HAL open science

Une méthode pour l'estimation désagrégée de données de population à l'aide de données ouvertes

Luciano Gervasoni, Serge Fenet, Peter Sturm

To cite this version:

Luciano Gervasoni, Serge Fenet, Peter Sturm. Une méthode pour l'estimation désagrégée de données de population à l'aide de données ouvertes. EGC 2018 - 18ème Conférence Internationale sur l'Extraction et la Gestion des Connaissances, Jan 2018, Paris, France. pp.83-94. hal-01667975

HAL Id: hal-01667975

https://inria.hal.science/hal-01667975v1

Submitted on 19 Dec 2017

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.

Une méthode pour l'estimation désagrégée de données de population à l'aide de données ouvertes

Luciano Gervasoni*, Serge Fenet**, Peter Sturm*

*Inria, Univ. Grenoble Alpes, LJK, F-38000, Grenoble, France
luciano.gervasoni@inria.fr, peter.sturm@inria.fr
**Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
serge.fenet@liris.cnrs.fr

Résumé. Nous présentons dans ce travail une méthode de désagrégation pour l'estimation de population à l'échelle locale à partir de données ouvertes globales. Notre but est d'estimer notamment le nombre de personnes résidant dans chaque bâtiment de la zone d'intérêt, à partir de données à plus grande échelle. Une description fine à l'échelle résidentielle est tout d'abord effectuée à partir des données d'OpenStreetMap. Les surfaces des bâtiments d'habitation ou d'usage mixte (habitation et activités) sont notamment identifiées. Nous effectuons ensuite une désagrégation à partir de données de grille de population à grande échelle (1km^2 par carreau), guidée par les surfaces des bâtiments compris dans chaque carreau de la grille. Ensuite, nous effectuons une désagrégation à partir de données de grille de population à grande échelle (1km^2 par carreau), guidée par les distributions spatiales découvertes à l'étape précédente. Nous utilisons exclusivement des données ouvertes pour favoriser la réplicabilité et pour pouvoir appliquer notre méthode à toute région d'intérêt, pour peu que la qualité des données soit suffisante. L'évaluation et la validation du résultat dans le cas de plusieurs villes Françaises sont effectuées à l'aide de données de recensement INSEE.

1 Introduction

De nombreuses recherches portant sur le contexte urbain, telles que l'étude du développement du commerce de proximité, les études démographiques ou la planification urbaine, nécessitent des données de population à une échelle assez fine, parfois jusqu'au niveau du bâti (Lu et al. (2010); Ural et al. (2011); Langford (2013); Sridharan et Qiu (2013)). Les modèles de transport par exemple, qui sont de plus en plus au cœur des questionnements urbains, nécessitent généralement des données décrivant les populations des habitations et de lieux de travail. En effet, la majorité du transport urbain est liée aux déplacements entre habitation et travail, et la précision des modèles de transport les décrivant est fortement corrélée à la connaissance des populations à l'échelle des bâtiments. Comme ces données ne sont que rarement disponibles, il est important de pouvoir efficacement les inférer à partir de données à plus grande échelle.

Les données de recensement contiennent des informations précieuses pour l'estimation d'une cartographie à haute résolution des population des bâtiments résidentiels et d'activités. Toutefois, la faible résolution temporelle -due à la fréquence de la collecte des données- et la faible résolution spatiale sont deux désavantages reconnus de cette source de données (Lu et al. (2010)). En conséquence, ce données sont souvent disponibles uniquement sous forme de projection en grille composées de carreaux très larges. Ces derniers sont trop grands pour permettre une estimation directe des populations à l'échelle des bâtiments, et nécessitent donc une étape de désagrégation à l'aide me méthodes diverses (Bakillah et al. (2014)). Sachant que cette étape a pour rôle de prendre en compte la topologie locale des bâtiments, ainsi que leur usage réel, elle nécessite donc des donnée additionnelles absentes des données de recensement.

