HAL
open science

Étude des quartiers : défis et pistes de recherche

Loïc Bonneval, Fabien Duchateau, Franck Favetta, Aurélien Gentil, Mohamed Nader Jelassi, Maryvonne Miquel, Ludovic Moncla

To cite this version:

Loïc Bonneval, Fabien Duchateau, Franck Favetta, Aurélien Gentil, Mohamed Nader Jelassi, et al.. Étude des quartiers : défis et pistes de recherche. Conférence Extraction et Gestion de Connaissances 2019 (EGC2019) Atelier DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis (DAHLIA 2019) -, Jan 2019, Metz, France. hal-02005923

HAL Id: hal-02005923

https://hal.science/hal-02005923

Submitted on 4 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Étude des quartiers : défis et pistes de recherche

Loïc Bonneval*, Fabien Duchateau**, Franck Favetta**, Aurélien Gentil*,
Mohamed Nader Jelassi**, Maryvonne Miquel**, Ludovic Moncla**

*Centre Max Weber, Université de Lyon, France
prénom.nom@univ-lyon2.fr
**LIRIS UMR5205, Université de Lyon, France
prénom.nom@liris.cnrs.fr

Résumé. Le projet Home In Love (HiL) s'intéresse à la recommandation de biens immobiliers, en particulier dans le cas où l'on ne connait pas sa future ville de résidence (e.g., mutation professionnelle). Si le choix d'un logement est facilité par les nombreuses ressources disponibles (e.g., sites web avec photos, visites virtuelles), cela reste compliqué de se faire une idée concrète des quartiers où se trouvent les logements disponibles. L'un des enjeux concerne donc la description et la comparaison de quartiers selon les domaines d'application (e.g., recherche immobilière, étude sociale, recensement du patrimoine). Cet article décrit les défis et les pistes de recherche (en informatique) liés à cette étude des quartiers.

1 Introduction

Le projet pluridisciplinaire Home In Love [1] (HiL) a pour objectif la recommandation de biens immobiliers pour des personnes ne connaissant pas ou peu leur future ville de résidence. Dans cette optique, le travail de caractérisation des quartiers potentiellement recommandés s'avère central.

Durant ces dernières décennies, les quartiers ont fait l'objet d'une attention particulière en sciences humaines et sociales (Authier et al. (2007)). Les cas d'utilisation liés au quartier occupent une place importante à l'ère des *"smart cities"*, avec des projets de simulation pour l'efficacité énergétique (Perez et al. (2016)), de gestion du patrimoine (Devernois et al. (2014)), d'étude d'impact de l'environnement sur la santé (Morland et al. (2002)) et la mobilité (Chaix (2014)), ou encore de reconstitution numérique du patrimoine (Pardoen (2015)). Par rapport au contexte du projet HiL, on peut citer des applications centrées sur la recommandation de quar-

Ce travail a été réalisé au sein du LABEX IMU (ANR-10-LABX-0088) de l'Université de Lyon, dans le cadre du programme "Investissements d'Avenir" (ANR-11-IDEX-0007) de l'Etat Français, géré par l'Agence Nationale de la Recherche (ANR).

1. http://imu.universite-lyon.fr/projet/hil

tiers comme Kelquartier [2], Cityzia [3], ville-ideale [4], vivroù [5] ou encore le site DataFrance [6] qui intègre des indicateurs fournis par l'INSEE et par d'autres sources (e.g., IGN, L'Express). Pour étudier ces quartiers, il est important de les décrire avec précision et de pouvoir les comparer.

