivedi, et K. Singh (2013). NoSQL, a solution for distributed database management system. *International Journal of Computer Applications (IJCA) 67*(2).

Laxmaiah, M. et A. Govardhan (2013). A conceptual metadata framework for spatial data warehouse. *International Journal of Data Mining & Knowledge Management Process 3*(3), 63–73.

Li, S., S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, et T. Cheng (2016). Geospatial big data handling theory and methods : A review and research challenges. *Journal of Photogrammetry and Remote Sensing (JPRS) 115*, 119 – 133.

Lotfy, A. E., A. I. Saleh, H. A. El-Ghareeb, et H. A. Ali (2016). A middle layer solution to support ACID properties for NoSQL databases. *Journal of King Saud University-Computer and Information Sciences 28*(1), 133–145.

Schneider, M. (2008). Fuzzy spatial data types for spatial uncertainty management in databases. In *Handbook of Research on Fuzzy Information Processing in Databases*, pp. 490–515. Information Science Reference.

Vitolo, C., Y. Elkhatib, D. Reusser, C. J. Macleod, et W. Buytaert (2015). Web technologies for environmental big data. *Environmental Modelling & Software 63*, 185–198.

Zadeh, L. A. (1965). Information and control. *Fuzzy sets 8*(3), 338–353.

Summary

With the quantitative explosion of digital data and the appearance of the Big Data concept, as relational database are not efficient, NoSQL databases are a good solution in the data storing and retrieval.

However, NoSQL databases do not offer strict data consistency. In addition, if data is geo-spatial and imprecise, additional complexities appear because of the complex syntax and semantic features of such data.

The paper presents a new methodology based on a syntax and semantic validation of the fuzzy geo-spatial entities before being stored in the document-based NoSQL databases.

Approche pour l'élaboration d'un modèle chronotopique urbain 2

Alain Guez*, Claire Lagesse**, Francis Rousseaux***

*LAA/LAVUE UMR CNRS 7218
118-130 avenue Jean-Jaurès
75019 Paris
guezal@wanadoo.fr
http://www.laa.archi.fr/

**Université Bourgogne - Franche-Comté
ThéMA UMR 6049, CNRS
UFR Lettres SHS, 32 rue Mégevand
25030 Besançon Cedex
claire.lagesse@univ-fcomte.fr
http://thema.univ-fcomte.fr/le-laboratoire-thema/

***URCA (Université Reims Champagne Ardenne)
UFR SEN, Moulin de la Housse
51100 Reims
francis.rousseaux@univ-reims.fr
http://www.univ-reims.fr

Résumé.
A partir des premières hypothèses de travail construites pour la projet *Exploration chronotopique d'un territoire parisien*, l'équipe de recherche s'est élargie à de nouvelles disciplines apparues comme nécessaires pour explorer les matériaux identifiées et les interpréter avec des outils adaptés. Deux terrains sont menés conjointement : une ethnographie visant à comprendre comment se constitue l'expérience du temps dans la ville contemporaine, et une approche « objectivante » cherchant à identifier les données et les indicateurs pertinents pour révéler les assises chronotopiques sous-jacentes à certaines expériences du temps vécu dans les lieux habités. Les données horaires et calendaires recueillies au cours d'une première phase de recherche ont pu être structurées en un modèle de base de données permettant d'une part un traitement informatique visant à repérer des *clusters* horaires et calendaires et, dans un deuxième temps, à interpréter, à travers une approche géomatique, ces *clusters* dans l'espace de la ville pour en affiner la lecture et construire une analyse croisée avec les résultats de l'ethnographie.

1 Chronotopie et expérience des temps urbains

La recherche en cours présentée ici, a été engagée dans le cadre de l'appel à projet Paris 2030 et s'intitule *Exploration chronotopique d'un territoire parisien*[1]. Cette recherche, a pu être exposée dans son état d'avancement de l'année dernière au cours des ateliers Gast 16[2]. Il s'agit d'explorer conjointement deux questions liées à l'expérience et aux assises spatiales et temporelles de la ville contemporaine : comment se forme l'expérience du temps dans une ville comme Paris ? Comment définir un chronotope et quel rôle peut-il jouer dans l'expérience temporelle de la ville ? Ces questions croisent d'emblée des approches différentes et articulées qui nécessitent de mobiliser des compétences et des disciplines variées pour tenter d'y répondre.

A travers l'approche ethnographique nous avons fait émerger des dimensions anthropologiques qui se déploient, dans des représentations et des pratiques signifiantes, à l'articulation de différentes échelles et horizons temporels, entre les biographies et la vie quotidienne, dans des territoires en transformation.

Par l'ancrage disciplinaire (entre anthropologie, urbanisme, paysage et architecture) et par positionnement culturel, nous abordons les temps des territoires à travers le vécu des habitants résidents et temporaires[3].

Ce positionnement a pour objectif de travailler au déploiement des temporalités territoriales signifiantes dans l'expérience humaine et à partir de laquelle il s'agit de repérer les faits et phénomènes participant de cette expérience dans les lieux. Comment l'expérience du temps est-elle liée à des « assises chronotopiques» observables, objectivables et mesurables ?

Il s'agit alors de chercher à saisir cette expérience et à la rendre lisible pour croiser ces données avec d'autres indicateurs territoriaux. Cette approche est chronotopique dans la me-

[1] Recherche nancée par la Mairie de Paris dans le cadre de l'appel à projet Paris 2030, et entamé depuis janvier 2015 par une équipe de quatre chercheurs du Laboratoire Architecture Anthropologie (Lavue UMR CNRS 7218) : Alain Guez, architecte-urbaniste, Professeur à l'Ensan – membre du LaaLavue UMR CNRS 7218 (porteur de projet responsable scienti que de la recherche) ; Alessia de Biase (architecte, docteur en anthropologie), directrice du Laa ; Federica Gatta (architecte, docteur en urbanisme) maître de conférence à Institut d'Urbanisme de Grenoble / Université Grenoble Alpes. Chercheur au Laboratoire Pacte UMR CNRS 5194 / associé au Laa-Lavue UMR CNRS 7218) , Piero Zanini (architecte, docteur en anthropologie), maître assistant associé à l'Ensaplv, chercheur au Laa. A cette première équipe originelle ont été associés Stefano Stabilini, chercheur de la Faculté d'architecture du Politecnico di Milano, expert de politiques temporelles urbaines italiennes, et plus récemment Francis Rousseaux, professeur en informatique à l'Université de Reims et chercheur associé à l'Ircam et depuis n 2015, Claire Lagesse, (géomaticienne, docteur en physique) maître de conférence à l'Université Bourgogne – Franche-Comté, membre du laboratoire ThéMA UMR 6049, CNRS.

[2] Guez, A., Rousseaux, F. (2016), « Approche pour l'élaboration d'un modèle chronotopique urbain », in Actes de l'atelier GAST – Gestion et Analyse de données Spatiales et Temporelles, Atelier inscrit dans la 16ème conférence internationale francophone sur l'extraction et la gestion des connaissances EGC 2016, sur coordonné par Éric Kergosien, Thomas Guyet, Christian Sallaberry, Reims, 19 janvier 2016, http://gt-gast.irisa.fr/gast-2016/

[3] Martinotti, G. (1993), Metropoli. La nuova morfologia sociale della città, Bologna : Il Mulino.

sure où elle cherche systématiquement à lier entre-elles les dimensions spatiales et temporelles des territoires habités.

A partir des trois principaux terrains[4] menés au cours des dix dernières années, apparaissent des questions récurrentes qui se posent en plusieurs termes :

- l'accord entre un mode de vie souhaité et les possibilités de l'environnement territorial ;
- les promesses et l'écart entre l'horizon d'un projet à venir et l'entre-temps de sa réalisation, vécu au quotidien ;
- les repères dans un environnement socialement et physiquement changeant ;
- la liberté, par rapport aux possibilités de faire des choix dans un environnement contraint.

Cette première synthèse des résultats obtenus au cours des différentes recherches menées, nous invite à nous questionner sur comment prolonger ce parcours de recherche, comment mobiliser d'autres savoirs qui pourront participer à repérer des relations entre expériences et assises chronotopiques territoriales.

De nombreuses données sont aujourd'hui disponibles pour temporaliser les territoires ainsi que des outils permettant leur interprétation et leur spatialisation. Ceci étant, dans l'approche que nous suivons, il existe un hiatus entre la complexité de l'expérience habitante et les données et outils mobilisables pour en rendre compte. Le récit habitant est une matière d'une grande richesse pour explorer l'épaisseur temporelle d'un territoire et en décrire et en déployer l'assise chronotopique.

Dans cette perspective, il nous semble important de placer cette approche anthropologique au centre de notre réflexion pour aborder l'observation et le projet territorial et de mobiliser les savoirs et savoir-faire des disciplines de projet qui intègrent ces multiples dimensions des temporalités territoriales.

C'est sur ces acquis de recherche que s'appuie la présente contribution qui vise à préciser les possibilités apparues à travers l'élargissement de notre équipe de recherche qui intègre désormais des compétences en informatique et en géomatique.

2 Structuration d'un modèle de base de données

Dans un premier temps de la recherche il s'est agi d'identifier des sources d'informations multiples permettant de constituer un corpus de données horaires et calendaires géoréférencées et qui couvrent des périmètres différents (voir figure 1 ci-après). Ces données ont été soit fournies par des opérateurs qui les constituent et les mettent à jour pour une exploitation commerciale[5], soit relevés par notre équipe afin d'avoir des références fiables et qui nous permettent également de vérifier les données qui nous ont été fournies par les opérateurs:

[4] (2006-2008) *Les réenchantements de La Courneuve. De l'ancienne cité des 4000 sud aux futurs quartiers de la Tour et Les Clos, (*2010-2012) *Qualifier la transformation ou comment se projette l'idée de qualité de vie dans le futur du Grand Paris,* (2015-2017) *Exploration chronotopique d'un territoire parisien.*

[5] Les données horaires et calendaires peuvent être très volatile avec une obsolescence relative aux évolutions urbaines et démographiques notamment. Les opérateurs qui recueillent ces données avec une visée commerciale ont intérêt à fiabiliser leurs bases en élaborant des stratégies de mise à jour au plus prêt de la réalité et de leur actualité.

- Des relevés horaires et calendaires élaborées par les membres de notre équipe et couvrant une partie de la tranche du territoire étudié – 1,3 km de large sur 3 km de long – et correspondant à environ 3 000 activités[6] ;
- Les horaires et calendriers relevés par notre équipe ont été associés à la base géolocalisée des activités de rez-de-chaussée, organisée par l'Atelier Parisien d'Urbanisme (APUR) qui a mis à notre disposition une extraction de ses données – sur le périmètre parisien de notre tranche de territoire – dans le cadre d'une convention ;
- Des horaires d'ouverture journalières sur la semaine, correspondant à un ensemble d'activités de rez-de-chaussée, couvrant toute l'étendue de la tranche de territoire étudié et comprenant environ 2 000 activités relevées[7] ;
- Des horaires d'ouverture journalières sur la semaine correspondant à des activités de rez-de-chaussée sur l'ensemble des arrondissements touchant le territoire étudié, ainsi que sur les communes d'Aubervilliers et Ivry-sur-Seine et comprenant environ 100 000 activités relevées dont 20 000 environ comportent des données horaires et calendaires à l'échelle de la semaine[8].

Il est important de souligner que ces données ont été collectées sur une même période – entre juin et octobre 2015 et qu'elles sont contemporaines. Ceci étant, nous savons aussi que les horaires évoluent avec l'environnement urbain, notamment dans les secteurs en transformation et que la variation de ces derniers ouvre de nouvelles questions et possibilités de recherche notamment en travaillant à des « coupes » synchrones permettant aussi de faire apparaître les permanences et variations horaires et calendaires des différentes parties de la ville étudiée.

À partir de ces sources, certains traitements ont pu être élaborés avec les outils intégrés dans le Système d'Information Géographique dont nous nous sommes dotés[9] et dont nous avons présenté les premiers résultats lors du précédent Atelier Gast 16[10].

[6] Ces relevés ont été e ectués par des stagiaires de notre Laboratoire entre avril et juin 2015.

[7] Ces données ont été fournies par LesHoraires.fr qui les a relevés et qui a mis à notre disposition une extraction de sa base.

[8] Ces données ont été fournies par Les Pages Jaunes qui les a relevés et qui a mis à notre disposition une extraction de sa base.

[9] Le logiciel Arcgis™ a été choisi en accord avec notre partenaire de la faculté d'Architecture du Politecnico di Milano.

[10] Guez, A., Rousseaux, F. (2016), op. cit

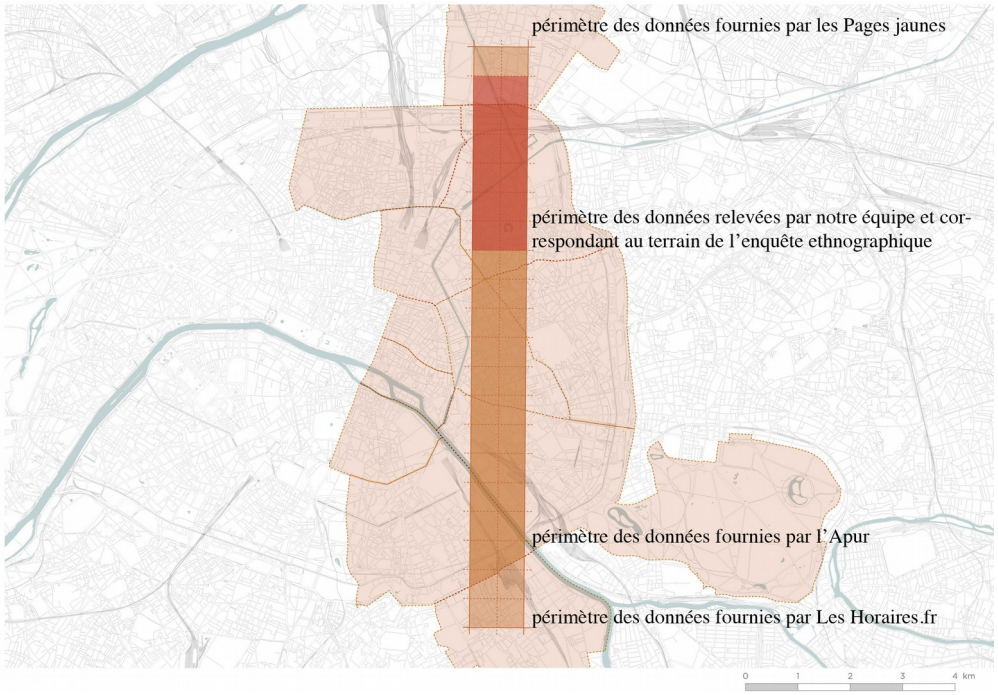


FIG 1 - Périmètre des différentes bases de données géolocalisées et datées des activités de rez-de-chaussée de la ville.

Afin de mieux exploiter ces données, l'équipe élargie, associant désormais des compétences en informatique et en géomatique, a élaboré conjointement un modèle de base de données dans la perspective d'une exploitation visant différents objectifs.

Ce modèle a été construit afin d'intégrer l'ensemble des données recueillies dans une même base (gérée avec le système de gestion de base de données PostGres/PostGIS). Les entités ainsi que les associations qu'elle regroupe sont détaillées dans la figure 2 ci-après. La présence des adresses, points géolocalisés, pour les bâtiments (abritant des commerces et des services), les marchés, les stations de métro et les terrasses nous permet de donner à notre étude une dimension géographique qui ouvre vers une analyse spatio-temporelle. Afin de la rendre possible, il est nécessaire de distinguer les activités (commerces et services) pour lesquels nous possédons les horaires de ceux pour lesquels l'information manque. Ces derniers portent néanmoins une indication intéressante : une activité d'un certain type existe en un point donné. Nous les conservons donc dans notre recherche.

Afin d'analyser plus finement les rythmes horaires, nous nous sommes concentrés sur les données issues des Pages Jaunes. Ces données nous permettent de mettre en relation les horaires, les types d'activité ainsi que les positions géographiques des commerces et services. Afin de pouvoir exploiter ces informations, nous avons dû regrouper les activités (aux intitu-

lés divers) en catégories, selon leur code APE[11]. Cette classification, données par l'INSEE, nous permet d'apprécier la répartition des types activités sur le territoire. Il existe plusieurs regroupements d'activités par exemple, en 22 catégories, dont 14 représentées sur l'espace que nous étudions. Cette classification agrège l'information et, ainsi, diminue le niveau de détail de nos données. Elle est néanmoins nécessaire si nous voulons étudier les activités par types, car il existe 375 codes APE différents pour les activités dont nous avons les horaires. Pour étudier plus finement certaines catégories d'activités, nous extrayons leurs codes APE particuliers. Ainsi, pour travailler sur les activités de restauration utiles pour le repas du midi, nous avons isolé les codes APE correspondants aux boulangeries, sandwicheries, restauration rapide, restauration traditionnelle et débit de boissons.

Nous présentons l'ensemble des données que nous avons recueillies sous la forme d'une modélisation HBDS (voir figure 2 ci-après). Cette conceptualisation (Hypergraph Based Data Structure) permet de synthétiser les entités, leurs attributs et les liens d'associations existants entre elles. Nous l'avons choisie car elle se prête idéalement à une analyse de construction de projet, où il est nécessaire d'avoir une vision globale des données à disposition, du détail auquel il est possible d'accéder et de l'évolutivité de cet ensemble.