Une première solution pour obtenir ces données de topologie urbaine repose sur l'usage du LiDAR (Light Detection And Ranging), mais cette solution demande des campagnes de mesure dédiées, reste extrêmement coûteuse, et comporte de nombreux biais (Harvey (2002)). Une seconde méthode se repose sur l'utilisation de campagnes de questionnaires, souvent utilisés pour la construction de modèles de transport. Mais là encore, un biais important est lié au très faible pourcentage de la population interrogée. Une autre source de données, les données d'usage et de couverture du sol (Land Use Land Cover) ont déjà été utilisées pour effectuer une estimation de la population par bâtiment (Mennis (2003); Reibel et Agrawal (2007)), mais elle ne peuvent être utilisées que pour des régions d'intérêt très précises. Elles peuvent par ailleurs être en partie inférées à partir des images satellites, plus faciles à obtenir. Mais si cette méthode est directement applicable au problème de l'inférence d'usage des sols, elle est plus difficilement applicable à l'obtention de donnée d'usage et de couverture du sol : l'intensité d'usage, qui est essentielle pour calculer des estimations de lieux d'emplois et de résidence réalistes, n'est pas directement extractible (Rodrigues et al. (2013)). Il a de plus été montré que les techniques basées sur l'utilisation de données satellitaires voient leur qualité baisser lorsque la densité de population augmente (Danoedoro (2006)). Ainsi, la reconstruction de la topologie des bâtiments reste très problématique sans utiliser des sources de données complémentaires. C'est ainsi le cas dans (Ural et al. (2011)) où il est noté que "l'empreinte et la hauteur des bâtiments sont d'abord déterminés à partir d'images aériennes, de modèle numériques de terrains, et de modèles de surface". Cependant la disponibilité de ce type de données et donc l'applicabilité de cette méthode sont très limitées.

Dans Liu et al. (2008), les auteurs estiment la densité de population à partir d'images satellitaires et de données de recensement. La relation entre la population et la topologie des bâtiments a aussi été étudiée dans le contexte de données LiDAR par Lu et al. (2010), où les auteurs ont montré que l'estimation de population était meilleure en utilisant la surface plutôt que le volume. Cela s'explique par l'homogénéité de l'habitat individuel, ainsi que par les erreurs de classification potentielles liées aux données volumétriques LiDAR. Dans ces travaux, les auteurs utilisent donc des données d'imagerie pour estimer la population locale à petit échelle. Ils ne descendent cependant pas à l'échelle de l'habitation individuelle. Leur méthode repose sur (1) l'identification du nombre d'unités d'habitation, (2) l'extraction des surfaces artificielles liées aux surfaces résidentielles, (3) la classification des types d'usage des sols, et enfin (4) l'extrapolation

des valeurs de population à partir des données spectrales de réflectance.

Une revue détaillée des méthodes d'estimation de population dans le contexte des Systèmes d'Information Géographique (SIG) à l'aide de données d'imagerie est disponible dans Wu et al. (2005).

La saisie volontaire de données géographiques (Voluntary Geographical Information (VGI)) est une importante source de données. Ce modèle, initialement critiqué pour des arguments de qualité, a vu un nombre croissant de contributions, une augmentation constante de qualité, et une très forte réactivité faire d'une plate-forme comme OpenStreetMap (OSM) un acteur incontournable du domaine. Les données VGI en source ouverte sont faciles d'accès et d'exploitation. Dans le cas particulier d'OSM, la plate-forme est maintenant connue pour la précision, la qualité, et la complétude de ses données, trois caractéristiques qu'on ne retrouve pas simultanément dans les jeux de données commerciaux et privés.

De nombreuses évaluations de la qualité de la base OSM ont été effectuées. Les travaux de Barrington-Leigh et Millard-Ball (2017) montrent que "dans de nombreuses applications, les chercheurs et les décideurs publiques peuvent compter sur la complétude d'OSM, ou pourront le faire très bientôt". Fan et al. (2014) évaluent les données de Munich et concluent que dans de nombreuses villes Européennes, les données d'OSM ont "une très forte complétude et précision sémantique". Par ailleurs, Touya et al. (2017) montrent que "les variations de contenu dans OSM sont parfois guidées par des désaccords entre les contributeurs, mais au final elles augmentent la qualité des données". Ainsi, la qualité supérieure des données d'OSM a été montrée pour le jeu de données officiel Meridian 2 pour la Grande-Bretagne (Haklay (2010)), et il a été montré que le réseau urbain Allemand est plus complet que celui décrit dans les jeux de données commerciaux (Neis et al. (2011)). Par exemple, celui de Hambourg est complet à 99.8% (Over et al. (2010)). Pour montrer la vivacité de cette communauté, on peut remarquer que le volume de points d'intérêt (Points Of Interest (POI)) en Chine a été multiplié d'un facteur 9 entre 2007 et 2013 (Liu et Long (2016)).