La notion de quartier (i.e., définition, délimitations, perception) est plurielle et sujette à différentes interprétations (Authier et al. (2007)), ce qui rend complexe leur utilisation pour un domaine d'application donné. Des projets s'attachent à décrire les quartiers d'une ville selon un point de vue, par exemple celui de ses habitants (Authier (2008)). Généralement, ces descriptions de quartiers se concentrent sur quelques villes, et ne permettent pas un traitement automatique à plus grande échelle. De même, la délimitation des quartiers n'est pas toujours disponible, et de nombreux travaux s'intéressent à la détection automatique de ces « frontières à l'interprétation variable ». Un défi important concerne donc la méthodologie à adopter pour décrire des quartiers et leur délimitation. Dans de nombreux cas d'utilisation, comme le regroupement de quartiers similaires, la classification des quartiers selon leur fonction, ou la recommandation, il est utile de comparer des quartiers, i.e., de déterminer si deux quartiers possèdent des caractéristiques semblables. Ce défi repose généralement sur l'utilisation d'un processus ou d'un algorithme de comparaison qui exploite la description des quartiers. Enfin, il est crucial de vérifier, en particulier dans le cas des traitements automatisés, la qualité des résultats produits par rapport à la réalité. Par exemple, on souhaite contrôler qu'un algorithme de regroupement ait bien associé des quartiers similaires ou vérifier qu'un algorithme de recommandation ait proposé des quartiers pertinents à un utilisateur. Pour cela, il faut réfléchir à l'évaluation et notamment la manière de valider les résultats, ce qui peut se révéler très complexe au vu de la perception et du flou qui entourent ces quartiers.

Dans cet article, nous dressons un panorama de l'étude des quartiers sous un angle informatique, en identifiant et décrivant les pistes de recherche principales que sont la description, la délimitation, la comparaison et l'évaluation en lien avec les quartiers.

2 Pistes de recherche

Dans cette section, nous décrivons quatre pistes de recherche que nous considérons comme cruciales pour l'étude des quartiers : la description, la délimitation, la comparaison et l'évaluation.

2.1 Description de quartiers

Afin de caractériser les quartiers, une première étape consiste à déterminer les informations pertinentes permettant leur description (pour un domaine d'application). Cette sélection nécessite des échanges entre les différents acteurs (chercheurs en SHS, informaticiens, etc.) afin d'établir la liste des données utiles pour l'application. Par exemple, l'INSEE quadrille le territoire français en IRIS [7] (Ilots Regroupés pour l'Information Statistique). Il existe plusieurs types d'IRIS, ceux d'habitation (entre 2000 et 5000 résidents), ceux d'activités (e.g., un

2. http://www.kelquartier.com/
3. http://www.cityzia.fr/
4. http://www.ville-ideale.fr/
5. http://www.vivrou.com/
6. http://datafrance.info/
7. http://www.insee.fr/fr/metadonnees/definition/c1523

campus universitaire) et ceux divers peu habités (e.g., un parc). L'INSEE fournit pour chacun des 50 000 IRIS de nombreux indicateurs (e.g. le nombre d'épiceries, la répartition par type de logement ou par catégorie socio-professionnelle). Certaines données seront agrégées (e.g., somme du nombre de boulangeries et du nombre d'épiceries pour représenter les "commerces de proximité"). Si l'on compare des quartiers très différents, il faut réfléchir à la normalisation des données (e.g., selon leur superficie, leur population, leur densité). Les données choisies peuvent donc être traitées de façon absolue (e.g., nombre de boulangeries dans le quartier) ou relatives (e.g., densité de commerces par rapport à celle du reste de l'agglomération). L'avantage de cette deuxième option est de rapprocher des quartiers d'agglomération différentes mais ayant une position comparable (e.g., un « quartier central » n'aura pas les mêmes caractéristiques à Lyon ou dans une ville moyenne, mais pourra avoir le même « rôle » ou fonction dans les deux cas). Une tendance récente concerne la collecte de données des réseaux sociaux, comme les *tweets*, les *likes* ou les *checkins* qui sont ou peuvent être localisés. Un corpus d'images habilement constitué peut également servir à représenter un quartier (Kennedy et Naaman (2008)). Enfin, le choix des informations pertinentes peut être actualisé et est dépendant de la disponibilité des données.

En effet, une fois les données identifiées, une deuxième difficulté est la recherche et la disponibilité des sources de données. Le web facilite aujourd'hui cette étape, et le mouvement "Données Ouvertes" (e.g., `data.gouv.fr`) permet de disposer de nombreux jeux de données. Cependant, il convient de vérifier la qualité de ces données (e.g., provenance, dates de mise à jour, utilisation par d'autres applications). Les jeux de données à caractère sensible ou ayant une potentielle valeur monétaire peuvent être plus difficiles à obtenir. Par exemple, pour la recherche immobilière, une information utile est le prix de vente (ou de location) d'un bien. Or cette donnée sur le prix est rarement disponible ou reste incomplète (e.g., uniquement fournie pour les plus grandes villes), ou elle n'a pas le bon niveau de granularité (e.g., au niveau d'un département, mais pas d'une ville ou du quartier). Sur les aspects techniques, certaines API peuvent être indisponibles à certaines périodes ou avoir des limitations en terme de requêtes. Enfin, les sources de données s'accompagnent d'une licence d'utilisation : des restrictions peuvent donc empêcher leur utilisation.