Moins technique qu'un diagramme de classes UML (Unified Modelling Langage), qui comprendra également les types d'attributs, les clés primaires et étrangères, et éventuellement quelques fonctions implémentées pour l'analyse, le diagramme HBDS reflète moins l'état exact de la base de données que toutes les possibilités qu'elle peut offrir. Il n'est donc pas réalisé pour illustrer une analyse en particulier, mais pour apporter une vision d'ensemble de toutes celles qu'il serait possible de réaliser.

[11] Des données existent également sur les équipements publics comme par exemple les écoles dont les rythmes sont spécifiques et qui impactent l'espace public à des moments particuliers d'entrée et sortie de classe, comme cela ressort des témoignages recueillis à travers les entretiens ethnographiques. Ces données pourront être également mobilisées afin de mettre en évidence les variations d'intensité de présence dans l'espace public.

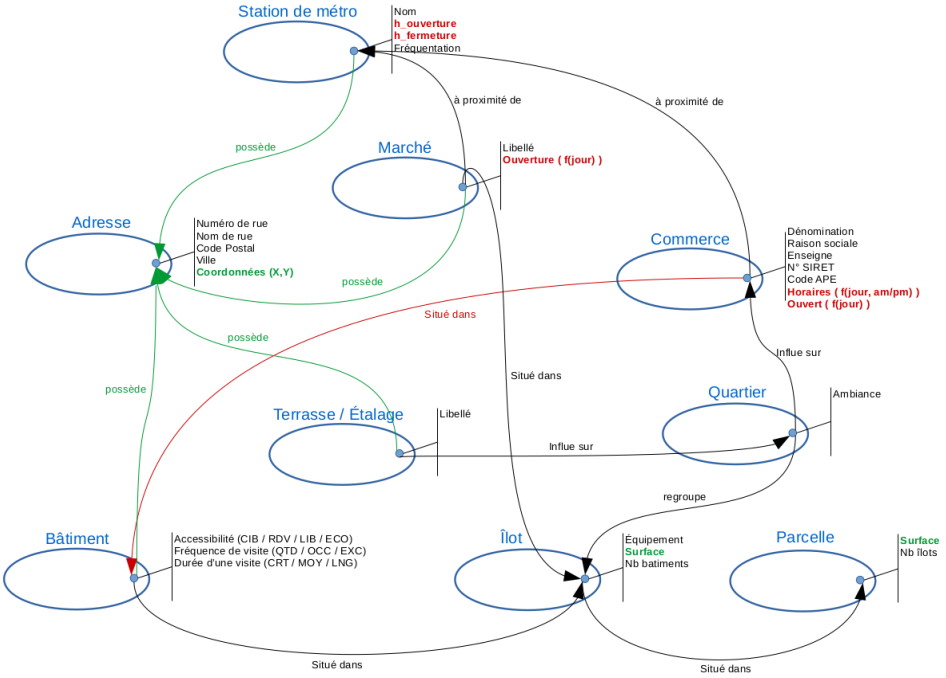


FIG 2 - Modèle conceptuel de données (HBDS).

3 L'identification de chronotypes des activités de la ville

Nous avons pu identifier, dans un premier temps, en exploitant les données horaires et calendaires géoréférencées et avec le SIG mobilisé, un certain nombre de caractéristiques spatio-temporelles du territoire étudié. Nous avons dénommé ces caractéristiques des chronotypes des activités de la ville[12] que nous distinguons des chronotopes.

Les chronotypes des activités de la ville sont des ensembles urbains dont les assises temporelles présentent, dans un périmètre à dimensionner, des caractéristiques horaires et calendaires semblables. En d'autres termes, ils se fondent sur l'hypothèse qu'on peut distinguer des ensembles d'activités, proches les unes des autres, qui ont des comportements temporels relativement semblables et sur des périmètres circonscrits. Par exemple, des ensembles plutôt du matin, d'autres plutôt du soir, ou encore des ensembles actifs le week-end et d'autres pas ou moins intenses pendant cette même période. Le SIG permet de spatialiser les données horaires et calendaires et de cartographier des périodes choisies afin de repérer ces chronotypes. Cet outil permet, par approximations successives, de repérer des pas temporels pertinents pour le territoire étudié.

Il est important de préciser que ces chronotypes des activités ne se superposent pas totalement aux chronotopes qui représentent eux, des ensembles plus complexes qui participent à structurer l'expérience du temps que l'on fait dans la ville, cette dernière étant explorée à travers l'ethnographie.

Afin d'identifier plus systématiquement les chronotypes des activités du territoire étudié, une première hypothèse de repérage de groupes ou *cluster* d'activité a été formulée. Cette hypothèse a été explorée à travers un traitement informatique des données horaires et calendaires et a permis de regrouper, parmi les activités de rez-de-chaussée présentes sur le périmètre des données Pages Jaunes, entre 55 et 60 *clusters* horaires pour un jour donné[13].

Les clusters sont déterminés en utilisant l'algorithme des K-means. Cet algorithme considère un espace de n dimensions, dans lequel sont représentés les points de l'étude. Dans notre cas n = 4, et les points de l'étude sont les différents commerces et services. Les quatre dimensions que nous retenons sont : l'heure d'ouverture, l'heure de fermeture, l'heure du début de la pause de midi et l'heure de fin de la pause de midi. Une fois l'espace défini, l'algorithme positionne K points-marqueurs aléatoirement. Le nombre K est déterminé par expérimentations successives, pour avoir des *clusters* aux profils homogènes. Nous avons procédé à une première expérimentation sur un échantillon limité de données, en nous restreignant à un nombre de *clusters* horaires compris entre 30 et 35.

Les K points-marqueurs étant positionnés dans notre espace à quatre dimensions, l'algorithme calcule pour chaque commerce et service le marqueur qui en est le plus proche. Une fois ce travail fait, le marqueur est déplacé pour se positionner à la position moyenne des ac-

[12] Comme précisé dans la note 11, notre analyse s'est pour le moment limitée aux commerces et services qui, dans une ville comme Paris, qualifient fortement l'espace public.

[13] Ce travail a été e ectué par Rémi Bardin, étudiant en master à l'UFR Sciences Exactes et Naturelles de l'Université de Reims Champagne-Ardenne, dans le cadre de son stage, qui s'est déroulé entre avril et juin 2016, sous la direction conjointe de Francis Rousseaux et de Alain Guez.

tivités qui lui ont été rattachées. L'opération est ensuite répétée jusqu'à ce qu'un état de stabilité soit atteint.

Par expérimentations successives, nous déterminons des clusters de sorte à ce qu'ils regroupent des catégories horaires pertinentes. Nous obtenons ainsi entre 55 et 60 *clusters* d'activités dont les rythmes horaires sont proches. Les calculs ont été réalisés sur une journée, le mardi (jour d'ouverture « normale » des commerces et services). Il serait toutefois intéressant d'étendre cette analyse aux rythmes calendaires, en rajoutant des dimensions à l'espace de l'algorithme.

Si nous voulons réaliser une cartographie de ces clusters, il est nécessaire de descendre leur nombre à 7, afin que l'œil humain puisse les différencier sur la carte. Cependant, en forçant leurs regroupements, les horaires des activités associés ne sont plus cohérents, et la clusterisation devient discutable. Il est donc nécessaire d'imaginer d'autres façons d'exploiter cartographiquement cette information, en restant sur une discrétisation de 55 à 60 clusters. Ceci constitue un des prochains objectifs de notre recherche.

Les premiers résultats obtenus ne sont pas spatialisés, mais peuvent l'être dans la mesure où les activités de la ville sont caractérisées par des attributs horaires et calendaires et également par des coordonnées géographiques et des caractéristiques qualitatives.

Par exemple, on peut croiser les résultats « clusterisés » avec des types d'activité afin de tenter de répondre aux questions suivantes : les boulangeries ou les cafés ont-ils tous les mêmes horaires - appartiennent-ils tous au même *cluster* sur l'ensemble du territoire étudié ? En d'autres, termes, y a-t-il une corrélation entre un type d'activité et des horaires et calendriers types ? Ou encore, si tel n'est pas le cas, les boulangeries et les cafés d'un même *cluster* horaire et calendaire sont-ils regroupés dans des ensembles urbains continues, identifiables et dimensionnables ? Ces questions sont notamment issues de premières observations faites grâce aux cartographies élaborées avec le SIG, outil avec lequel il est par exemple possible de repérer, sur l'avenue de Flandres à Paris, une synchronocité, localisée, entre les horaires d'ouverture matinale des cafés et ceux des stations de métro[14].

Plus généralement, l'étude des corrélations croisées entre type d'activité, rythme horaire et position géographique pourrait nous amener à observer la *qualité* et le *rythme* d'ensembles urbains présentant des caractéristiques horaires et calendaires semblables. Nous recherchons d'éventuelles synchronicités ou a-synchronicités sur le terrain étudié pour les ouvertures / fermetures (qui se répondent potentiellement). Quant aux types d'activités, il serait intéressant d'observer les critères d'attraction (spécialisation de certains quartiers) ou de répulsion (homogénéisation dans l'espace pour être dans une concurrence plus lointaine). Un de nos objectifs est ainsi l'analyse de la continuité d'un service tant au niveau temporel qu'à celui spatial.

[14] Guez, A., Rousseaux, F. (2016), op. cit

4 Croiser l'approche ethnographique et les assises chronotypiques identifiables à travers la base de données

Le parcours de recherche que nous présentons ici regroupe différentes approches qui ne sont pas toutes simultanées et nous ont posé la question de l'articulation et du séquençage dans le temps de ces dernières. En effet, chacune des approches développées nécessite un certain temps de préparation, de traitement, d'interprétation. Nous avions fait l'hypothèse, ambitieuse, de travailler dans un premier temps sur les chronotypes des activités de la ville et de focaliser, à partir de premiers résultats, notre attention et notre ethnographie sur des questions apparaissant à travers cette première approche factuelle. Cette hypothèse ne s'est pas totalement vérifiée, même si nous pouvons mesurer aujourd'hui l'intérêt d'engager dès le début de la recherche l'exploration des chronotypes notamment pour identifier les données utiles, leur existence et les sources de données, contacter les opérateurs détenteurs de ces données et négocier avec eux leur mise à disposition pour la recherche (les données horaires n'étant pas à ce moment de la recherche en *open data)*, tester des outils et des technologies, identifier des partenaires scientifiques et organiser le terrain ethnographique.

Ce processus ne peut pas être totalement et précisément planifier a priori dans un calendrier de recherche. Conscients de cette difficulté, nous avons décidé d'engager parallèlement plusieurs tâches et de travailler celles-ci en temps masqué, ce qui permet de gagner du temps par rapport à une démarche qui aurait privilégié la succession de ces dernières.

Si nous sommes à présent en mesure de caractériser des ensembles d'activités urbaines en fonction de leurs horaires, calendrier hebdomadaires et type (service, commerces,…), il est important de territorialiser ces informations et notamment pour essayer de répondre à des questions relatives à l'étendue, au dimensionnement et aux limites spatiales de ces chronotypes d'activité : les chronotypes ont-ils une dimension ? Avec quel grain peut-on les repérer ? Sont-ils multiscalaires ?

Ces questions sont directement issues de l'ethnographie qui fait apparaître que certains lieux identifiables peuvent être caractérisés comme des chronotypes (par exemple un marché), voire même comme des chronotopes (par exemple le bassin de La Villette) et que Paris, la ville entière, est parfois décrite comme un chronotope, souvent en la comparant à d'autres villes, par exemple en ce qui concerne la vie nocturne (en définissant la singularité de Paris par rapport à Berlin ou à Barcelone) ou en ce qui concerne la désertification lors des vacances scolaires (en comparant par exemple Paris à Lyon).

Les caractéristiques des lieux urbains émergents des récits des habitants constituent ce que l'on peut considérer comme un premier modèle descriptif des chronotypes et des chronotopes. Sur cette base, nous pouvons poursuivre dans deux directions fertiles pour la recherche. D'une part en systématisant le modèle descriptif des chronotypes et des chronotopes, issu de l'analyse ethnographique et relatifs à un territoire donné, et, d'autre part, en cherchant les données et les indicateurs pertinents pour prendre la mesure et rendre lisibles les assises chronotypiques et chronotopiques sous-jacentes à ces récits ethnographiques.

L'ethnographie a permis de repérer des éléments caractéristiques des chronotopes urbains dont certains ont probablement un lien étroit et plus ou moins direct avec les chronotypes des activités et nous faisons l'hypothèse que l'outillage et les modèles de données préparés depuis le début de la recherche doivent pouvoir nous aider à le révéler.

Les matériaux ethnographiques ont permis de thématiser les récits des habitants en faisant émerger six grandes familles d'indicateurs significatifs pour l'expérience des temps urbains : les temps structurants, ménager son propre temps, proximité et mouvements temporalisés, repères chronotopiques, les temps de la transformation, Paris comme chronotope.

Parmi ces indicateurs significatifs, nous avons repéré quelques singularités qui peuvent nous aider à vérifier l'hypothèse de la relation entre expérience du temps et caractéristiques chronotypiques des activités urbaines. Notamment des périodes particulières bien identifiées dans les témoignages et qui constituent des chronotypes, voire même des chrontopes. Pour avancer dans l'exploration, nous nous proposons de tester la lisibilité de trois périodes que sont : la nuit, le week-end, la pause midi en semaine.

Certains secteurs présentent des variations importantes entre l'activité en semaine et le week-end. Ces variations sont perçues comme des renversements de l'espace urbain entre deux configurations qui peuvent être diamétralement opposées.

Au sein même du week-end, des variations existent entre le samedi et le dimanche et notamment en ce qui concerne l'offre commerciale qui, dans certains secteurs par exemple du XIXème arrondissement de Paris, présentent une inversion entre fermeture le samedi ou le dimanche du fait d'une importante communauté israélite dont le jour de congé hebdomadaire est le samedi et non le dimanche.

De la même manière, les quartiers touristiques du centre de la capitale sont clairement identifiés pour leur activité décalée par rapport aux rythmes de la vie quotidienne et notamment par les ouvertures le dimanche, rendues possibles par les pouvoirs publics justement pour les secteurs définis comme zones touristiques internationales[15].

A l'échelle infra-hebdomadaire un moment particulier à la vie parisienne est clairement exprimé : il s'agit de la pause midi qui est culturellement installée dans la vie professionnelle parisienne et probablement plus marquée dans certains secteurs caractérisés par une forte proportion de locaux professionnels - notamment par des activités professionnelles (professions libérales et/ou créatives) dont l'étendue sur une grande amplitude horaire justifie la pause méridienne qui présente une importance plus significative. Intégrée au temps du travail, la pause midi émerge ainsi des entretiens comme un moment très particulier à la fois structurant et plastique de la journée, sensible tant au contexte, en termes d'offre, qu'a une série de contraintes temporelles, économiques, ou liées aux modes de vie individuels, qui la chargent de sens.

L'atelier Parisien d'URbanisme à établie une base cartographique qu'il sera intéressant de croiser avec nos propres données : la cartographie des destinations qui établie, à la parcelle, la mixité entre activité et logement, et distingue trois familles de destination de parcelles : Forte dominante Activité (3/4 et plus) ; Forte dominante Logement (3/4 et plus) ; Autre mixité Logements/Activités. Il sera intéressant de vérifier par exemple comment l'offre de restauration s'organise en fonction de cette caractérisation de la mixité des parcelles : les secteurs plus fortement marqués par de l'activité présentent-ils une offre de restauration plutôt ouverte en semaine à midi, alors que d'autres plus mixtes présenteraient des offres ouvertes également les soirs de week-end dans la mesure où la part des logements y est plus importante ? Il

[15] Loi dite Macron du 6 Août 2015 dont l'article Art. L. 3132-24.-I. stipule que *Les établissements de vente au détail qui mettent à disposition des biens et des services et qui sont situés dans les zones touristiques internationales peuvent donner le repos hebdomadaire par roulement pour tout ou partie du personnel, dans les conditions prévues aux articles L. 3132-25-3 et L. 3132-25-4.*

est clair que pour relater cette réalité, il faut croiser différentes données et les données horaires et calendaires dont nous disposons sont particulièrement précieuses pour cela. La carte représentant les densités des activités de restauration ouvertes un mardi à 13h (boulangeries, sandwicheries, débits de boissons, restauration rapide et traditionnelle), réalisées sur le périmètre pour lequel nous disposons de données horaires, superposée à la carte des destinations de l'APUR, permet un premier aperçu des relations qui peuvent être étudiées entre différentes formes d'occupations temporaires de la ville et qui concourent à mieux comprendre les variations d'intensité de présences sur l'espace public (voir figure 3 ci-après).

FIG 3 – Carte des densités des activités de restauration ouvertes un mardi à 13h (boulangeries, sandwicheries, débits de boissons, restauration rapide et traditionnelle), superposées à la carte des destinations de l'APUR.