Avant que la qualité des données d'OSM n'atteigne un tel niveau, des données relatives à l'usage des sols à la fois précises et à jour étaient difficiles à trouver, notamment aux échelles continentale et régionale. Heureusement, ce n'est maintenant plus le cas.

2 Méthodologie

Contribution : Nous proposons dans ce travail une méthode basée sur l'évaluation des surfaces d'habitation pour estimer la population des bâtiments en utilisant uniquement des données libres. Cette méthode vise une applicabilité à n'importe quelle région d'intérêt, pour peu que les données OSM associées possèdent une qualité et une complétude suffisante.

Méthode générale : Tout d'abord, la liste des bâtiments de la zone d'intérêt est extraite de la base OSM. Cette liste est ensuite filtrée selon l'usage déclaré : bâtiment résidentiel ou bâtiment d'usage mixte. Ensuite, la surface d'habitation de chacun est évaluée à partir des annotations des "building parts" composant chaque bâtiment. Enfin, des données annexes de population disponibles sous forme de grille sont utilisées pour

calculer le nombre d'habitants de chaque bâtiment, en prenant en compte les ratios des surfaces de ses différents usages. Dans ce travail, nous ré-agrégeons ensuite ces données sous forme de grille, afin de pouvoir comparer la qualité du résultat avec une vérité terrain, mais cette dernière étape n'est pas indispensable pour une application réelle.

2.1 Requête OSM

Les données OSM portant sur la région d'intérêt sont obtenues par l'intermédiaire de l'API Overpass, suivant la méthode de Boeing (2017). La région d'intérêt peut être définie de plusieurs manières :

- Par l'intermédiaire d'une forme polygonale géo-référencée.
- En utilisant une référence administrative définie dans OSM (nom de ville, limite administrative, etc.).
- À l'aide de coordonnées géographiques et d'une distance définissant un cercle.
- En définissant un cercle par une adresse postale de centre et une distance.
- En utilisant la liste des coordonnées géographiques des coins d'une boite englobante.

La requête auprès de la base OSM renvoie les données suivantes :

- Les bâtiments de la zone d'intérêt sélectionnée (Buildings OSM), en utilisant la même méthode que Boeing (2017).
- Les sous-parties de bâtiments (Building parts OSM), afin de reconstruire la surface réelle associée à chaque bâtiment.
- Les polygones d'usage des sols (Land use polygons OSM), afin d'effectuer l'inférence de l'usage si les données annotées dans la base sont insuffisantes.
- Les points d'intérêt (POIs OSM) associés aux usages des sols d'activité. Ils permettent d'identifier avec plus de précision le cas de bâtiments ayant un usage mixte d'habitation et d'activité.

Certaines sous-parties de bâtiments sont filtrées si elles se révèlent inutiles pour le traitement ("*building : part*"="*no*" et "*building : part*"="*roof*"), ou bien si le bâtiment a déjà été identifié. Ce dernier cas se pose par exemple lorsqu'à la fois le bâtiment et ses sous-parties sont annotées en tant que bâtiment.

Toutes les données géographiques sont projetées dans le système de coordonnées UTM (Universal Transverse Mercator) pour réduire les biais en fournissant des distances correctes indépendamment de la zone d'intérêt choisie.

2.2 Classification

Les structures géographiques importées à partir de la base OSM (bâtiments, partie de bâtiments, polygones et points d'intérêts) sont classifiées en fonction de leur usage des sols : "activité", "résidentiel", "mixte" ou "autre". Cela permet de reconstituer les structures des bâtiments ayant un usage résidentiel complet ou partiel.

Informations d'étiquettes : Les bâtiments, parties de bâtiments et points d'intérêt sont classés en fonction de leur étiquette. Nous suivons pour cela le processus décrit dans le

		Bâtiment		
		Activité	Résidentiel	Usage mixte
POI	Activité	Activité	Mixte	Mixte
	Résidentiel	Mixte	Résidentiel	Mixte
	Usage mixte	Mixte	Mixte	Mixte

Tab. 1: Adaptation de la classe d'usage des bâtiments en fonction des points d'intérêt qu'ils contiennent.

wiki d'OSM [1]. Nous effectuons ensuite une étape d'inférence d'usage de bâtiment pour prendre en compte les données manquantes et gérer les étiquetages incomplets. Cette procédure est décrite en détail dans Gervasoni et al. (2016).