Le troisième problème est l'intégration de ces données hétérogènes pour faciliter leur exploitation (Halevy et al. (2006)). Les concepts qui décrivent les données sont habituellement comparés en utilisant des techniques d'appariement de schémas ou d'ontologies (Bellahsène et al. (2011)). Pour les données elles-mêmes, le *record linkage* ou appariement de données permet de détecter les informations équivalentes et donc d'éviter la redondance (Christen (2012)). Des données plus complexes (e.g., textuelles, multimédia) utiliseront des outils de détection d'entités nommées (Shen et al. (2015)), d'extraction d'informations (e.g., spatio-temporelles dans Strötgen et al. (2010)) ou des annotateurs automatiques (Zhang et al. (2012)). Dans le cas des données sur les IRIS de l'INSEE, les indicateurs sont fournis dans des dizaines de fichiers, qui ne sont pas tous organisés de la même manière (interprétation différente des concepts, hétérogénéité des libellés, regroupement ou division d'IRIS, etc.). Les solutions pour l'intégration peuvent être manuelles (e.g., saisie des données dans un tableur, codage d'un script) ou automatisées en utilisant des outils tels qu'OpenRefine [8], Talend [9], Karma ((Gupta et al., 2012)),

8. `http://openrefine.org/`
9. `http://fr.talend.com/products/data-integration/`

BigGorilla (Chen et al. (2018)). La description des quartiers à partir de différentes sources agrégées doit donc être stockée dans un format approprié (e.g., GeoJSON).

2.2 Délimitation des quartiers

Comme indiqué précédemment, la définition d'un quartier n'est pas fixée et elle dépend du contexte (e.g., économique, historique, politique) et de la perception (e.g., point de vue de l'administration, des habitants). La délimitation d'un quartier (e.g., liste de coordonnées géographiques formant un polygone) fait partie de sa description, mais nous la considérons à part car c'est un défi particulièrement complexe et qui apparaît comme optionnel dans certains cas d'utilisation. Selon les données disponibles, la délimitation d'un quartier consiste à identifier son contour en :

— Se basant sur les définitions de l'administration ;
— Regroupant des zones plus petites (e.g., les IRIS) pour former un quartier, ou en divisant une zone (e.g., un arrondissement, un grand IRIS) en plusieurs quartiers ;
— Exploitant des cartes ou des systèmes d'information géographique (SIG) ;
— Exploitant des sources semi-structurées, textes, ou images (aériennes) ;
— Exploitant les *likes* et *checkins* de réseaux sociaux comme Foursquare, ;
— Réalisant des enquêtes auprès des populations.

L'administration met aujourd'hui à disposition certaines informations sur les quartiers tel que le contour cartographique. C'est par exemple le cas dans le cadre de la définition des quartiers prioritaires de la politique de la ville [10]. Nous pouvons également citer des communes (e.g., Sèvres, Nanterre, Saint-Quentin) qui mettent à disposition les informations de délimitations de leurs quartiers sur `data.gouv.fr`. La méthodologie détaillant le découpage n'est pas toujours accessible, et ces données ne prennent en compte que le point de vue administratif. Mais ces informations peuvent servir de base à une étude ou à un comparatif par rapport à d'autres perceptions.