En ce qui concerne la nuit, nos interlocuteurs identifient clairement cette période et l'associent à des secteurs de la capitale, différenciés par leur activité nocturne. La chronotopie de la nuit semble un indicateur important dans la comparaison entre les singularités de différentes métropoles et le terme de chronotopie s'adapte particulièrement bien à cet espace-temps de la nuit qui relève à la fois d'une assise spatiale et d'une organisation temporelle, faisant de la nuit une sorte de hors temps social dont l'activité festive s'étend au-delà de la durée même de la nuit solaire, isolant les noctambules des rythmes circadiens dans des enveloppes étanchent à la lumière et dans lesquels les *beats* remplacent les repères de l'horloge ou du soleil.

Les ambiances peuvent s'inverser entre le jour et la nuit, comme par exemple rue Oberkampf, dans le 11ème arrondissement de Paris, caractérisée par une activité nocturne repérable à l'échelle de la ville entière et dont la renommée et l'attractivité dépasse largement les limites de Paris.

La densité urbaine et la diffusion du logement font que des conflits apparaissent dans les pratiques simultanées et décalées de la ville et ne se stabilisent que dans des secteurs spécialisés (par exemple rue Oberkampf) où les pratiques nocturnes ne peuvent être qu'acceptées par les riverains, devenant clairement dominantes.

Par ailleurs, des seuils temporels sont clairement repérés, notamment le soir, et qui sont en partie marqués par l'horaire de fin des activités commerciales et de services bordant l'espace public. Certaines activités de rez-de-chaussée faisant même office de balise, surtout lorsque les journées sont plus courtes, dans les périodes extrêmes du matin ou du soir.

Enfin, il sera intéressant de croiser les recherches menées par Claire Lagesse, dans le cadre de sa thèse[16] au sein du groupe de recherche Morphocity[17], sur la caractérisation du réseau des voies urbaines, afin de vérifier s'il y a des corrélations entre géométrie, histoire des voies et organisation des activités de rez-de-chaussées. Par exemple, en reprenant la typologie construite sur la base des classements élaborés par l'APUR et catégorisés par nos soins et faisant apparaître des distinctions entre activités quotidiennes, occasionnelles ou exceptionnelles. Ou encore en spatialisant les *clusters* horaires et calendaires des activités urbaines et en les croisant avec la caractérisation du réseau des voies du territoire étudié.

Une première description chronotopique du territoire étudié est désormais possible à travers l'ethnographie dont le terrain a pu être interprété. Cette description pourrait être la base d'une montée en généralité pour un modèle descriptif, mais celle-ci est certainement en partie spécifique au territoire étudié et il faudrait mener d'autres terrains pour en valider les aspects fondamentaux et repérer les dimensions variables selon le contexte. Sur le territoire étudié, il est important de bien avoir conscience de la partialité (en termes d'analyse chronotopique) des résultats que l'on peut obtenir à partir des données horaires et calendaires disponibles, mais il est toutefois intéressant d'essayer d'en épuiser le croisement possible avec les dimensions de l'expérience du temps apparues à travers l'ethnographie.

La recherche présentée ici, associant différentes disciplines, permet de tester les complémentarités entre ces dernières, tout en donnant une conscience aigue des limites de chacune d'entre-elles surtout lorsqu'on essaie d'identifier les données concourant à façonner une expérience humaine. Dans cette perspective, la collaboration et les itérations entre chercheurs entre disciplines, entre différents outils, sont une démarche nécessaire si l'on veut essayer de construire un modèle descriptif de la ville habitée tendue entre les faits observables et mesurables et les expériences vécues.

---

[16] Lagesse C., 2015, Lire les lignes de la ville. Méthodologie de caractérisation des graphes spatiaux, Thèse de doctorat, Université Paris-Diderot-Sorbonne-Paris Cité.

[17] Groupe animé conjointement par Philippe Bonnin (CNRS-AUS-LAVUE) et Stéphane Douady (CNRS-MSC). Recherche soutenue par le CNRS dans le cadre du PIRVE (Programme Interdisciplinaire de Recherches Ville-Environnement), puis par le programme ANR MoNuMoVi.

Références

Bakhtine, M. (1978), Esthétique et théorie du roman, Paris: Gallimard.

Benjamin, W. (1989), Paris, capitale du XIXe siècle – le livre des passages, Paris: Le Cerf.

Berque, A. (2000), Écoumène, introduction à l'étude des milieux humains, Paris: Belin.

Bonfiglioli, S., Mareggi, M. (dir.) (1997), Il tempo e la città fra natura e storia. Atlante di progetti sui tempi della città, Urbanistica Quaderni, Rome : Inu Edizioni, n. 12.

Guez, A., Rousseaux, F. (2016), « Approche pour l'élaboration d'un modèle chronotopique urbain », in Actes de l'atelier GAST – Gestion et Analyse de données Spatiales et Temporelles, Atelier inscrit dans la 16ème conférence internationale francophone sur l'extraction et la gestion des connaissances EGC 2016, sur coordonné par Éric Kergosien, Thomas Guyet, Christian Sallaberry, Reims, 19 janvier 2016, http://gt-gast.irisa.fr/gast-2016/

Guez, A., Biase de, A., Gatta, F., Zanini, P. (2016), Chronotopic exploration of a parisian landscape, in *Temporalités de la ville*, à paraître aux éditions Europia.

Guez, A., Mareggi, M. (1997), Représentations de temporalités urbaines, in S. Bonfiglioli, M. Mareggi (dir.), *Il tempo e la città fra natura e storia. Atlante di progetti sui tempi della città, Quaderni di Urbanistica*, Rome: Inu Edizioni.

Guez, A., Stabilini, S., Zedda, R. (2000), Les temps italiens se réorganisent, in *La Recherche*, N° 337, supplément ville.com, décembre 2000.

Hägerstrand, T. (1970), What about people in regional science ? : *Papers of the Regional Science Association*, 24 : 6– 21.

Lagesse C., 2015, Lire les lignes de la ville. Méthodologie de caractérisation des graphes spatiaux, Thèse de doctorat, Université Paris-Diderot-Sorbonne-Paris Cité.

Lefebvre, H. (1992), Éléments de rythmanalyse, Paris: Syllepse.

Lynch, K. (1972), What time is this place ? Cambridge : MIT Press.

Martinotti, G. (1993), Metropoli. La nuova morfologia sociale della città, Bologna : Il Mulino.

Rousseaux, F. (2006), La collection, un lieu privilégié pour penser ensemble singularité et synthèse, *revue EspacesTemps.net*: http://www.espacestemps.net/articles/la-collection-un-lieu-privilegie-pour-penser-ensemble-singularite-et-synthese/

Rousseaux, F., Saurel, P., Petit, J. (2014), Knowledge Engineering or Digital Humanities? Territorial Intelligence, a Case in Point. Innovation in Intelligent Machines-4, *Recent Advances in Knowledge Engineering: Paradigms and Applications*, Springer-Verlag, Vol. 514, pp. 129-187, C. Faucher and L-C. Jain eds.

Rousseaux, F., Legrand, J., Soulier, E., Bugeaud, F., Saurel, P., Neffati, H. (2012), A New Methodology for Collecting and Exploiting Vast Amounts of Dynamic Data, *3rd International Conference on Emerging Intelligent Data and Web Technologies*, 2012.

Rousseaux, F., Soulier, E., Saurel, P., Neffati, H. (2012), Agencement multi-échelle de territoires à valeur ajoutée numérique : des Hétérotopies foucaldiennes aux Complexes simpliciaux, Politiques publiques, Systèmes complexes, Edité par Danièle Bourcier, Romain Boulet et Pierre Mazzega, Editions Hermann, pp. 169-192, Paris.

Van Schaick, J. (2011), Timespace matters. Exploring the gap between knowing about activity patterns of people and knowing how to design and plan urban areas and regions, Delft: Eburon Academic Publishers. URL : http://www.eburon.nl/timespace_matters.

Wajcman, G. (1999), Collection, Paris: Nous.

Vers la représentation cartographique d'une extraction des motifs de flux mondiaux issue d'une régionalisation

Françoise Bahoken *

*Université Paris-Est, AME, SPLOTT,
IFSTTAR, F-77447, Marne-la-Vallée, France
UMR 8504 Géographie-Cités, Equipe PARIS.
francoise.bahoken@ifsttar.fr,

Résumé. L'une des principales familles de solutions apportées à l'objectif de réduction de la complexité de la carte de flux consiste à réduire la résolution des données. Les solutions liées à l'agrégation spatiale conduisent à un changement d'échelle géographique, à la production d'un maillage territorial de niveau régional issu d'une classification fonctionnelle car fondée sur les valeurs des flux. Si l'approche est efficace, on souhaite s'interroger sur les raisons de l'absence d'analyse factorielle réalisées en amont de ces partitions de l'espace par des flux.

1 Introduction

L'analyse (carto)graphique [1] des matrices de flux ou des mouvements, c'est-à-dire des quantités mesurables transférées entre des couples de lieux d'origines et de destination, dans un espace assorti d'une métrique, par exemple géographique, comporte de nombreux défis liés à la complexité intrinsèque de l'objet. En présence de données de flux massives, la difficulté consiste, en effet, à produire une carte descriptive ou explicative des motifs des flux qui soit lisible et intelligible, autrement dit dotée de sens dans un registre thématique donné. Pour y remédier, trois familles d'approches sont traditionnellement déployées. La première approche consiste à filtrer l'information de flux disponible de manière à n'en représenter qu'une sélection significative. La seconde approche consiste à fusionner visuellement les tracés symbolisant les flux de manière à réduire leur complexité graphique. La troisième approche s'inscrit dans un objectif de réduction de la résolution des données qui conduit à mobiliser, voire à générer, un nouveau maillage territorial associé aux flux, à changer d'échelle géographique. Si ces approches descriptives s'inscrivent toutes dans le registre des analyses statistiques uni et bi variées appliquées aux données de flux, la troisième d'entre elles ouvre néanmoins la voie vers les analyses multivariées de type classification ascendante. En inscrivant notre communication dans cette troisième approche, nous souhaitons présenter un exemple classique de régionalisation fonctionnelle par les flux, i.e. la production d'un maillage territorial issu d'une procédure de classification ascendante hiérarchique (CAH), qui entraîne un changement d'échelle géographique des flux représentés, par conséquent une réduction de la complexité graphique de la

1. Ce terme signale un problème graphique que l'on appréhende au niveau cartographique, i.e. en tenant compte de la composante spatiale de localisation des lieux (et non des sommets ou objets, termes non spécifiquement géographiques).

cartographie associée. L'objectif de notre communication est d'ouvrir une discussion sur les conditions de l'application de méthodes d'analyses factorielles, en amont ou en complément d'une CAH, devant permettre d'extraire les principaux motifs de flux et de mouvements à cartographier. Nous rappelons en premier lieu l'objectif général de la cartographie des matrices de flux denses. Nous présentons ensuite la famille de solutions liées aux procédures d'agrégation, que nous illustrons par une méthode de régionalisation fonctionnelle. Le cas d'application est une matrice de flux commerciaux internationaux mondiaux, dont le niveau élémentaire d'expression (pays*pays) va être réduit à un niveau interrégional.

2 L'objectif général de réduction de la complexité graphique

L'objet « matrice de flux », associé à un espace géographique, est formé d'objets vectoriels (le couple des (i, j) lieux d'origine et de destination) en sus des scalaires (la distribution des (i) lieux de la zone d'étude) à représenter. Cette particularité en fait un objet complexe à cartographier dans la mesure ou la structure multidimensionnelle de la matrice nécessite l'articulation d'au moins quatre dimensions à représenter : les positions des lieux $(i), (j)$ définies par un couple de coordonnées géographiques (x, y) dans un système de projection donné ; le tracé du lien caractéristique du couple $L(i, j)$; la symbolisation de la valeur du flux $F(i, j)$ comme composante attributaire ou thématique caractérisant le type de flux dont il est question (par exemple, migratoire ou commercial) ; ainsi que la temporalité (t), soit : (x_i, y_i) ; (x_j, y_j) ; F_{ij} ; t.
Cette complexité intrinsèque de l'objet pose de nombreuses difficultés d'exploitation, d'une part, sur les aspects graphiques et sémiologiques de la représentation (la réalisation de tracés pertinents et signifiants au niveau thématique) et, d'autre part, sur les impératifs du traitement des données numériques : de l'information de flux (disponible dans la matrice de départ) et de la donnée cartographique (le maillage territorial). Toutes constituent autant de défis, souvent multiples, à résoudre qui sont renforcés dans le cas de données spatiales massives et/ou complexes ; la complexité étant entendue comme l'existence de matrices « empilées » sur un même espace géographique, autrement dit de sous-matrices qui correspondent à un découpage de la quantité de flux (Z) en catégories thématiques (par exemple : en types de produits transportés) et/ou en (T) périodes (par exemple, lors ces données sont disponibles à différentes dates).
La difficulté de cartographier ces matrices denses tient, en effet, dans la conjonction de plusieurs effets bien connus, parmi lesquels les effets graphiques suivants (Bahoken, 2015) :

- un « effet de surcharge graphique », caractéristique d'une densité de figurés par ailleurs hétérogènes trop importante ;
- un « effet de couverture spatiale », caractéristique d'une densité de teinte (surface visuelle) par rapport au fond trop élevée ;
- un « effet d'enchevêtrement » lié à la présence de figurés linéaires fortement enchevêtrés en raison d'orientations angulaires variées .

La combinaison de ces effets se répercute ainsi au niveau graphique par une complexité qui empêche, d'après D. Fairbairn (2006), la compréhension aisée de l'information. Cette situation explique que la clarté de ces représentations spatialisées de flux soit d'ailleurs considérée par (Tobler, n.-d.) comme une « propriété générale » de la carte de flux, comme l'objectif de la figure. La cartographie de flux est par conséquent fortement sensible aux choix réalisés en amont de la représentation proprement dite : lors des traitements numériques et cartographiques réa-

lisés pour produire une figure lisible et pertinente sur le plan thématique.
En matière de flux, indépendamment du caractère descriptif ou explicatif de l'approche mobilisée, plusieurs directions sont traditionnellement privilégiées pour réduire sa complexité analytique (voir Bahoken, 2016 : 205-206) : (i) agir sur les aspects graphiques ; (ii) filtrer les données affichées ; (iii) agréger les données, en réduisant leur résolution. Les procédures de type fusion graphique intervenant sur les liens (edge bundling : Holten, 2006, 2009 ; Lambert, Bourqui, Auber, 2010) et/ou sur les noeuds s'inscrivent dans la catégorie (i) et permettent de donner l'illusion d'un mouvement dans l'espace géographique. Suivant la catégorie (ii), la sélection des valeurs jugées significatives pour le propos et/ou par la sélection des objets spatiaux réduisent naturellement la quantité d'objets à représenter. Les procédures d'agrégation spatiale qui nous intéressent appartiennent à la (iii) catégorie, car elles regroupent les objets en communautés pertinentes suivant des critères qui portent sur les lieux géographiques et/ou sur les liaisons entre les lieux.

3 Les solutions liées aux procédures d'agrégation des flux

L'un des éléments d'explication de la complexité graphique de la carte de flux étant d'ordre géographique (l'augmentation du nombre de liens à représenter, $N(N-1)$, est une fonction quadratique du nombre N de lieux de la zone d'étude), l'une des possibilités consiste à réduire le nombre théorique de liens, en regroupant les lieux.
Cette éventualité intéresse les questions liées à la partition des matrices de flux : par exemple, la création de sous-tableaux (des sommes marginales) ou l'identification de blocs internes (formé d'un ensemble de lieux), considérant une matrice externe des positions (x, y) de ces lieux. Ces blocs peuvent être appréhendés comme des classes, i.e. comme des régions, regroupant un certain nombre de couples de lieux qui partagent une caractéristique d'appartenance commune. La production de ces blocs régionaux intéresse la problématique du changement d'échelle géographique, car la production d'un maillage territorial peut être issue de données de flux. D'après C. Grasland (1997), un maillage territorial est une partition qui discrétise l'espace en deux catégories binaires : un couple de lieux appartient ainsi ou non à un ensemble régional, ce qui « [...] *implique que l'on peut définir des maillages soit à travers l'étude des attributs d'un ensemble d'éléments [par exemple, ceux des lieux, d'origine ou de destination, des flux], soit à travers l'étude des relations entre les éléments d'un ensemble [...] le maillage est à la fois un réseau de relation et une partition* ».
Cette solution, liée aux procédures d'agrégation, lorsqu'elle est apportée au problème de la complexité graphique des cartes de flux est appréhendée soit du point de vue des lieux, soit du point de vue des flux. Dans le premier cas, la partition est préexistante aux flux : la carte décrit alors les flux observés à l'échelle de pays voire de régions, échelons formés d'unités spatiales définies [2] indépendamment des flux. Dans le second cas, le niveau régional, qui sert de fond de carte, est issu d'une classification préalable des valeurs de flux observées à l'échelle des pays. C'est cette seconde possibilité qui nous intéresse ici.