Points d'intérêt : La base OSM utilise les points d'intérêts pour étiqueter des caractéristiques particulières géo-localisées. Étant donné qu'ils sont essentiellement associés à un usage des sols d'activité, il sont donc utiles pour déterminer l'usage réel de nombreux bâtiments. Ainsi, dans notre méthode le bâtiment contenant un POI se verra assigner un nouvelle classe telle que défini dans la table 1. Cette procédure est motivée par le fait que de très nombreux bâtiments à usage mixte contiennent leurs activités sous forme de POI.

2.3 Calcul de la surface résidentielle

Niveaux et étages des bâtiments : Si le nombre d'étage d'un bâtiment est renseigné dans OSM, on peut l'utiliser directement. Ce n'est hélas pas souvent le cas, et l'on utilise alors les données relatives aux hauteurs des bâtiments et de leurs différentes parties. Pour approcher la relation entre le nombre d'étages et la hauteur, nous utilisons les recommandations d'un outil récent développé pour la visualisation 3D des empreintes de bâtiments à partir de données OSM qui propose une hauteur de $3m$ en moyenne par étage [2]. Le nombre d'étages est donc calculé pour chaque bâtiment en suivant cette méthode ainsi que les recommandations d'OSM [3]. Lorsque les données manquantes sont trop nombreuses, une hauteur par défaut de $3m$ ou d'un étage est supposée.

Calcul de la surface d'usage résidentiel : La surface résidentielle d'un bâtiment est calculée en fonction de sa classification et de celle de ses sous-parties. Dans le cas de bâtiments d'usage mixte sans informations plus détaillées sur leurs sous-parties, nous supposons que la moitié de sa surface totale est utilisée pour un usage résidentiel.

Afin de calculer la surface totale de résidence, nous multiplions le nombre d'étages par la surface associée à chaque bâtiment ou sous-partie de bâtiment. La surface associée à chaque usage des sols est calculée incrémentalement en considérant chaque composant de chaque bâtiment. Cela nous permet de prendre en compte le fait que

1. http://wiki.openstreetmap.org/wiki/Map_Features
2. https://www.mapbox.com/blog/mapping-3d-buildings/
3. http://wiki.openstreetmap.org/wiki/Key:building:levels

différentes sous-parties peuvent être associées à différents usages, chacune contribuant différemment aux surfaces associées.

3 Estimation désagrégée de population

La procédure de désagrégation utilise l'information de surface totale d'habitation de chaque bâtiment, ainsi qu'un jeu de données de population sous forme de grille géo-référencée. En l'absence de toute donnée socio-économique géographique, nous faisons l'hypothèse d'une consommation constante de surface par habitant pour tous les bâtiments. Cela introduit un biais qui pourra être levé dans le futur en prenant en compte notamment le type d'habitation et les catégories socio-professionnelles.

Nous utilisons dans ces travaux des données de population libres disponibles sous forme de grille géo-référencée. Un premier jeu de données représente la population mondiale avec une résolution de $1km^2$(Doxsey-Whitfield et al. (2015)). Dans le cas particulier de la France, les données de population INSEE sont utilisables, avec une résolution spatiale de $200m^2$.

Calcul de la population par bâtiment : Nous répartissons cette population proportionnellement en fonction de la surface résidentielle totale calculée dans chaque carré G_i de la grille échantillonnant la population. Les bâtiments à cheval sur plusieurs carrés de cette grille voient leur surface totale affectée à chaque carré en fonction du pourcentage d'inclusion de leur empreinte.

Si ϕ est l'opération d'obtention de la surface d'un polygone, et si $R(x)$ est la fonction qui détermine la quantité totale de surface d'habitation du bâtiment x, alors la surface résidentielle totale du bâtiment B_j se situant dans le carré G_i est déterminée par :

$$R(B_j, G_i) = R(B_j) * (\ \phi(B_{j,Polygon}\ \cap\ G_{i,Polygon}\ )\ /\ \phi(B_{j,Polygon})\ ) \tag{1}$$

Ainsi, la somme dees surface résidentielles définit la surface totale incluse dans G_i :

$$\sum_{B_j} R(B_j, G_i) \tag{2}$$

Enfin, la population de chaque carré de la grille de population est répartie au sein de chaque bâtiment appartenant à ce carré selon pourcentage de leur surface résidentielle.