Un quartier peut résulter d'un regroupement ou d'une division d'autres unités géographiques. Dans de nombreux travaux, un IRIS est tout simplement assimilé au quartier. Par exemple, dans Préteceille (2009), l'IRIS est considéré comme l'unité spatiale la plus pertinente pour étudier la ségrégation car il correspond mieux au quartier vécu des habitants. Cependant, selon l'objectif de la recherche, cette simplification n'est pas toujours opératoire. En effet, cette vision s'oppose à celle des travaux de Maurin (2004), dans lesquels un découpage plus fin est préconisé (e.g., similaire à l'ancienne unité géographique appelée îlot [11] et définie par l'INSEE comme « *un pâté de maison en zone dense ou un ensemble limité par des voies en zone périphérique* »). Dans Barret et al. (2019), les auteurs utilisent également les IRIS mais prennent en compte le voisinage (IRIS adjacents à un IRIS) pour lisser les données, caractériser plus finement les espaces étudiés et s'approcher au mieux de la notion de quartier. Enfin, il est possible de regrouper ou de diviser des unités géographiques qui ne sont pas adaptés au domaine d'application. Dans Actif et al. (2013), des « grands quartiers » sont ainsi créés à partir des IRIS, mais la méthode n'est pas discutée. Le problème avec cette réorganisation des unités géographiques est la validité des données, qui ne sont donc plus au même niveau et peuvent éventuellement introduire des biais dans les analyses.

10. `http://www.data.gouv.fr/fr/datasets/5a561801c751df42d7fca9b6/`
11. `http://www.insee.fr/fr/metadonnees/definition/c1656`

Plusieurs fournisseurs géographiques proposent de visualiser les quartiers. Chez Google Maps, des noms de quartiers avec leur contour apparaissent dans les villes, mais la méthodologie utilisée pour les déterminer n'est pas décrite. Les noms de quartiers sont également présents sur Bing Maps ou Here Maps, mais seul un point représente le quartier, ce qui est insuffisant. OpenStreetMap [12] et Wikimapia [13] offrent pour certaines zones un découpage en quartiers. Comme ces sites sont collaboratifs, la délimitation est de fait subjective et potentiellement incomplète à l'échelle d'une ville. Dans la base de données géographique Geonames [14], les quartiers sont généralement présents à travers des unités géographiques plus importantes. Par exemple, le quartier « *Serin* » à Lyon est associé au quatrième arrondissement, car il est cité dans la page Wikipédia de cet arrondissement. De plus, la visualisation d'un quartier (sur un fond de carte Google Maps) est très approximative car représentée par un rectangle.

Dans les sources de données semi-structurées, Wikipédia figure parmi les sites les plus utilisés. On y trouve une page [15] « catégorie : quartier de ville en France ». Mais la délimitation du quartier accompagne rarement ces descriptions, qui de plus ne concernent que quelques dizaines de villes. Le « Linked Open Data » [16] relie sémantiquement les entités de nombreux jeux de données. Dans la base de connaissances Wikidata, les quartiers sont de type « district » [17], dont la définition trop générale (« type de division administrative existant dans certains pays, de tailles variables allant du quartier à la région ») illustre là encore la difficulté d'uniformiser cette notion de quartier et ne permet pas d'exploiter cette source de données sans traitement spécifique. Les sources de données textuelles sont largement disponibles sur le web, en particulier des sources touristiques ou patrimoniales. Par exemple, le site `http://www.patrimoine-lyon.org/` décrit les (vieux) quartiers de Lyon avec de nombreux détails spatiaux et cartographiques, mêlant textes, cartes et photographies. Une façon d'exploiter de telles sources est détaillée dans Brindley et al. (2014), où des noms de quartiers sont extraits d'un grand volume de documents, et reliés ensuite à des codes postaux. Cela permet d'étudier l'évolution du contour des quartiers au fil du temps. Pour exploiter des documents textuels, il est nécessaire d'appliquer des méthodes de traitement automatique du language (TAL), notamment pour l'extraction d'informations géographiques (e.g., lieux, bâtiments, monuments, rues), comme le soulignent Miller et Han (2009); Leidner et Lieberman (2011). Les annonces immobilières ainsi que les offres de location provenant de sites tels que AirBnB [18] comportent parfois des descriptions détaillées du quartier, avec le point de vue d'un expert ou d'un habitant (Guérois et Madelin (2017)). Dans le rapport de Tang et Sangani (2015), il est question de prédire le quartier et le prix d'une annonce (pour la ville de San Francisco), et il est envisageable de déterminer les limites d'un quartier en fonction de ces prédictions. Enfin, les images aérienne et/ou satellite sont l'objet de recherches pour la détection automatique d'objets (e.g., véhicules) ou de phénomènes (e.g., déforestation, glissement de terrain). Les bâtiments peuvent être détectés et classés par type en les comparant à des exemples annotés (Du et al. (2015)) ou en exploitant les images « Street View » (Kang et al. (2018)). La détection de quartiers à partir d'images reste un défi majeur, mais des progrès sont réalisés,