2. Un exemple de cartes de flux mondiaux qui en résulte est proposé dans (Bahoken 2016 :317-322).

3.1 La génération de partitions territoriales issues des flux

Aborder la production de maillages du point de vue des flux est devenu un impératif, depuis que S. Openshaw (1977) a mis en évidence la manifestation d'un effet du problème des unités spatiales modifiables (Modifiable area unit problem MAUP) à partir de données de flux modélisées sur un maillage territorial inadéquat. Si la production de partitions territoriales par les flux permet de répondre aux prérequis des modèles (Combes et al. 1996), elle répond aussi à l'objectif de réduction de la complexité de la carte, par la réduction du nombre de lieux concernés par les flux.
Sur le plan méthodologique, la génération de maillages à partir d'unités spatiales élémentaires porte soit sur leur redécoupage, par modification de leur morphologie (forme, taille…), soit sur leur regroupement en régions fondé sur un ou plusieurs critères d'agrégation (contiguïté des zones, taille minimale ou maximale de la région, similitude ou non des zones intégrées, nombre de régions...). Dans le contexte de l'analyse géographique, ces procédures d'agrégation, qu'elles soient structurelles [3] ou fonctionnelles [4] , conduisent toutes à un changement d'échelle géographique, c'est-à-dire à la production d'un nouveau zonage de résolution spatiale supérieure.
L'agrégation entraîne donc un changement conceptuel fort, car elle modifie la résolution des données et leur thématique : l'analyse de flux intercommunaux différant de celle de flux inter régionaux. La procédure qualifiée de « régionalisation fonctionnelle » consiste donc à regrouper des unités spatiales élémentaires en régions qui sont fonction du type et de l'intensité des échanges entre les couples de lieux qui les composent ; ces régions correspondant à des classes définies selon différents critères d'homogénéité. L'un des intérêts de ces méthodes de régionalisation fonctionnelle réside donc dans la construction d'un espace multi scalaire fondé sur l'analyse des flux à différentes échelles géographiques, en incluant des critères spatiaux : de distance, de voisinage ou de contiguïté. La génération du maillage proprement dit intervenant après une opération de classification des flux qui conduit au regroupement des unités spatiales élémentaires en régions.

3.2 Une typologie des méthodes de partition territoriale par les flux

Plusieurs méthodes de partition territoriale par les flux sont envisageables, en fonction du modèle théorique sous-jacent, un présentation est disponible dans Grasland et al. (2009). Aussi pouvons-nous évoquer a minima trois grandes approches, sans compter les approches mixtes développées par Coombes et al. (1986). Il s'agit de l'approche nodale, de l'approche centre-périphérie et de la régionalisation fonctionnelle. L'approche nodale est inscrite dans le cadre d'analyse des flux dits majeurs et dominants. Elle est fondée sur la méthode initialement proposée par J. D. Nyusten et M. F. Dacey (1961), de définition de régions nodales autour des lieux les plus attractifs d'un système d'échanges. En révélant une organisation spatiale multipolaire, la méthode décrit l'arborescence des relations (et non des flux), de même que l'aire d'influence des pôles dominants. L'approche étant populaire, mais jugée restrictive, plusieurs variantes en

3. Les procédures structurelles sont fondées sur des caractéristiques morphologiques (géométrie, taille…) ou attributaires des lieux (appartenance à une même région administrative de niveau supérieur).

4. Les procédures fonctionnelles sont fondées sur les couples de lieux, dans notre cas sur l'analyse de l'intensité de leur relation (la valeur de flux), de la distance qui les sépare, etc. Elles pourraient également être fondées sur l'existence de relations, comme par exemple dans les approches de clustering de type Théorie des graphes.

ont été proposées (voir par exemple Cattan, 2004). L'approche centre-périphérie diffère de la précédente en ce sens qu'elle est fondée sur les valeurs de flux et mobilise en outre des données structurelles pour définir des zonages totalement indépendants des découpages administratifs. En France, le découpage en Aires urbaines, par exemple, est élaboré par l'INSEE à partir de tableaux de navettes domicile-travail. Sa cartographie décrit l'aire d'influence des territoires urbains, en les distinguant des territoires ruraux.
L'approche fonctionnelle est fondée sur le principe général de la CAH appliqué aux données de flux. Plusieurs versions en ont été proposées par l'INSEE, à la suite de la méthode Zonage élaborée dès 1965 pour générer un maillage fondé sur la hiérarchie des échanges. On peut citer les méthodes de Denimal et de C. Terrier. La première mobilise une matrice d'adjacence entre les lieux et non de flux, comme dans l'approche de J. D. Nyusten et M. F. Dacey. La méthode Mirabelle de C. Terrier (INSEE 1998) est, quant à elle, plus étendue que la méthode Zonage d'origine, car elle vise la production d'un « *zonage adéquat* » permettant l'analyse des configurations spatiales issues des volumes de flux (avec différentes variantes de pondération) ainsi que des liens hiérarchiques. Mirabelle propose ainsi « *des zonages de finesse différente plus ou moins emboités* » (Terrier, 1998).
L'une des méthodes anglo-saxonne très populaire [5] de régionalisation par les flux est la méthode Intramax (« *Intrazonal interaction maximisation* ») et sa variante Intramin développée par Masser et Brown (1975).

4 La méthode Intramax/min, principe et application

La méthode *Intramax/min* de régionalisation par les flux a ceci de particulier qu'elle n'est pas fondée sur un critère de contiguïté, elle ne mobilise que l'intensité des flux et une acception de la distance fonctionnelle.

4.1 Principe de l'Intramax/min

Le fondement de la méthode Intramax/min réside dans l'analogie du concept de « région fonctionnelle » avec l'hypothèse gravitaire de décroissance de l'intensité du flux avec l'augmentation de la distance parcourue. De ce fait, l'intensité des interactions entre deux zones sera d'autant plus forte que la « *distance fonctionnelle* » qui les sépare est faible. Ainsi, une zone quelconque sera agrégée avec une autre zone au sein d'une région si leur « *distance fonctionnelle* » est faible et leurs échanges intenses.
Appliquée pas à pas, la procédure permet de générer $(n-1)$ partitions territoriales par les flux dont le critère d'homogénéité maximise les flux intra régionaux cumulés à l'origine de la formation des régions, autrement dit le volume total des interactions entre les lieux regroupés dans les régions (la version Intramin étant une minimisation de ces flux intra régionaux). A la fin de la procédure, comme dans une CAH classique, l'ensemble des unités spatiales se retrouve dans la même zone.
Les résultats produits par la méthode sont : un dendrogramme, un rapport d'agrégation, sur les $(n-1)$ étapes précisant le pourcentage de flux intra régionaux cumulé à chacune des étapes ;

5. La popularité de cette méthode est probablement en partie liée à son implémentation dans le logiciel Flowmap© (Van der Zwan et al., 2003).

éventuellement des matrices de flux inter régionaux issues de l'état de l'agrégation par la fonction somme des valeurs de flux aux différentes étapes.

4.2 Application

Pour illustrer le fonctionnement de la méthode Intramax, nous l'appliquons à une matrice de flux internationaux mondiaux décrivant l'ensemble des échanges commerciaux, en valeur (US dollars), réalisés entre des pays ou des groupes de pays. Les données sont issues de la base CIN de CHELEM du CEPII (2004, 2005, 2006). Préparées dans le cadre du projet Eurobroadmap (2009-2011), elles ont été tri-annualisées entre 2004 et 2006 et décrivent ainsi la moyenne mobile de l'ensemble des flux commerciaux internationaux sur les trois années. La Fig. 1 présente le niveau géographique élémentaire de description de ces flux.

Au début de la procédure, les 220 unités spatiales (pays ou groupes de pays) sont indépendantes et l'interaction intra régionale est de 0%. La densité de la matrice est de 36% dont 32% de cellules différentes de zéro (20% de l'ensemble total). Le nombre maximal d'étapes est de 172.

Les Fig. 2 et 3 présentent deux des niveaux de partition de l'espace mondial formés par la classification de ces flux commerciaux. On observe que à 51,5% d'interactions intra régionale (Fig.2), l'espace commercial mondial est formé de sept grandes régions homogènes que sont une région Amérique, une région Europe du nord, une région formée de l'Europe occidentale, septentrionale incluant les pays du pourtour méditerranéen et une grande région englobant le proche orient, l'Asie et Australie ; une région fédération de Russie et une région Afrique sub-saharienne. A 77% d'interactions intra régionale (Fig.3), seules trois grandes régions subsistent à savoir : les amériques-Asie du sud et Australie ; l'Europe ? Afrique du nord et Russie ; l'Afrique sub-saharienne. Ces régions peuvent ainsi être interprétées comme autant d'aires d'influence commerciales.
Une des questions principales consiste donc à choisir la fonction de régionalisation qui conviendra le mieux aux données de flux disponibles, en particulier les éventuelles contraintes spatiales liées à l'échelle d'expression des flux (locale, régionale, mondiale) et la disponibilité des données. Plusieurs fonctions ont ainsi été définies dans les années 1970 de manière à maximiser ou minimiser le pourcentage cumulé de ces flux intra-zones (ou interzones), une revue est proposée par S. Alvanidès (2000).
La méthode Intramax, développé initialement à l'échelle locale, est également efficace pour générer des partitions territoriales à l'échelle mondiale (voir Fig. 2 et 3.). Elle conduit à une forme de segmentation des valeurs de flux qu'il est possible de cartographier de manière classique : en symbolisant les relations qui se produisent entre des lieux appartenant à une même région et celles qui se produisent entre des régions, jouant alors le rôle de *hub* permettant de passer d'une région à une autre.

5 Conclusion

Le fait est que cette approche descriptive soit directement mise en oeuvre sur les valeurs de flux et qu'elle conduise in fine à un filtrage des valeurs par l'espace (il est possible de ne représenter que les valeurs qui s'expriment entre ou au sein de certaines régions) peut être

préjudiciable pour l'analyse. De même que l'on peut s'interroger sur les raisons de l'absence d'analyse factorielle réalisée au préalable de cette partition de l'espace par des flux.
En effet, la régionalisation fonctionnelle n'intègre pas de module de traitement multivarié qui consisterait à extraire préalablement les axes factoriels, en l'occurrence les composantes principales de ces flux que l'on pourrait interpréter comme un « résumé » des motifs principaux des flux. Quelles en sont les raisons ? Est-il possible de trouver une (ou plusieurs) méthodes complémentaires à ces classifications ?
Pour la poursuite de ces travaux exploratoires sur les méthodes multivariées liées à l'agrégation des données de flux, nous faisons l'hypothèse que l'extraction des composantes principales d'une matrice de flux bilatérale, éventuellement catégorielle ou temporelle, permettrait de résumer efficacement l'information qu'elle contient et d'en proposer une représentation cartographique simplifiée, tout en mobilisant l'ensemble des valeurs disponibles. Cette analyse serait d'ailleurs plus robuste, car fondée sur l'analyse de la variance de données dont on sait qu'elles sont fortement dissymétriques.

6 Références mentionnées

Alvanides S., Openshaw S., Duke-Williams O., 2000, Designing zoning systems for flows datas, In : Atkinson P., Martin D. (eds.), 2000, Geocomputation, Part II : Zonation and Generalization, Taylor and Francis Group, Ed. CRC Press, pp. 115-134.

Bahoken, F., (2011), Evolution des limites de régions fonctionnelles Européennes liées aux flux internationaux, Proceedings of the EuroBroadMap International Conference, Session C : « Frontières, réseaux et régionalisations floues du monde », Rouen, France, 1-3 Décembre 2011, pp. 14-20.

Bahoken, F., (2015), Eléments pour une représentation (carto)graphique des matrices de flux, M@ppemonde, n°115.

Bahoken, F., (2016), Contribution à la cartographie d'une matrice de flux, Thèse de doctorat en géographie, Sciences des territoires, Université Paris-Sorbonne (Paris 7), 520 p.

Cattan, N. (2004), Le monde au prisme des réseaux aériens, Flux, vol. 4, n°58, pp. 32-43.

Coombes M. G., Green, A. E., Openshaw, S. (1986) : An efficient algorithm to generate official statistical reporting areas : The case of the 1984 travel-to-work-areas revision in Britain. Journal of the Operational Research Society, 37(10), pp. 943-953.

Fairbairn, D. (2006), Measuring Map Complexity, Cartographic Journal, vol. 43, n°3, pp. 224 - 238.

Grasland, C. (1997), Contribution à l'analyse géographique des maillages territoriaux, Habilitation à Diriger des Recherches en géographie, Volume inédit, Université Paris I, 373 p.

Grasland, C., Bahoken, F, Beauguitte, L., Pion, G., Van Hamme, G. (2009), Toolbox for flows and network analysis (Methodological Paper). Deliverable D.5.1. EuroBroadMap.Vision of Europe in the World. Small or medium scale focused project FP7-SSH-2007-1, 100 p.

Holten, D. (2006), Hierarchical edge bundles : Visualization of adjacency relations in hierarchical data, Proceedings of the IEEE Transactions on Visualization and Computer Graphics, vol. 12, n°5, pp. 741-748.

Holten, D., Van Wijks, J. (2009), Force-Directed Edge Bundling for Graph Visualization, in : Hege, H.-C., Hotz, I., Munzner, T., (Guest editor), Journal compilation of The Eurographics Association /IEEE-VGTC Symposium on Visualization, Vol.28, n°3, pp. 983-990.

INSEE (1998) : « Les zonages : enjeux et méthodes » - INSEE Méthodes n°83 - décembre 1998

Lambert, A., Aubert, D., Mélançon, G. (2010), Living flows : enhanced exploration of edge-bundled graphs based on GPU-intensive edge rendering, Proceeding of the 14th Interna-

tional Conference on Information Visualization, pp. 523-530.

Masser, I., Brown, P.J.B. (1975), Hierarchical aggregation procedures for interaction data, Environment and Planning, A7, pp. 509-523.

Openshaw, S., (1977) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling, Revue Transactions of the institute of british geographers, Institute of British Geographers, pp. 459-472.

Tobler, W. (1969), Geographical Filters and their Inverses, In : Geographical Analysis, Vol. 1, n° 3, pp. 234-253.

Tobler, W. R., (n.-d.) On viewing flow map, Note de recherche non datée, 6 p.

Summary

One of the main families of solutions for reducing the complexity of flow mapping (the cluttering of a flowmap) is to reduce the spatial resolution of flow data. Solutions linked to spatial aggregation procedures lead to a change of geographical scale, they produce a territorial framework at a regional level, a functional regionalization. If the approach is efficient, we want to question the reasons for the absence of factorial analysis carried out before these partitions of space by flows.

FIG. 1 – *Le découpage du monde selon CHELEM.*

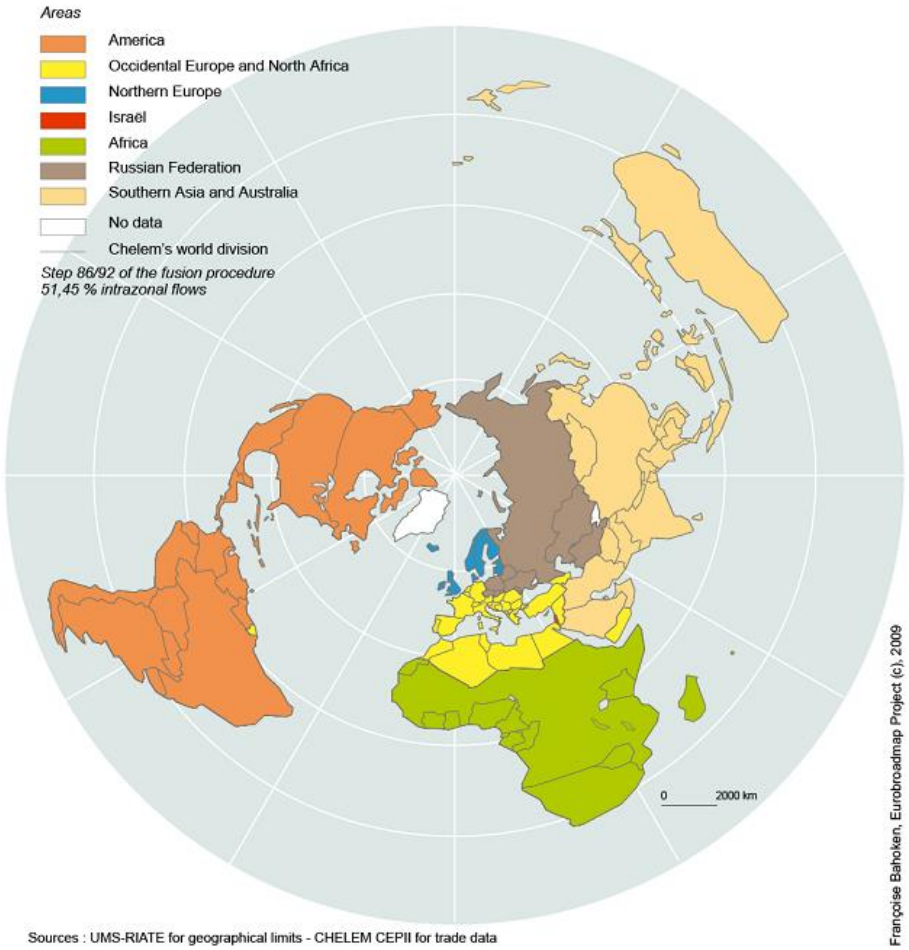


FIG. 2 – *Exemples de partition de l'espace mondial issue d'une régionalisation fonctionnelle par des flux commerciaux mondiaux : 51,45% de flux intra zonaux cumulés*

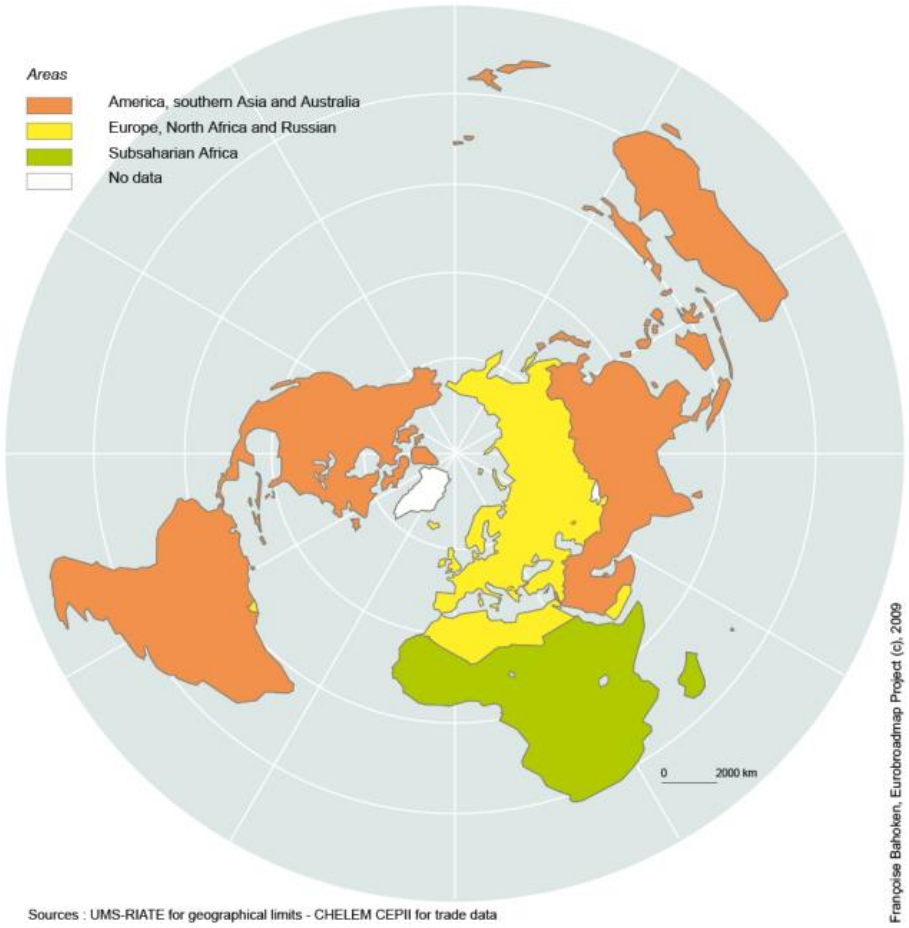


FIG. 3 – *Exemples de partition de l'espace mondial issue d'une régionalisation fonctionnelle par des flux commerciaux mondiaux : 77,04% de flux intra zonaux cumulés*

The Semantic Evolution of Observation Data: Visualizing spatiotemporal water data

Sylvain Bouveret*, Philippe Genoud*
Kim Hobus**, Danielle Ziebelin*

*Université Grenoble-Alpes CNRS LIG
<prenom.nom>@imag.fr
http://www.liglab.fr/fr/presentation/equipes/steamer
**Yukon Water Board
Kim.Hobus@gov.yk.ca
http://www.yukonwaterboard.ca

Résumé. The Coordinated Online Information Network (COIN) incorporates semantic web technology that integrates, publishes and visualizes time series water data allowing users to access a multitude of datasets in order to compare the data and draw conclusions. It employs an efficient and systematic data storage and retrieval system and can convert the data to a standard format for visualization. COIN utilizes a number of standards from OGC (Open Geospatial Consortium) and W3C (RDF, OWL, SPARQL and GeoSPARL) and benefits from generic ontologies transforming semantics to data, enriching data sets and making it available and interoperable via WFS and WMS standards. These principles facilitate publication and exchange of data across the web, increasing transparency and interpretability. Through modernized data submission and retrieval we hope to break down the silos of data, allowing users to visualize time series water quality and hydrometric data from multiple sources to increase knowledge in relationship to impacts on Yukon water.

1. Introduction

The study of the evolution of observation data is a central task in environmental management; it is essential for researchers to review and compare other data. Environmental management has become a multidisciplinary field in which stakeholders have to consult a multitude of different sources (water/soil chemical analyses, agricultural elements, land use etc). However, the heterogeneity of models, data, metadata and formats and their change over time, remains a major difficulty in integrating different sources. Currently, after the tremendous growth of Web 2.0, we are witnessing an evolution of the World Wide Web to what the W3C refers to as web data: a model for simple, flexible and powerful data. The Resource Description Framework (RDF) (Cyganiak et al, 2014), is based on web infrastructure and facilitates publication and exchange of data across the web. The representation models and the ontologies, expressed in RDFS (Brickley and Guha 2004) and OWL (Hitzler et al., 2012), give a semantics to data. In this paper, we present how data sources could be integrated in an RDF graph, visually presented in a time series format, and processed by associating a semantic model from ontologies, enriched by linking to other data sets.

COIN was developed in order to help Yukon Water Board officials in their evaluation of water license applications. Officials need to synthesize multiple information concerning a particular area and its immediate surrounding area. COIN allows us to understand the various steps necessary for such an approach and utilizes different technologies necessary for its implementation and according to the OGC and W3C web services description and orchestration recommendations, COIN use: XML, HTTP, WFS, WCS, WMS, and CSW. For observations and measurements: SensorML, O&M ISO 19156 were studied. For the exchange of hydrological time-series, as we did in COIN, WaterML 2.0 is the standard exchange format that we used. Our case study focused on the Yukon ecosystems: biological, chemical and physical data come either from water monitoring stations or sampling results. The purpose of these data sets is to help researchers to answer questions about water regulation, water quality, water management, and identify changes in the Yukon environment. Over the years a number of datasets have been collected, sometimes different types and forms, stored in various file formats. For easy operation, we propose a semantic web architecture that follows the Semantic Web stack [W3C 2005], implements processing functionalities in our application with queries, ontologies, inference rules and analysis services based on an RDF format and then links them to other external data sources. See Fig.1 for a visualization of this process.

2. Architecture of COIN

The basic idea of open and shared data is to establish a way to create an open and extendable infrastructure, which gives free access to information, its use and re-use. That means nonproprietary formats with low barriers and an open data license to ensure the possible re-use by anyone. It is crucial to link information and data with their context as it creates new knowledge. Tim Berners-Lee in 2010 presented his 5* model to explain the cost and the benefit of such architecture:

‘*’ Information is available on the Web (any format) under an open license

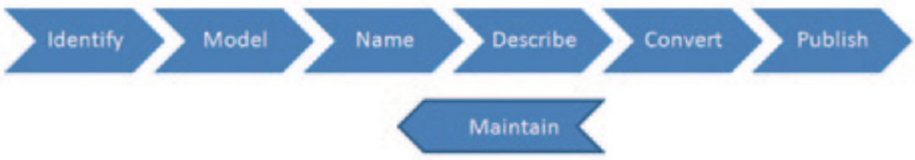
‘**’ Information is available as structured data (e.g. Excel instead of an image scan of a table)

‘***’ Non-proprietary formats are used (e.g. CSV instead of Excel)

****’ URI identification is used so that people can point at individual data

‘*****’ Data is linked to other data to provide context

W3C recommend by the W3C linked Data Cookbook the following steps to publish data:



To meet these objectives, we chose to build a triple store web application, as a web of data approach. According to the W3C, web services are self-contained and self-described application components using XML and HTTP for communication with other applications by

using open protocols. To be OGC compliant the web services need to be developed in the following way:

- for spatial data WFS for vector and discrete data,
- WMS for image and maps,
- CSW for catalogue service and metadata
- WCS for continuous and raster data access

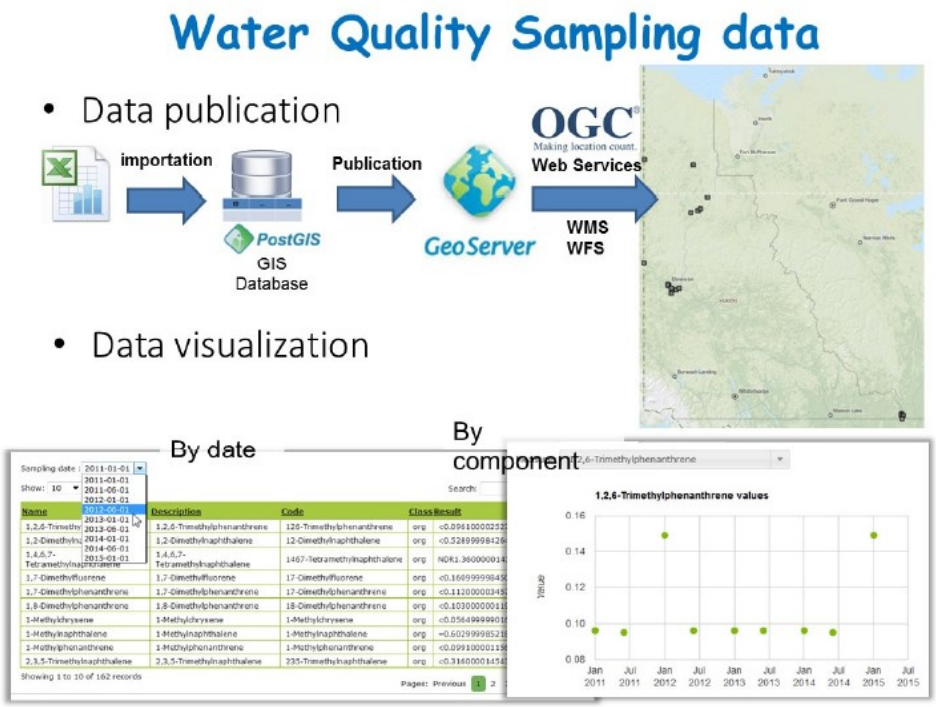


FIG. 1 - *The objective of COIN is to help the public, government, First Nation governments, and licensing officers to answer questions about water quality, water management and water allocations. To reach this objective we created an online information network with an ontology-based interface to visualize data and information source content.*

Web services are essential in the orchestration of internet-based workflows. Currently, two main architectural styles are most commonly used: SOA and REST. SOA services use remote procedure calls to invoke functions on remote systems. REST, or Representational State Transfer, is an alternative architectural model where each resource has a URI.

We have developed the following features:

- Interactive map of Yukon
- Display hydrometric data in a graph as well as allow for downloading data
- Display water quality data in a table or graph
- Search water quality data by chemical as well as by geographic area
- Set data submission standards for water quality data
- Filter map layers to anything within a geographic area

- Display and search Yukon Water Board (YWB) water license information
- Display water flow direction
- Link to Wikipedia for information regarding chemistry
- Tool to determine water allocated within a watershed.

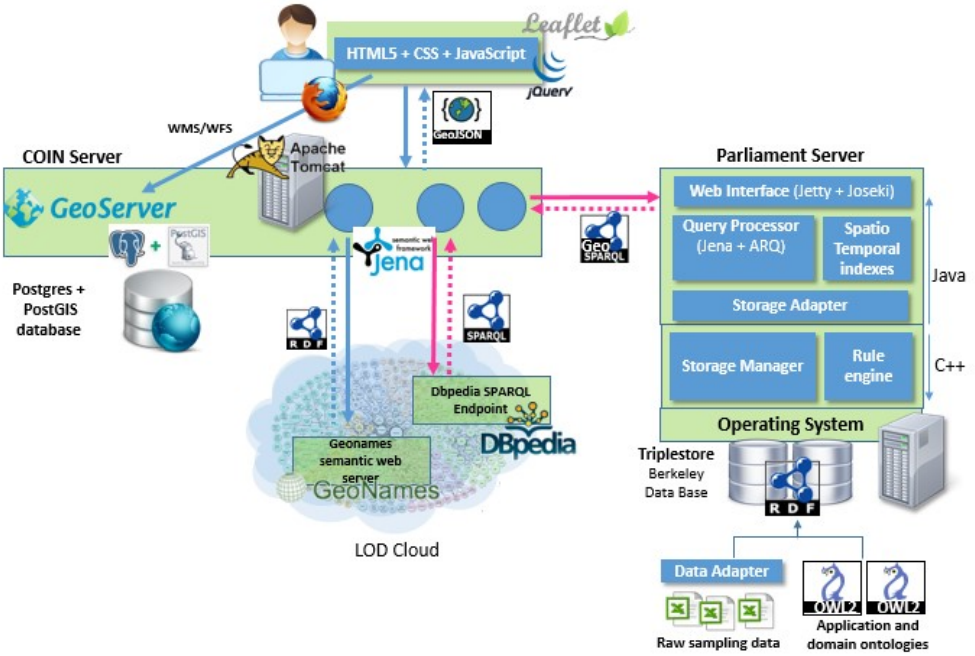


FIG. 2 – *The architecture shows how the different elements are organized. The RDF triple store for storing and integrating spatiotemporal datasets and the use of the web services provide access to the map and visualization interfaces though GIS layers. The RDF representation with the ontology is connected though data on the web.*

3. COIN application data

Numerous datasets are available from different sources in different formats ie. database, excel, and pdf. Figure 1 presents an example of sampling stations: a set of measures at different dates, in different locations and with different technology. The COIN application can then make this data accessible to users in an easy way with a graphical web user interface using maps. According to the OGC and W3C web services description and orchestration recommendations, COIN uses: XML, HTTP, WFS, WCS, WMS, and CSW as mentioned above. For observations and measurements: SWE and SOS, SensorML, O&M ISO 19156 and for exchange hydrological time-series, WaterML 2.0 is the standard exchange format. GeoSciML is a standard of data format. These are the principles that we applied to environmental data on water quality in northern Canada and which were used to study the impact on water in the Yukon Territory (Lim al., 2008).

The objective is to store different and heterogeneous types of observations (hydrometric stations, water licenses, water quality analysis etc). These data structures which store selected fields, values and tables are the key elements of the data model.

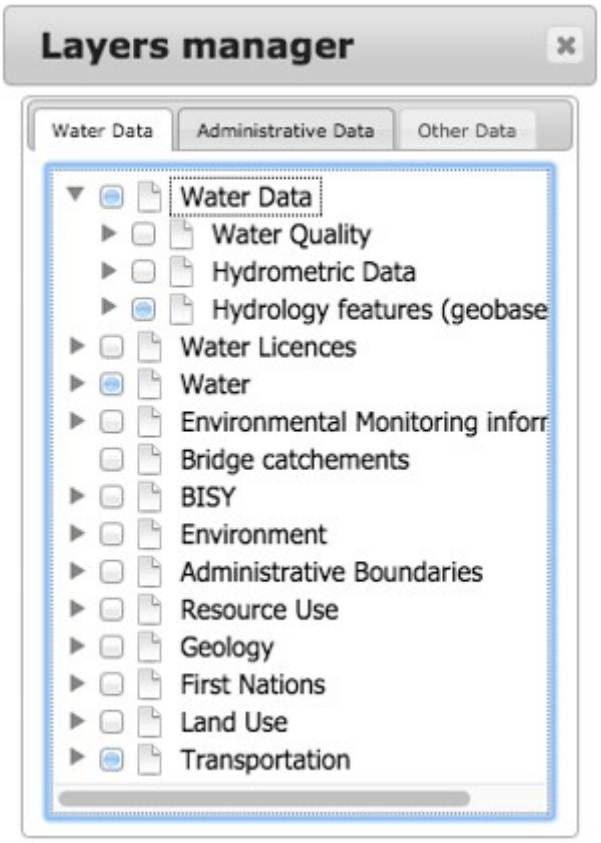


FIG. 3- *The data collected at sampling and collecting sites include: geographical description (administrative and geographic description), of station, of monitoring sites, type and size of the sampling area (ponds, lakes, springs, rivers, wetlands, watersheds), date of sampling.*

To build our model we propose a data dictionary that describes all tables and fields. The data series identify which variables have been measured at which locations and for what time stamps or periods. The data series are sets of observation values with specific properties (continuous or discontinuous measurements, units). These data series illustrate the dataflow and derivations and estimations are inferred associated to these data values. An appropriate method allows to obtain computed data values, provide logical grouping of data and make explicit the relationship between different monitoring and observation sites. It is critical that the different data sources are carefully documented and annotated with an adequate metadata structure. This structure will be used by different actors to collect data sets, and will enable easy access to the information

The aim of open and linked data is to make accessible the datatsets on a website but also to connect them to other scientific, economic, social and possibly political data sources in relationships with these elements. COIN's datasets provide different representation of sites : raw data, water quality, water microbiological information, environmental status of the site, information about agricultural and mining waste etc from different land and territory status. In order to ensure that these data sets are available to be reused and correlated to other data sets, they must be available as open and linked data. The initiative of open and linked data (Linked Open Data) (Bizer et al., 2009) follows this line, whose principles were set out by Tim Berners-Lee (Berners-Lee 2006):

(1) use of URIs (Uniform Resource Identifiers) to name (identify) things,
(2) use of HTTP URIs to consult these addresses,
(3) when a URI is accessed, provision of useful information using open standards (RDF, SPARQL, ...), and
(4) inclusion of links to other URIs in order to discover more linked data

4. Ontologies and models

To model data in the form of an RDF graph, an OWL ontology has been specifically defined for this application. This ontology includes a number of classes and properties for representing observational data. To represent spatial information (coordinates of sampling sites, geometry regions) we relied on the GeoSPARQL standard proposed by OGC (OGC, 2012) (Battle Kolas, 2012). Figure 5 shows in the form of a UML diagram the various classes and relationships (owl: ObjectProperties) defined for our application.