12. `http://www.openstreetmap.org/`
13. `http://wikimapia.org/`
14. `http://www.geonames.org/`
15. `http://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Quartier_de_ville_en_France`
16. `http://linkeddata.org/`
17. `http://www.wikidata.org/wiki/Q149621`
18. `http://www.airbnb.fr`

par exemple pour détecter des villages « absorbés » par l'expansion rapide des grandes villes chinoises (Huang et al. (2015)).

Une autre piste pour détecter les contours est l'exploitation des données de géolocalisation issues de réseaux sociaux, mouvement qui a connu un engouement certain depuis une quinzaine d'années. Avec Livehoods [19], la délimitation de quartiers est possible en analysant les *checkins* des réseaux sociaux (Cranshaw et al. (2012)). Dans Hoodsquare (Zhang et al. (2013)), les activités des habitants sont analysées et associées à l'une des 300 catégories de la hiérarchie de Foursquare pour en déduire le ou les raisons de fréquenter un quartier, et d'en déterminer les contours. Les aspects temporels (e.g., période la journée) ainsi que les activités des touristes sont également pris en compte. À Beijing, les fonctions d'une zone géographique (e.g., résidentiel, éducation, diplomatique, culturel ou historique) sont découvertes en considérant ce problème comme la découverte de thématiques pour un document (Yuan et al. (2012)). Les zones sont regroupées en fonction de la distribution des points d'intérêt (*via* un algorithme de *clustering*), et les activités humaines de chaque groupe servent à identifier sa ou ses fonctions principales. Comme l'indiquent les auteurs de ces travaux, les données peuvent avoir un biais (e.g., plus forte probabilité de *liker* un bar que son lieu de travail).

Les enquêtes auprès de populations sont la dernière piste pour identifier les contours de quartiers. Des travaux permettent à l'utilisateur de dessiner sur une carte interactive son quartier (ou « région ») d'intérêt, et différentes informations sont alors affichées comme les activités principales ou points d'intérêt représentatifs ((Kumar et al., 2015)). Les habitants peuvent aussi décrire les points d'intérêt représentatifs de leur quartier afin d'en déduire ses frontières (Berjawi et al. (2013)). Pour obtenir des résultats valides, et en particulier des délimitations « sans trous », il faut récolter l'avis de nombreuses personnes. Côté sciences humaines de sociales, il y a peu d'enquête systématique sur ce thème. Seuls les travaux de Pan Ké Shon (2005) exploitent une question de l'enquête permanente « *condition de vie de 2001* », menée par l'INSEE (la question étant "pouvez-vous me dire en quelques mots ce que représente votre quartier pour vous ?").

2.3 Comparaison de quartiers

Une troisième piste de recherche concerne la comparaison de quartiers. En effet, de nombreux cas d'application ont besoin d'établir la ressemblance entre quartiers. Ce processus repose essentiellement sur la description des quartiers, dont les données serviront à mesurer un degré de similarité.

Dans un premier temps, les données utiles à la comparaison doivent être sélectionnées parmi toutes celles qui caractérisent les quartiers. Le voisinage d'un quartier peut être exploité, par exemple pour estimer le climat urbain ou la co-occurrence d'activités dans une zone.

Différentes techniques permettent d'établir des comparaisons entre objets. La similarité cosine et la mesure de Jaccard sont les algorithmes les plus connus pour réaliser cette opération. Ils permettent de calculer directement le degré de ressemblance entre deux quartiers décrits comme des vecteurs de valeurs (Yu et al. (2016) et Zhang et al. (2013)). La distance *Earth mover* (EMD) mesure l'effort pour "transformer" un quartier en un autre (Le Falher et al. (2015)).