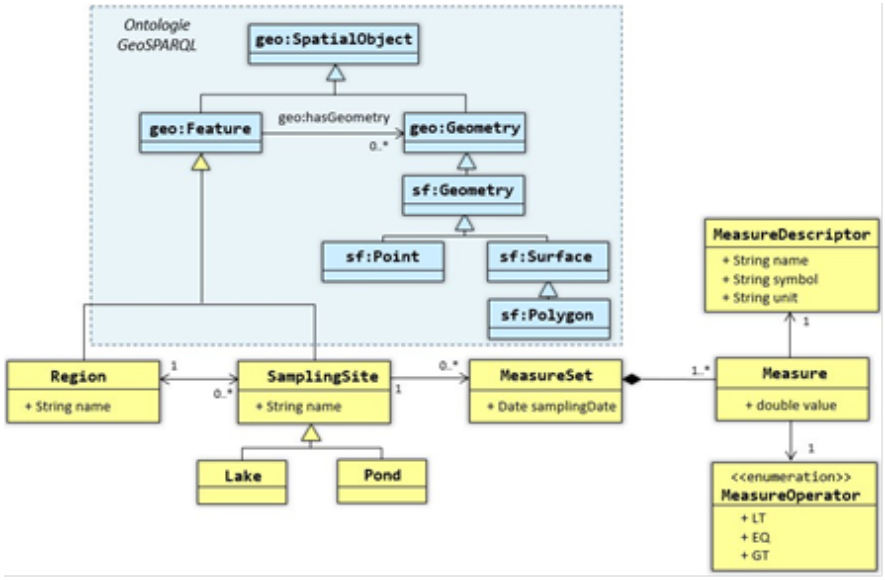


FIG. 4: *The ontology of the application and its links with GeoSPARQL ontology*

We have extended our application ontology with a general ontology from hydrology used by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) (Couch et al., 2014).

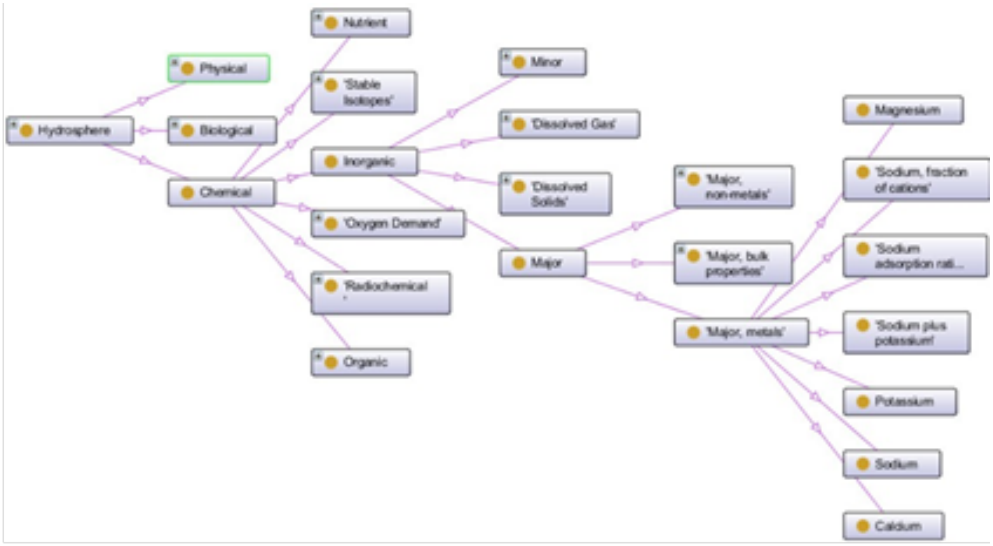


FIG. 5: *An extract of the ontology of the CUAHSI hydrosphere. Only the classes corresponding to the major metal concepts are developed in this hierarchy.*

This ontology defines a taxonomy that can hierarchically structure more than 4,000 words describing physical, chemical and biological measures related to water. It is used by the System Information CUAHSI (CUAHSI-HIS Hydrologic Information System) and consists of a set of servers and databases connected to client applications such as web services to facilitate the discovery of time series data collected at a given point. We have taken this ontology, defined in tabular form, and translated it into an OWL class hierarchy (Figure 5). The use of this ontology in our model is made by combining corresponding measures in the terminology of CUAHSI with the descriptor measures identified during sampling.

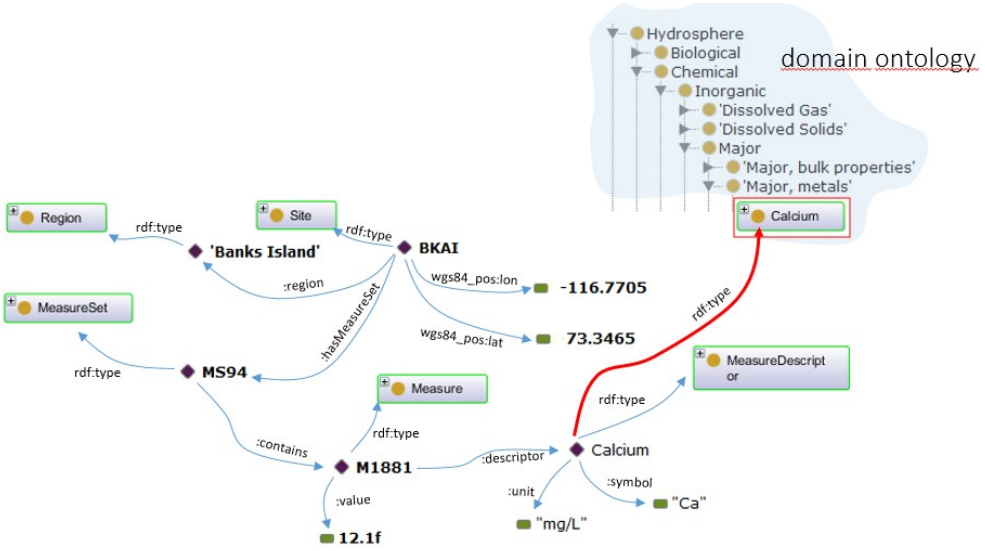


FIG. 6: *On the BKAI site at Banks Island a set of MS94 measures was taken. The M1881 12.10 measurement value is associated with a descriptor indicating that this is a measure of Calcium and its unit is mg / l. This descriptor is connected to the domain ontology. CUAHSI by a relationship rdf: type.*

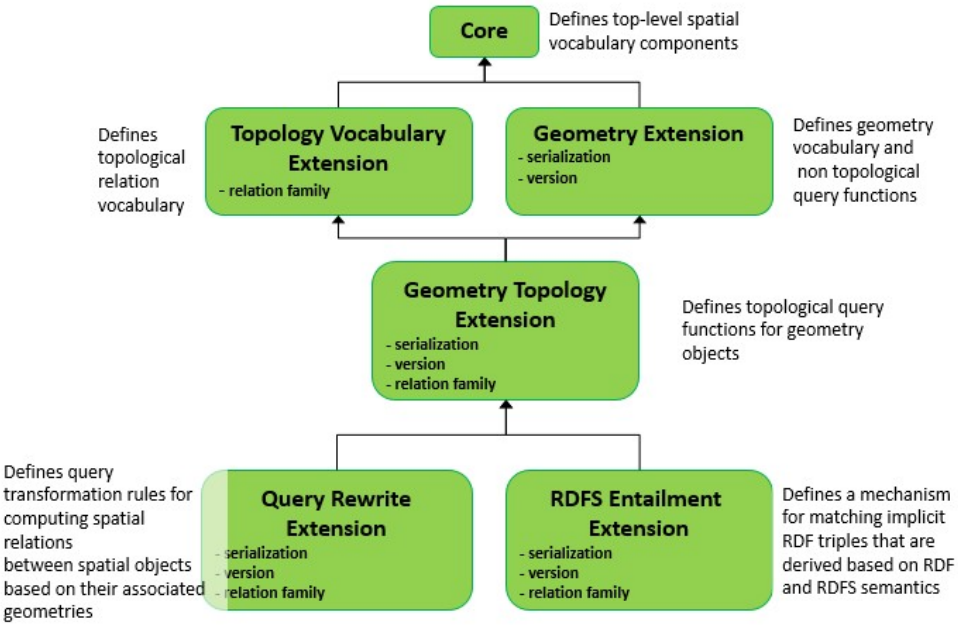
5. Exploring and analysing data

To explore and analyze the observational data thus represented, we used a web mapping interface for data queries through SPARQL and its spatial extension GeoSPARQL.

Our sampling results, stored in excel, were converted to RDF and stored in a specially designed database for storage and data recovery, called "triplestore". Although published in 2012, few triplestores currently support the standards GeoSPARQL (Athanasiou et al, 2013.) Strabo, (Kyzirakos et al, 2012.) USeekM and Parliament. Our choice fell on Parliament, which has a relatively good balance between ease of installation and use, supports GeoSPARQL and has acceptable performance (although far from the performance offered by the spatial databases (Patroumpas et al.,2014)). On the Parliament server, an RDF graph is created in which are loaded:

– Ontologies used: the domain ontology (prefix cuahsionto :), the application ontology (prefix ccionto :);

– Observational data from Excel files and converted into RDF using vocabularies defined by previous ontology and GeoSPARQL vocabulary for their spatial dimension.



When loading, Parliament automatically performs a number of inferences: RDFS inference over a number of OWL inferences (equivalent classes or properties; inverse properties: symmetrical, transitive or functional). Once loaded, the data can be queried via GeoSPARQL requests transmitted (via http) to the access point of the Parliament server (Jetty server + Joseki).

The interface of the COIN application allows the user to visualize the different sampling sites on an interactive map. The selection of a sampling site on the map provides access to the various comments made at each site. In the dialog box that appears the user has three tabs (FIG. 7) that allow the user:

- To filter the measures to display. Filtering is achieved through the hierarchy of concepts of the domain ontology. The RDFS inferences made in loading data help to automatically add a link ("rdfs: subClassOf") between each superclass of the Cuashi ontology to the descriptors measures. Only comments with a measurement concept as a descriptor are displayed, with a subclass descriptor of the selected concepts.
- To view all the comments for a given date.
- To view all the comments for a given measurement (time series).

A second enrichment of the initial data that allows their representation in the form of an RDF graph allows the ability to link with other external data sets published, respecting the principles of open and linked data. To demonstrate the potential of such enrichment we linked our data represented using our domain ontology with data from DBpedia (Lehmann et al., 2015). With the links to DBpedia , the user can access additional information that the application will look for dynamically in the web data (FIG 7). The DBpedia link allows for a given observation in Wikipedia to provide a description of the item measured.

FIG. 7: *Filtering observations using the domain concepts of hierarchy are displayed as comments regarding Major metals and pesticides*

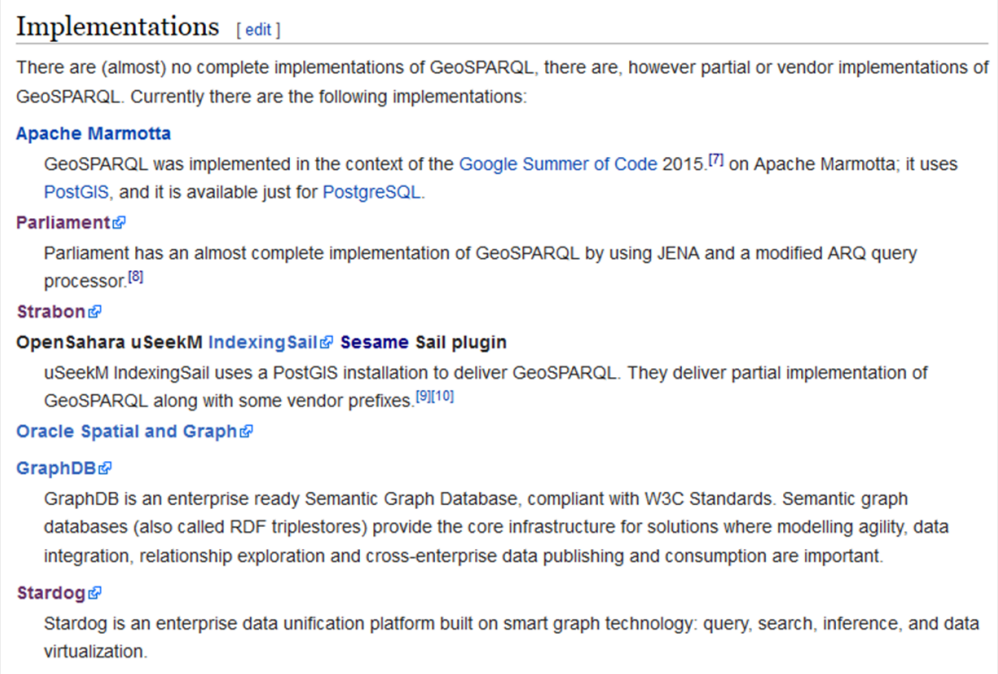
Furthermore, the use of GeoSPARQL allows for queries combining both a spatial component and a semantic component. For example, the following query "Find all the pond sites located in an area for which there is a case for a heavy metal whose value is greater than 15.2 mg / l" can be expressed using a query GeoSPARQL:

- Using the RDF Schema type inferences to select only the "Pond" type sites (Pond) (RDF triplet pattern Site rdf: type ccionto: Pond?) and have an observation corresponding to a heavy metal (triplet pattern? md rdf: type cuahsionto: C2268 where md is a measure descriptor and cuahsionto: C2268 URI concept "Major Metal" in the domain ontology (ontology CUAHSI)). The ontology includes synonyms such as “heavy metals” and “major metals”.

- GeoSPARQL uses quantitative spatial processing capabilities by selecting only sites whose geometry is inside a selected area: triplets patterns geo Site:

hasGeometry siteGeom?. and geo siteGeom: asWKT siteWKT?. possible to recover the geometry of a site in the variable WKT siteWKT; the spatial filter FILTER

(geof: sfWithin (siteWKT, "<http://www.opengis.net/def/crs/OGC/1.3/CRS84> Polygon ((- 132.35 69.74, 75.26 -132.35, -132.35? 69.74, 69.74 -132.35)) "^^ geo: wktLiteral)) allows to select only sites whose geometry is within the selected area.



Implementations [edit]

There are (almost) no complete implementations of GeoSPARQL, there are, however partial or vendor implementations of GeoSPARQL. Currently there are the following implementations:

Apache Marmotta

GeoSPARQL was implemented in the context of the Google Summer of Code 2015.[7] on Apache Marmotta; it uses PostGIS, and it is available just for PostgreSQL.

Parliament

Parliament has an almost complete implementation of GeoSPARQL by using JENA and a modified ARQ query processor.[8]

Strabon

OpenSahara uSeekM IndexingSail Sesame Sail plugin

uSeekM IndexingSail uses a PostGIS installation to deliver GeoSPARQL. They deliver partial implementation of GeoSPARQL along with some vendor prefixes.[9][10]

Oracle Spatial and Graph

GraphDB

GraphDB is an enterprise ready Semantic Graph Database, compliant with W3C Standards. Semantic graph databases (also called RDF triplestores) provide the core infrastructure for solutions where modelling agility, data integration, relationship exploration and cross-enterprise data publishing and consumption are important.

Stardog

Stardog is an enterprise data unification platform built on smart graph technology: query, search, inference, and data virtualization.

FIG. 8 : *GeoSPARQL implementation from wikipedia.*

6. Conclusion

The application described in this article addresses the hydrologic and environmental observation data to analyze and publish these data. The results of COIN's analysis enable us to understand better how the water constituents move through the watersheds. COIN software shows how semantic heterogeneity in heterogeneous data can be easily accessed and interpreted by using the web of data techniques. Beyond the analyses and the computing of data time series the system allows the user's questions to be answered by using ontologies and spatial relations. Additional linked data extend COIN's analysis to quantify human modification and land use effects on hydrologic and hydrochemical water information. The additional information included by linking data techniques increases the ability of the system to answer these types of question and greatly facilitates extensive analysis.

7. Références

Athanasiou et al, 2013 : Athanasiou S., Bezati L., Giannopoulos G., Patroumpas K., Skoutas D. (2013). Market and Reaserach Overview. GeoKnow Deliverable 2.1.1.

Berners-Lee 2010 5* model : http://5stardata.info/en/

Brickley and Guha 2004: Brickley D. Guha R. V. (2004). RDF vocabulary description language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/tr/rdf-schema/.

Couch et al., 2014 : Couch A., Hooper R., Pollak J., Martin M., Seul M. (2014), Enabling Water Science at the CUAHSI Water Data Center, 7th Int'l Congress on Env. Modelling and Software, San Diego, California, USA.

Cyganiak et al, 2014 : Cyganiak B., Wood D., Lanthaler M. (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. http://www.w3.org/tr/rdf11-concepts/.

Hitzler et al., 2012 : Hitzler P., Krötzsch M., Parsia B., Patel-Schneider P.F., Rudolph. S. (2012). OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation 11 December 2012. http://www.w3.org/tr/owl2-primer

Kyzirakos et al, 2012 : K. Kyzirakos, M. Karpathiotakis, M. Koubarakis : Strabon : A Semantic Geospatial DBMS In: Proceedings of the 11th International Semantic Web Conference (2012).

Lim al., 2008 : Lim D., Smol J., Douglas M. (2008). Recent environmental changes on Banks Island (N.W.T., Canadian Arctic) quantified using fossil diatom assemblages. Journal of Paleolimnology, vol. 40, n° 1, p. 385-398.

O&M ISO 19156 : http://www.opengeospatial.org/standards/om

Patroumpas et al.,2014 : Patroumpas K., Giannopoulos G., Athanasiou S. (2014). Towards GeoSpatial semantic data management: strengths, weaknesses, and challenges a head. Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, Texas, USA.

SensorML : http://www.opengeospatial.org/standards/sensorml

WaterML 2.0 : http://www.opengeospatial.org/standards/waterml

W3C 2005, Semantic Web stack : https://www.w3.org/Consortium/techstack-desc.html

Analyse et gestion de l'occupation de places de stationnement par vision artificielle

Karim Hammoudi*, Halim Benhabiles**, Mahmoud Melkemi*, Fadi Dornaika***,****

*Université de Haute-Alsace (UHA), LMIA (EA 3993), MAGE
6 rue des Frères Lumière, 68093 Mulhouse, France
karim.hammoudi@uha.fr, mahmoud.melkemi@uha.fr
http://www.lmia.uha.fr/

**LRDSI Laboratory, Faculty of Sciences, Saad Dahlab University, Blida, Algeria
halim.benhabiles@univ-blida.dz

***Department of CS & AI, University of the Basque Country, 20018 San Sebastiàn, Spain
****IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
fadi.dornaika@ehu.es

Résumé. Cet article présente un système de surveillance basé sur la vision pour le développement de services de gestion de places de parking. Le système présenté est un système adaptable pour l'analyse de places de stationnement dans des parkings de différentes configurations. Dans ce but, des expérimentations ont été menées sous différentes prises de vue en utilisant une caméra connectée à une station de travail mobile. Les résultats obtenus montrent la faisabilité du système dans l'analyse et dans la gestion des emplacements de parking avec des véhicules.

1 Introduction et motivations

Face à l'augmentation continuelle du nombre de véhicule dans le monde, il est nécessaire d'apporter de nouvelles solutions de gestion du trafic. Il devient parfois extrêmement difficile de trouver une place de stationnement disponible, notamment en zones urbaines denses. De plus, ces phénomènes de congestion du trafic engendrent d'autres problèmes de nature économique et écologique. Dans ce travail, nous avons conçu un système de vision visant à simplifier la recherche d'une place de parking par l'analyse en temps réel de places vacantes ; le résultat de cette analyse pouvant être alors communiqué à travers des écrans numériques ou des applications mobiles.

Par ailleurs, nous nous intéressons aux systèmes d'analyse de places de parking basés vision car ils pourraient s'avérer être une alternative intéressante pour compléter ou remplacer les systèmes actuels exploitant des infrastructures physiques (capteurs de pression au sol ou autres systèmes mécaniques), relativement complexes à adapter aux différents types de parking et parfois coûteux en terme de déploiement et de maintenance.

2 État de l'art

Durant cette dernière décennie la communauté de la vision par ordinateur a montré un grand intérêt en ce qui concerne le développement d'approches de surveillance de places de parking. Dans ce contexte, une grande partie des approches existantes sont fondées sur des techniques d'apprentissage (e.g. ; Wu et Z. (2006), Huang et Sheng-Jyh (2010), Ichihashi et al. (2009)). Elles exploitent principalement un classifieur entraîné à partir d'une base de vérité-terrain composée d'images représentant des places de parking (disponibles ou occupées). Généralement, la différence principale entre ces approches réside dans le choix de l'algorithme d'apprentissage et le protocole adopté pour la création de la base de vérité-terrain et pour la construction des données d'apprentissage.

Wu and Zhang (Wu et Z. (2006)) ont proposé une approche basée sur ce modèle. Plus spécifiquement, un modèle d'apprentissage multi-classes est construit en utilisant un classifieur SVM (Support Vector Machine). Dans ce but, ils ont utilisé un ensemble de données d'apprentissage où chaque instance consiste en un vecteur de caractéristiques basé sur la couleur de pixels qui représente un patch de trois places de parking adjacentes. Par conséquence, le classifieur entraîné possède huit sorties correspondant à huit états possibles du patch de véhicules. Les auteurs ont montré que l'apprentissage de l'état des patchs à l'endroit d'une place de parking a permis l'amélioration de la robustesse du classifieur en terme de changement de la luminosité. Plus précisément, le classifieur a atteint 83.75% de succès.

Ching-Chun et Sheng-Jyh (Huang et Sheng-Jyh (2010)) ont également proposé une approche établie sur l'apprentissage pour l'identification de l'état des places de parking rangée par rangée. Leur modèle d'apprentissage est fondé sur un framework bayésien hiérarchique. Il inclut trois étapes à savoir le modèle de classification, le modèle d'adjacence et le modèle sémantique. Les auteurs ont montré que le modèle qu'ils ont proposé permet de gérer efficacement le problème d'occlusion inter-objet. En effet, sur 1500 images de test, les taux de fausses acceptations et de faux rejets sont respectivement 3.2% et 2%.

Similairement aux travaux précédents, Ichihashi et al. (Ichihashi et al. (2009)) ont proposé un système de surveillance de places de parking. La différence principale avec les travaux précédents réside dans l'algorithme d'apprentissage utilisé pour apprendre l'état des places de parking. Dans ce travail, les auteurs ont utilisé un algorithme de classification Fuzzy C-Mean Sadaaki et al. (2008). Pour augmenter la robustesse du système, les auteurs ont pris en considération les conditions climatiques dans les données d'apprentissage. Dans ce but, ils ont collecté des images qui couvrent un large éventail de conditions climatiques incluant notamment la pluie, le soleil, le jour, la nuit. Le classifieur entraîné a montré sa performance à travers un taux de détection d'occupation des places de parking dépassant 80%.

Certains travaux récents proposés dans la littérature évitent l'utilisation de techniques d'apprentissage (voir Suhr et al. (2010) and Shih et Wen-Hsiang (2014)). Leurs mises en oeuvre peuvent s'avérer délicates en raison de la complexité d'implémentation et de paramétrage pour l'obtention de solutions efficaces.

Dans ce qui suit, nous proposons un système qui est simple à implémenter et qui limite l'usage d'opérations coûteuses en mémoire et en calcul tout en préservant des performances comparables à celles des travaux précédemment décrits dans l'état de l'art. Nous présentons ci-dessous une synthèse de travaux publiés concernant les mécanismes de notre système de surveillance de places de parking (Hammoudi et al. (2016b) and Hammoudi et al. (2016a)).

3 Système de surveillance de places de parking présenté

Le système que nous présentons consiste en une application temps réel qui permet d'analyser l'occupation d'un ensemble de places de parking visualisé par une caméra statique. Il n'y a aucune contrainte sur le positionnement de la caméra mis à part le fait qu'elle doive couvrir dans son champ de vision des places de parking ciblées. En conséquence, elle peut être fixée en hauteur sur un poteau ou bien sur un bâtiment, ou même encore derrière la vitre d'un bâtiment.

Afin d'initialiser le système de surveillance, une étape de sélection manuelle des places de parking est réalisée dans une image de référence. Cette image de référence est une image de parking pour lequel les places surveillées sont vacantes ou bien une image d'un parking entièrement non occupée. Ainsi, le système de surveillance sera en mesure d'analyser automatiquement l'occupation des places de parking (analyse spatiale) en comparant chaque couple de places sélectionné sur l'image de référence avec son homologue dans le flux vidéo. Pour améliorer l'analyse des places, nous avons intégré dans notre système un mécanisme d'analyse temporelle. Les étapes et les mécanismes de notre système d'analyse et de gestion des places de parking sont décrits ci-dessous et illustrés en Figure 1.

3.1 Sélection des places de parking

Comme indiqué précédemment, cette étape a pour but de sélectionner manuellement chaque place de parking sur l'image de référence à travers l'interface graphique. À cette fin, un opérateur trace un quadrilatère sur chaque emplacement (i.e. ; étape de focalisation). La taille du quadrilatère est flexible mais doit couvrir au moins une partie significative de la place considérée. Les quadrilatères tracés sur l'image de référence sont projetés directement sur le flux de vidéo puisque nous supposons que la caméra est fixe. Il n'est pas nécessaire de sélectionner l'empreinte complète d'un emplacement de parking pour analyser son occupation. Le fait de focaliser l'analyse sur la partie centrale des emplacements permet de limiter des effets de bords (véhicules mal stationnés) et rend le système plus efficace en temps de calcul.

3.2 Analyse de l'occupation des emplacements

Une fois les places de stationnement sélectionnées, notre système procède au calcul d'une mesure de dissimilarité entre les régions des quadrilatères situés sur l'image de référence et leurs correspondantes sur le flux de vidéo temps réel. Nos études ont conduit à proposer deux stratégies qui mènent à des résultats concluants en ce qui concerne l'analyse de l'occupation des places (occupées ou vacantes).

3.2.1 Stratégie 1

La première stratégie a consisté à exploiter des mesures de dissimilarités connues en photométrie pour leur robustesse (e.g. ; Chen et al. (2003)). Par exemple, nous avons employé des mesures basées sur la différence des intensités entre les couples de quadrilatères (image référence et homologues vidéo) de la façon suivante.

Soient $I_{p_{(i)_{ref}}}$ et $I_{p_{(i)_{flux}}}$ les intensités du $i^{ème}$ pixel situé respectivement dans la région

dans le flux vidéo. La différence d'intensité entre les deux régions notée ΔI est calculée comme suit :

$$\Delta I = \sum_{i=1}^{n} (\left[I_{p_{(i)_{ref}}} - \overline{I}_{ref} \right] - \left[I_{p_{(i)_{flux}}} - \overline{I}_{flux} \right])^2$$

où n est le nombre de pixels contenus dans le quadrilatère tracé autour d'une place de parking et :

$$\overline{I}_x = \frac{1}{n} \sum_{i=1}^{n} I_{p_{(i)_x}}, x\ est\ ref\ ou\ bien\ flux$$

Une grande valeur de ΔI indique une différence significative entre les deux places ce qui signifie potentiellement que la place est occupée. Le contraire est vrai ; une petite de valeur de ΔI signifie que la place est disponible. Cette mesure (i.e. ; Zero-mean Sum of Squared Differences (ZSSD)) commence par calculer l'écart des intensités par rapport à la moyenne pour l'image de référence d'une part et pour l'image vidéo d'autre part. Ceci permet de limiter l'effet de variation d'éclairage qui survient au cours du temps. Cette mesure calcule alors la somme totale des écarts (Sum of Squared Differences (SSD)) puis un seuil global est appliqué pour déterminer le statut d'occupation des emplacements de stationnements. Le score calculé pour chaque emplacement est normalisé par rapport à sa taille. Ce seuil d'occupation est déterminé par expérimentation après un étalonnage initiale.

Malgré l'usage d'une mesure limitant les effets de variations d'éclairage, nous avons observé que l'exposition pouvait changer de manière importante d'une heure à l'autre. Par conséquent, nous avons proposé un algorithme de calcul d'un seuil adaptatif qui rafraichit la valeur du seuil de façon dynamique en retenant la valeur centrale entre les moyennes des scores des emplacements vacants et ceux des emplacements occupés. Cette méthode a permis d'augmenter la robustesse de notre système aux changements de luminosité.

3.2.2 Stratégie 2

La seconde stratégie a consisté à employer une mesure basée sur la présence de contours dans les images analysées. En effet, une place occupée est caractérisée par un nombre de contours important. Pour extraire des contours, nous utilisons un détecteur de Canny qui est basé sur l'analyse de la magnitude de gradients des images. Ensuite, pour comparer les régions d'un emplacement (référence et flux vidéo), nous calculons la différence entre cardinalité de leurs contours $\Delta Card_{\varsigma}$ comme suit :

$$\Delta Card_{\varsigma} = |Card(\varsigma_{ref}) - Card(\varsigma_{flux})|$$

où ς représente l'ensemble des pixels appartenant aux contours extraits respectivement d'une place de référence et sa correspondante dans le flux vidéo. Similairement à ΔI, une grande valeur de $\Delta Card_{\varsigma}$ signifie potentiellement que la place est occupée. Si cet écart $\Delta Card_{\varsigma}$ (nombre de pixels contour apparants) dépasse un certain pourcentage du nombre de pixels du

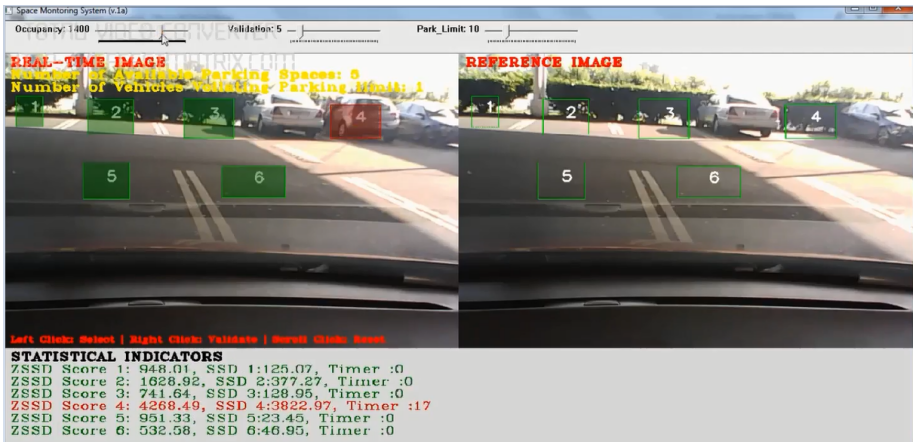


FIG. 1 – *Prototype de l'application développée pour la gestion et la surveillance en temps réel des places de parking.*

Cette seconde stratégie basée sur une analyse de primitives géométriques s'est avérée plus robuste que l'usage seul d'une mesure de similarité exploitant des propriétés photométriques (e.g. ; celle employée Stratégie 1).

Ces deux stratégies comprennent des mécanismes qui ont permis de déterminer en temps réel des statuts d'occupation de places de parking corrects en employant des critères basés image simples à implémenter.

3.3 Filtrage des occultations et horodatage des emplacements

De plus, comme la caméra observe un environnement dynamique incluant des objets en mouvement tels que des piétons ou bien des voitures, les places de parking peuvent être couvertes partiellement ou bien complètement par ces objets a priori pour une courte durée. De ce fait, les scores calculés peuvent être faussés ce qui conduit à une analyse erronée de l'état d'une place de parking. Pour faire face à ce problème, un minuteur est déclenché à chaque fois qu'une place est potentiellement occupé (e.g. ; score d'un emplacement dépassant le seuil fixé). La décision est validée si et seulement si le score de l'emplacement reste stable pendant un certain temps. De plus, le minuteur permet d'horodater chaque emplacement. Ce mécanisme d'horodatage basé vision peut permettre de déclencher une alerte de dépassement du temps alloué. Notre système peut aussi permettre d'informer les usagers lorsqu'un parking est entièrement occupé.

4 Expérimentations et résultats

La Figure 1 montre l'interface de l'application de gestion des places de parking développée. Le panneau situé dans la partie haute comprend trois paramètres ajustables dans le cadre

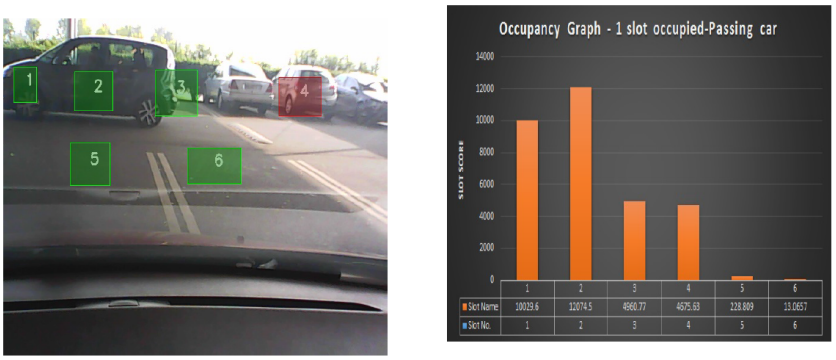


FIG. 2 – *Étude de la robustesse de notre système face aux objets dynamiques occultants. Scène de parking avec un véhicule occultant des emplacements de stationnement (partie gauche). Histogramme des scores (stratégie* 1*) associés à chaque emplacement (partie droite).*

pour la prise de décision (ici 1400). Les paramètres “Validation” et “Park_Limit” sont liés aux minuteurs déclenchés pour chaque emplacement. Un emplacement ayant un score supérieur au seuil pré-défini pendant une durée supérieure au temps de validation (ici 5 secondes) sera considéré comme occupé. L’emplacement sera alors colorisé en rouge. Un emplacement ayant un score inférieur au seuil pré-défini signifie que l’emplacement est vacant. L’emplacement est colorisé en vert. “Park_Limit” permet de déclencher un bip sonore (ou une alerte textuelle) après un temps de stationnement autorisé. L’image de gauche correspond à l’image acquise en temps réel et l’image de droite correspond à une image de référence. Le panneau du bas présente les scores calculés avec différentes métriques ainsi que les minuteurs qui permettent d’obtenir à partir d’analyses visuelles l’horodatage automatique pour chaque emplacement surveillé.