19. `http://livehoods.org/`

Pour comparer des quartiers, il est également possible de les regrouper, par exemple en utilisant des algorithmes de partitionnement ou de *clustering*. Les algorithmes de partitionnement comme KMeans, Affinity Propagation ou Spectral Clustering nécessitent de spécifier le nombre de groupes. Au contraire, les algorithmes de *clustering* comme DBSCAN et ses alternatives estiment automatiquement le nombre de groupes, mais requièrent un paramètre ϵ représentant la distance au-dessus de laquelle de nouveaux groupes sont créés. Les auteurs de Livehoods se basent sur l'algorithme *spectral clustering* pour comparer des quartiers (Cranshaw et al. (2012)).

Le *case-based reasoning* détecte des cas similaires pour rapprocher des quartiers. Par exemple, la situation d'une personne (e.g., composition du ménage, distance maison-travail) à la recherche d'une résidence est analysée pour proposer des quartiers où vivent des résidents dans une même situation (Yuan et al. (2013)).

Des algorithmes de classification peuvent être utilisés, en particulier avec des données issues des réseaux sociaux. Les travaux de Le Falher et al. (2015) utilisent *Information Theoretic Metric Learning* (ITML) et *Large Margin Nearest Neighbor* (LMNN) pour déterminer une matrice contenant les lieux les plus proches à partir des activités humaines réalisées dans ces lieux (vecteurs d'entrée) et des catégories de Foursquare (classes).

2.4 Évaluation

Pour certaines études, par exemple sociologiques, le processus informatique sert souvent d'outil dont le résultat (i.e., tableaux, visualisation, diagrammes) est ensuite analysé et interprété par des experts. Dans d'autres cas, le résultat produit par un processus informatique, en particulier la délimitation du quartier et la comparaison, a besoin d'être vérifié et validé, parfois de manière automatique.

Les enquêtes, déjà mentionnées en section 2.2, sont un moyen répandu pour l'évaluation, que ce soit sous forme de questionnaire ou d'entretien. Par exemple, pour mesurer l'impact d'un quartier sur l'activité physique, des enquêtes ont été envoyées à une centaine de résidents de deux quartiers de San Diego (Saelens et al. (2003)). Dans Lovejoy et al. (2010), les enquêtes permettent d'évaluer le degré de satisfaction des résidents vis à vis de leur quartier. Dans l'article de Yuan et al. (2012), des habitants de Beijing (au moins six ans de résidence) sont questionnés pour annoter des zones géographiques avec leur fonctions d'utilité. Cette expertise est ensuite comparée aux résultat de l'algorithme. Le web facilite aujourd'hui la création et le traitement d'enquêtes type questionnaire, et permet de toucher un plus grand nombre d'individus.

Pour une évaluation automatique, un résultat expertisé ou estimé est requis. Par exemple, pour évaluer la comparaison de quartiers, il est possible de disposer d'un jeu de données manuellement rempli qui liste les quartiers similaires. L'évaluation de la délimitation est particulièrement complexe du fait de la définition floue du quartier. Un cas d'utilisation plus précis peut permettre de trouver un mode d'évaluation adapté et fiable. L'évaluation automatique est aussi rendue possible par le micro-travail. En effet, on peut disposer de jeux de données annotés ou vérifiés par des personnes rémunérées grâce à des outils en ligne comme Amazon Mechanical Turk (Kittur et al. (2008)). Les villes développent parfois des planifications urbaines, et dont les informations peuvent être utiles pour l'évaluation. À Belo Horizonte, une méthodologie fournit des indicateurs sur la qualité de vie d'une dizaine de quartiers. Le nombre de services culturels, de santé, d'environnement, etc. est donc connu et mis à jour régulièrement.

Les travaux de Smarzaro et al. (2017) cherchent à vérifier si les fournisseurs cartographiques (Facebook, Foursquare, Google Places et Yelp) confirment ces statistiques expertisées. Dans les travaux de Yuan et al. (2012), l'algorithme d'annotation des zones géographiques est comparée à la planification urbaine de Beijing. Bien que les deux cartes se recoupent effectivement à certains endroits, une étude plus approfondie sur l'ensemble de la ville est nécessaire.