La Figure 2 illustre le phénomène d’occultation des emplacements de stationnement qui peut être occasionné par des objets mobiles. Dans l’histogramme des scores, nous observons que les valeurs des scores sont élevés (supérieures au seuil global pré-défini) mais les couleurs des emplacements n’ont pas changé grâce au paramètre de validation. De cette façon, notre système est robuste aux occultations de cette nature.

Par ailleurs, la Figure 3 expose un phénomène d’ombrage qui peut parfois survenir et mettre en défaut un système de ce type. Le véhicule à l’emplacement 2 se gare. L’ensoleillement important fait apparaître une zone d’ombre significative dans l’emplacement 1 (colorisé en bleu). Ce cas illustre l’importance d’ajuster le seuil dans le cadre de la stratégie 1.

D’autres expérimentations de notre système ont été menées en appliquant la stratégie 1. Ces dernières ainsi que les expérimentations réalisées en appliquant la seconde stratégie sont

FIG. 3 – *Effet d'ombre sur un emplacement de stationnnement. Le stationnement d'un véhicule a produit une ombre sur un emplacement adjacent causant alors une mauvaise détection (partie gauche). Histogramme des scores (stratégie* 1*) associés à chaque emplacement (partie droite).*

5 Conclusions et perspectives

Nous avons présenté un système basé vison pour l'analyse et la gestion des places de parking. Ce système peut être employé pour automatiquement connaître la disponibilité de places de stationnement. Il peut être également employé pour contrôler le temps d'occupation des places par un horodatage visuel. Ce système exploite des mécanismes d'analyse d'occupation simples à mettre en oeuvre par l'utilisation de métriques image basées sur l'étude de propriétés photométriques ou géométriques à partir de données brutes (e.g. ; images obliques non prétraitées). Des recherches (Hammoudi et al. (2016b) and Hammoudi et al. (2016a)) ont permis de valider le fonctionnement des mécanismes d'analyse présentés pour chacune des stratégies.

Les résultats montrent une robustesse aux objets dynamiques occultants ainsi qu'aux faibles variations lumineuses. En perspective, nous escomptons confronter les deux stratégies et réaliser des expérimentations dans des conditions rudes (intempéries, faible éclairage, ombrages par diverses sources). Dans ce sens, nous menons des expérimentations pour rafraîchir à différents moments les intensités des emplacements vacants de l'image de référence pour limiter l'effet des changements d'éclairage.

Références

Chen, J., C. Chen, et Y. Chen (2003). Fast algorithm for robust template matching with m-estimators. *IEEE Transactions on Signal Processing 51*, 230–243.

Hammoudi, K., H. Benhabiles, A. Jandial, F. Dornaika, et J. Mouzna (2016a). Developing a vision-based adaptive parking space management system. *International Journal of Sensors,*

Hammoudi, K., H. Benhabiles, A. Jandial, F. Dornaika, et J. Mouzna (2016b). Self-driven and direct spatio-temporal mechanisms for the vision-based parking slot surveillance. In *Science and Information Conference (SAI Computing)*, pp. 1327–1329. IEEE.

Huang, C. et W. Sheng-Jyh (2010). A hierarchical bayesian generation framework for vacant parking space detection. *IEEE Transactions on Circuits and Systems for Video Technology 20*, 1770–1785.

Ichihashi, H., A. Notsu, K. Honda, T. Katada, et M. Fujiyoshi (2009). Vacant parking space detector for outdoor parking lot by us-ing surveillance camera and fcm classifier. In *International Conference on Fuzzy Systems)*, pp. 127–134. IEEE.

Sadaaki, M., H. Ichihashi, et K. Honda (2008). *Algorithms for fuzzy clustering: methods in c-means clustering with applications.* Springer.

Shih, S. et T. Wen-Hsiang (2014). A convenient vision-based system for automatic detection of parking spaces in indoor parking lots using wide-angle cameras. *IEEE Transactions on Vehicular Technology 63*, 2521–2532.

Suhr, J., H. Jung, K. Bae, et J. Kim (2010). Automatic free parking space detection by using motion stereo-based 3D reconstruction. *Machine Vision and Applications 21*, 163–176.

Wu, Q. et Y. Z. (2006). Parking lots space detection. machine learning). Technical report, Carnegie Mellon University.

Summary

This paper presents mechanisms of a vision-based monitoring system for the development of parking space management services. The presented system is an adaptive system dedicated to the analysis of parking spaces having various configurations. In this way, experiments have been conducted under several fields of view by using a camera connected to a mobile workstation. Obtained results show the feasibility of this system for analyzing and managing slots of car parkings.

Application mobile de visualisation 3D temps réel des compétitions sportives en extérieur (LIS3D)

Thierry Pagès *, Régis Beuscart **,*** Julien Soula **,*** Anne Quesnel-Barbet**,***

*SAS Live In Sport, 1, rue Cournut, 34570 PIGNAN, France
neo@liveinsport.com
http://www.liveinsport.com
**Univ. Lille, EA 2694-Santé Publique, épidémiologie et qualité des soins,
F-59000 Lille, France
rbeuscart@univ-lille2.fr
http://cerim.univ-lille2.fr
*** CHU Lille, Maison Régionale de la Recherche Clinique, Pôle de Santé Publique,
F-59000 Lille, France
aquesnel@univ-lille2.fr
http://www.chru-lille.fr
http://quesnela.free.fr

1 Introduction - Etat de l'art

Une solution innovante de dossards connectés permettant la sécurisation du parcours sportif et le suivi en direct de l'ensemble des participants est proposée aux organisateurs d'épreuves sportives. L'innovation repose sur le suivi par système de positionnement global (GPS) et la visualisation bidimensionnelle (2D) et tridimensionnelle (3D) temps réel.

Cette solution correspond à deux besoins des organisateurs non satisfaits. Le premier besoin est de renforcer la sécurité et les délais d'intervention du fait d'un nombre croissant de participants. Le deuxième besoin est de renouveler leur offre dans le contexte numérique actuel et permettre de mieux valoriser les compétitions sportives.

Cette solution permet également au public de suivre en direct ou en différé [1] l'ensemble des participants (500 et plus) ou bien un participant en particulier. Le direct peut être regardé sur place ou à distance grâce à un client web connecté à Internet (écran géant, mobile ou téléviseur).

Le suivi GPS portable existe depuis une dizaine d'années, mais n'a été utilisé qu'à petite échelle (c'est-à-dire que le suivi concerne moins de cinquante participants) du fait de couvertures réseau insuffisantes des opérateurs. Il est à noter cependant, qu'en 2016, deux acteurs industriels, Tech4Race et Race-Result, ont annoncé entreprendre le développement de produits, pouvant équiper plusieurs centaines de participants simultanément venant ainsi concurrencer

1. (France Régionale Télévision Poitou-Charente (FR3), 2016) : `https://drive.google.com/open?id=0B7QU9mZfFEEgRkhlMjRHalhxZVk`

notre produit en cours de développement [2] [3]. Enfin, il faut préciser que le développement récent de l'Internet des Objets (IOT) et de la généralisation de l'itinérance des réseaux mobiles a permis le développement de sociétés spécialisées dans la gestion des réseaux multi-opérateurs à des prix adaptés aux communications sans intervention humaine (machine to machine ou M2M) [4] [5].

A notre connaissance, la visualisation 3D virtuelle et dynamique de suivi en temps réel d'une compétition sportive semble peu représentée voire inexistante. Nous rappelons toutefois que le quotidien anglais « The Guardian » a publié en 2015 un ensemble de vues statiques 3D de la montée de l'Alpes d'Huez du Tour de France [6].

2 Objectifs, matériels et méthodes

A travers nos travaux réalisés en 2016 et en cours, nous comptons nous démarquer de la concurrence par la qualité du traitement et la restitution visuelle 2D et 3D des données GPS grâce au procédé des dossards connectés associé à une plateforme informatique de traitement de l'information en temps réel.

Tout d'abord, le dossard connecté se compose d'un procédé électronique intégré au dossard du participant. Il comprend une puce GPS et une carte de type module d'identité par abonnement « Subscriber Identity Module (SIM) » télétransmettant grâce au protocole de transfert des données sur internet (TCP/IP) à des fréquences comprises entre 30s et 60s et assurant une autonomie du système sur un créneau horaire de 10h à 20h.

En complément, nous avons développé plusieurs modes de visualisation 2D : à l'aide de l'interface de programmation (API) Google Maps [7], Fig. 1, ou à l'aide de projections 2D, Fig. 2, toutes obtenues à partir des flux temps réel des coordonnées GPS des participants pendant l'épreuve. Nous avons dû résoudre les problèmes méthodologiques de calcul de la distance parcourue par le participant au regard de ses sorties de parcours liées aux imprécisions des positions GPS et des tracés des parcours, à la perte de repère (ou la tricherie) des participants.

Nous nous sommes fixés également pour objectif en 2017 de créer une application mobile où le spectateur bénéficie d'une immersion virtuelle 3D de l'épreuve en temps réel comparable à celle, réelle, du Tour de France. Pour cela, une modélisation 3D restreinte aux bornes du parcours (parallélépipède rectangle) a été réalisée. Nous avons utilisé un environnement logiciel open source QGIS [8] et le plugin Qgis2threejs [9] de la librairie ThreeJS [10] (Fig. 3). À cet environnement, nous avons associé des couches de données relatives à l'information géographique comprenant des modèles numériques de terrain issus des bases de données ALTI ou

2. (Race result Allemagne et al., 2017) : `https://www.raceresult.com/fr-fr/home/index.php`
3. (Tech4race.com, 2017) : `http://tech4race.com/wordpress/`
4. (Emnify.com, 2017) : `https://www.emnify.com/features/`
5. (Matooma.com, 2017) : `http://www.matooma.com/fr`
6. (?) : `https://www.theguardian.com/sport/ng-interactive/2015/jul/23/tour-de-france-the-climb-of-alpe-dhuez-interactive`
7. (Google, 2017) : `https://enterprise.google.com/intl/fr/maps/?utm_source=google&utm_medium=cpc&utm_campaign=2016-geo-emea-endor-gmedia-search-crossreg-`
8. (QGis, 2017) : `https://www.qgis.org/fr/site/`
9. (Akagi, 2017) : `https://github.com/minorua/Qgis2threejs`
10. (mrDoob, 2017) : `https://github.com/mrdoob/three.js/` `https://threejs.org/` `http://mrdoob.com/`

RGE ALTI, de l'orthophotographie Haute Résolution ou de l'orthophotographie par protocole de communication « (WMS) service de cartes web » [11] [12].

3 Résultats / démonstration

C'est dans ce contexte que nous proposons une analyse descriptive et une démonstration de visualisation 3D d'épreuves sportives suivies en 2016 parmi lesquels un TRAIL de 53km (département du Cantal), un SWIM & RUN de 44km (département des Pyrénées Orientales), un marathon (département de la Charente) et une CYCLO-SPORTIVE de 164km (département des Bouches du Rhône) avec un nombre de participants compris entre 100 à 500 suivant les épreuves.

La qualité visuelle des blocs 3D a été obtenue grâce aux paramétrages de QGIS et du plugin Qgis2threejs et aux couches de données associées. Le rendu global obtenu (bloc 3D avec suivi dynamique des sportifs ou agents) permet aujourd'hui de simuler une caméra en mouvement à une cinquantaine de mètres au-dessus du parcours [13].

4 Perspectives et verrous

Les prochaines étapes, qui pourront faire l'objet de discussions, consistent à pouvoir descendre la caméra au niveau des participants avec un rendu (bloc 3D avec suivi dynamique des sportifs) de l'environnement immédiat du participant le plus réaliste possible tout en gardant une fluidité de l'animation. Cela pose vraisemblablement des problématiques d'annotations et de modélisation de cet environnement associées à un niveau de calcul compatible avec une application mobile.

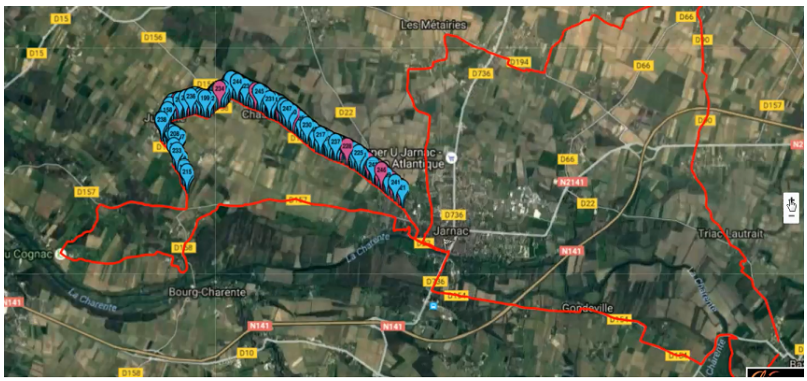


FIG. 1 – *Vue Google Map (Marathon du Cognac, Sources : Google (Google, 2017), LiveInSport).*

11. (CRAIG, 2017) : https://www.craig.fr/
12. (IGN, 2017) : http://professionnels.ign.fr/donnees
13. (Pagès, 2017) http://liveinsport.com/ https://www.facebook.com/LiveinSport-354227534966467/

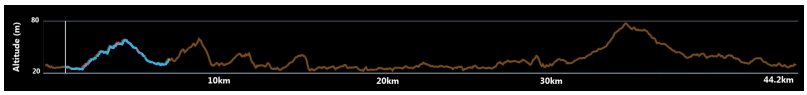


FIG. 2 – *Vue profil 2D (Marathon du Cognac, Source : LiveInSport (Pagès, 2017)).*

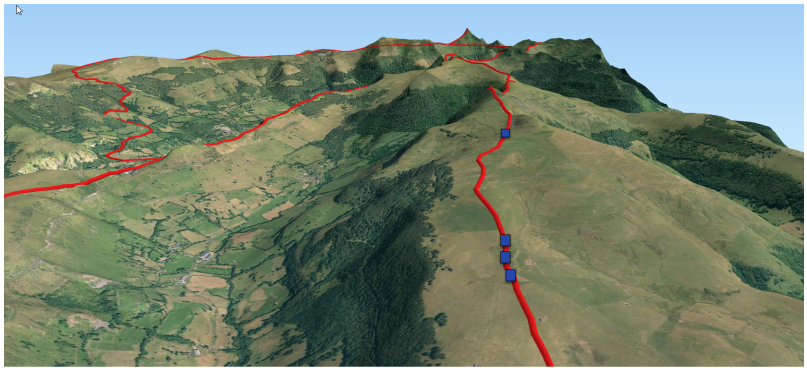


FIG. 3 – *Vue 3D (Trail de la Pastourelle, Sources : MNT-IGN (IGN, 2017), WMS-CRAIG (CRAIG, 2017), Univ.Lille, CHU Lille, LiveInSport (Pagès, 2017)).*

Références

Akagi, M. (2017). Qgis2threejs plugin exports terrain data, map canvas image and vector data to your web browser. [qgis plugin].

CRAIG (2017). Centre national auvergnat de l'information géographique.

Emnify.com (2017). Iot & m2m connectivity management features.

France Régionale Télévision Poitou-Charente (FR3) (2016). Reportage marathon du cognac (02' 03"). Temps du reportage : 2mn 03.

Google (2017). Api google maps.

IGN (2017). Les données de l'institut national géographique et forestier.

Matooma.com (2017). Connectez et gérer vos objets par carte sim.

mrDoob (2017). Threejs : a javascript 3d library which makes webgl simpler.

Pagès, T. (2017). Live in sport.

QGis (2017). Système d'information géographique libre et open source. qgis is a free, open source, cross platform (lin/win/mac) geographical information system (gis).

Race result Allemagne, G. Chalamet, et Race result France (2017). Tracking et chronométrage avec la même puce.

Tech4race.com (2017). Le suivi en temps réel des événements sportifs outdoor.

Summary

Sport event competitors have been fully equipped with GPS/GSM devices. Dynamic 3D representations of resulting GPS data flux will be shown and discussed focusing on 3D model quality and mobile computational abilities.

Index

Akdag,Herman, 41

Bahoken, Françoise, 64
Benhabiles, Halim, 81
Beuscar, Régis, 89
Bosch,Martì, 2
Bouveret, Sylvain, 68

Curé,Olivier, 24

De Runz, Cyril, 41
Dornaika, Fadi, 81

Faiz,Sami, 41
Fenet,Serge, 2

Genoud, Philippe, 68
Gervasoni,Luciano, 2
Grabar,Natalia, 37
Guez,Alain, 49

Hammoudi,Karim, 81
Hobus, Kim, 68

Kergosien,Éric, 37
Khallfi, Besma, 41

Lagesse,Claire, 49
Lhuez, Jérémy, 24

Mathieu,Roche, 1
Melkemi, Mahmoud, 81

Pagès, Thierry, 89

Quesnel-Barbet, Anne, 89

Rousseaux,Francis, 49

Severo,Marta, 37
Sturm,Peter, 2

Valmond-Leblanc, Nathalie, 37

Ziebelin, Danielle, 68