3 Conclusion et perspectives

Dans cet article, nous avons dressé un panorama des pistes de recherche informatique pour l'étude des quartiers en sciences humaines et sociales : description, délimitation, comparaison et évaluation. Ces grands axes sont utilisables dans de nombreux cas d'utilisation, que ce soit le regroupement ou la classification de quartiers, la recommandation, ou la prédiction de l'évolution d'un quartier.

Le projet HiL s'intéresse à plusieurs de ces pistes avec l'objectif de recommander le quartier idéal lors d'une recherche immobilière. Pour la description des quartiers, nous nous basons sur les données IRIS que nous agrégeons lorsque celles-ci renvoient à des découpages trop fins. La prise en compte du voisinage d'un IRIS permet d'étendre sa description en considérant les caractéristiques situées à proximité. Dans les grandes villes (où les IRIS sont souvent petits), le voisinage se rapproche d'une forme de quartier. En parallèle, nous étudions la possibilité de délimiter des quartiers en regroupant des IRIS adjacents partageant des descriptions similaires. Enfin, la comparaison de deux IRIS est essentielle pour la recommandation, et nous avons expérimenté quelques algorithmes (e.g., mesure cosine, DBSCAN) que nous devons affiner en exploitant les profils utilisateur. L'évaluation et la justification des recommandations sont des problèmes encore ouverts.

Références

Actif, N., A. Levet, S. Hoarau, H. Maillot, F. Andy, M. Boyer, C. Calteau, L. Trentin, et C. Ory (2013). Des quartiers inégaux face à la précarité. In *Cartographie sociale des territoires*. Insee.

Authier, J.-Y. (2008). Les citadins et leur quartier. *L'Année sociologique 58*(1), 21–46.

Authier, J.-Y., M.-H. Bacqué, et F. Guérin-Pace (2007). *Le quartier*. La Découverte.

Barret, N., F. Duchateau, F. Favetta, M. Miquel, A. Gentil, et L. Bonneval (2019). À la recherche du quartier idéal. In *EGC (à paraitre)*.

Bellahsène, Z., A. Bonifati, et E. Rahm (2011). *Schema matching and mapping*. Springer.

Berjawi, B., M. Colomb, T. Joliveau, F. Favetta, F. Duchateau, et M. Miquel (2013). Outil de repérage urbain à travers la prise de points de repère. Prototype, Laboratoires EVS et LIRIS.

Brindley, P., J. Goulding, et M. L. Wilson (2014). A data driven approach to mapping urban neighbourhoods. In *SIGSPATIAL*, pp. 437–440. ACM.

Chaix, B. (2014). Quartiers, mobilité et santé : l'étude record. *Les Cahiers de l'IAU*, 170–171.

Chen, C., B. Golshan, A. Y. Halevy, W.-C. Tan, et A. Doan (2018). BigGorilla : An Open-Source Ecosystem for Data Preparation and Integration. *IEEE Data Eng. Bull. 41*(2), 10–22.

Christen, P. (2012). *Data matching : concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Cranshaw, J., R. Schwartz, J. Hong, et N. Sadeh (2012). The livehoods project : Utilizing social media to understand the dynamics of a city.

Devernois, N., S. Muller, et G. Le Bihan (2014). *Gestion du patrimoine urbain et revitalisation des quartiers anciens : l'éclairage de l'expérience française*. Agence française de développement (AFD).

Du, S., F. Zhang, et X. Zhang (2015). Semantic classification of urban buildings combining vhr image and gis data : An improved random forest approach. *ISPRS journal of photogrammetry and remote sensing 105*, 107–119.

Guérois, M. et M. Madelin (2017). Comment les hôtes et clients d'Airbnb parlent-ils des lieux ? Une analyse exploratoire à partir du cas parisien. In *EXCES-EXtraction de Connaissances à partir de donnéEs Spatialisées*.

Gupta, S., P. Szekely, C. A. Knoblock, A. Goel, M. Taheriyan, et M. Muslea (2012). Karma : A system for mapping structured sources into the semantic web. In *Extended Semantic Web Conference*, pp. 430–434. Springer.

Halevy, A., A. Rajaraman, et J. Ordille (2006). Data integration : the teenage years. In *VLDB '06 : Proceedings of the 32nd international conference on Very large data bases*, pp. 9–16. VLDB Endowment.

Huang, X., H. Liu, et L. Zhang (2015). Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing 53*(7), 3639–3657.

Kang, J., M. Körner, Y. Wang, H. Taubenböck, et X. X. Zhu (2018). Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Kennedy, L. S. et M. Naaman (2008). Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th international conference on World Wide Web*, pp. 297–306. ACM.

Kittur, A., E. H. Chi, et B. Suh (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453–456. ACM.

Kumar, C., W. Heuten, et S. Boll (2015). Visual overlay on openstreetmap data to support spatial exploration of urban environments. *ISPRS International Journal of Geo-Information 4*(1), 87–104.

Le Falher, G., A. Gionis, et M. Mathioudakis (2015). Where Is the Soho of Rome ? Measures and Algorithms for Finding Similar Neighborhoods in Cities. *ICWSM 2*, 3–2.

Leidner, J. L. et M. D. Lieberman (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special 3*(2), 5–11.

Lovejoy, K., S. Handy, et P. Mokhtarian (2010). Neighborhood satisfaction in suburban versus traditional environments : An evaluation of contributing characteristics in eight California neighborhoods. *Landscape and Urban Planning 97*(1), 37–48.

Maurin, E. (2004). Le ghetto français. *Enquête sur le séparatisme social*.

Miller, H. J. et J. Han (2009). *Geographic data mining and knowledge discovery*. CRC Press.

Morland, K., S. Wing, A. D. Roux, et C. Poole (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine 22*(1), 23–29.

Pan Ké Shon, J.-L. (2005). La représentation des habitants de leur quartier : entre bien-être et repli. *Économie et statistique 386*(1), 3–35.

Pardoen, M. (2015). Les oreilles à l'affût ! restitution d'un paysage sonore : œuvre de l'imaginaire ou recherche d'authenticité. *Silences et bruits du Moyen Âge à nos jours : Perceptions, identités sonores et patrimonialisation. L'Harmattan*.

Perez, N., A. Mailhac, C. Inard, et P. Riederer (2016). Outil d'aide à la décision multicritère pour la conception de systèmes énergétiques à l'échelle du quartier. In *IBPSA France Conference*.

Préteceille, E. (2009). La ségrégation ethno-raciale a-t-elle augmenté dans la métropole parisienne ? *Revue française de sociologie 50*(3), 489–519.

Saelens, B. E., J. F. Sallis, J. B. Black, et D. Chen (2003). Neighborhood-based differences in physical activity : an environment scale evaluation. *American journal of public health 93*(9), 1552–1558.

Shen, W., J. Wang, et J. Han (2015). Entity linking with a knowledge base : Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on 27*(2), 443–460.

Smarzaro, R., T. F. d. M. Lima, et C. A. Davis Jr (2017). Could data from location-based social networks be used to support urban planning ? In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1463–1468. International World Wide Web Conferences Steering Committee.

Strötgen, J., M. Gertz, et P. Popov (2010). Extraction and exploration of spatio-temporal information in documents. In *Workshop on Geographic Information Retrieval*, pp. 16. ACM.

Tang, E. et K. Sangani (2015). Neighborhood and price prediction for san francisco airbnb listings.

Yu, M., G. Li, D. Deng, et J. Feng (2016). String similarity search and join : a survey. *Frontiers of Computer Science 10*(3), 399–417.

Yuan, J., Y. Zheng, et X. Xie (2012). Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194. ACM.

Yuan, X., J.-H. Lee, S.-J. Kim, et Y.-H. Kim (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems 38*(2), 231 – 243.

Zhang, A. X., A. Noulas, S. Scellato, et C. Mascolo (2013). Hoodsquare : Modeling and recommending neighborhoods in location-based social networks. In *Social Computing*, pp. 69–74. IEEE.

Zhang, D., M. M. Islam, et G. Lu (2012). A review on automatic image annotation techniques. *Pattern Recognition 45*(1), 346–362.

Summary

The French project Home In Love (HiL) aims at recommending real estates (for buying or renting), especially when people move in a new city. With the growing amount of websites (including pictures, virtual tours), one can easily search for and select an ideal property. However, summarizing information about the neighborhood(s) remains an issue. In this paper, we identify and describe the main challenges related to the study of neighborhoods (description, border detection, comparison and evaluation).