4 Résultats

Nous avons évalué notre méthode de désagrégation de données de population sur plusieurs villes. Pour ce faire, nous avons utilisé des données de population sur une grille d'une résolution de $1km^2$ pour Manhattan, New York [4], ainsi que pour des villes Françaises avec une résolution plus fine de $200m \times 200m$ [5].

4. http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
5. https://www.insee.fr/fr/statistiques/2520034

Aucune des sources de données n'est exempte d'erreurs, ainsi les estimations de la population ne seront pas nécessairement cohérentes entre elles.

Comme la vérité terrain (population par bâtiment) n'est pas disponible, nous évaluons notre méthode en calculant des données de population à une résolution plus grossière à partir des données connues à une résolution plus fine, en agrégeant 5×5 carrés. Puis, nous appliquons notre méthode de désagrégation en partant sur ces données à résolution grossière, ce qui nous permet de comparer le résultat à une vérité terrain et de produire des histogrammes d'erreurs.

Notons que dans le cas de données manquantes (carrés sans données sur la population), l'agrégation est simplement effectuée sur les carrés pour lesquels la population est connue.

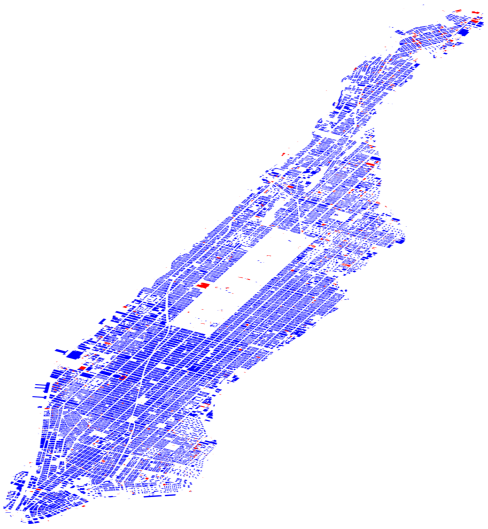


Fig. 1: Carte de hauteurs pour Manhattan, New York. Les bâtiments avec une information associée sur la hauteurs ou le nombre d'étages, sont montrés en bleu, les autres en rouge.

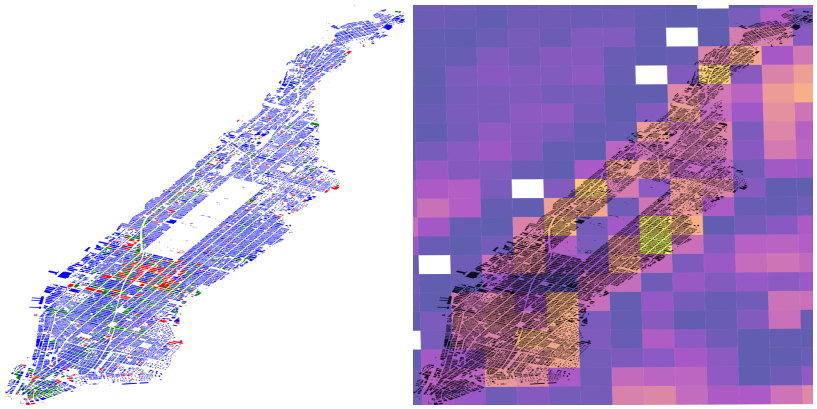
La procédure d'agrégation est effectuée de manière itérative afin d'éviter une superposition des carrés de la grille. Pour chaque carré de la grille à $200m \times 200m$, nous indexons les carrés voisins afin de créer un carré simulé de cotés $1km \times 1km$, centré dans le carré initial. La procédure de désagrégation est ensuite appliquée en utilisant l'union des carrés comme géométrie et la somme de leur population comme entrée.

Les estimations de la population par bâtiment sont agrégées selon la résolution de la vérité terrain. Pour des bâtiments à cheval sur plusieurs carrés, la répartition de la population est faite de manière proportionnelle comme dans l'équation 1.

Remarquons que l'erreur croît aux bords de la région analysée pour une raison simple : des polygones larges pourraient ne pas être retrouvés aux bords de la région puisque seuls des polygones complètement contenus dans la boîte englobante sont trouvés. Ainsi, la procédure d'inférence de l'usage du sol peut donner lieu à une mauvaise classification, due à ces données manquantes.

4.1 Manhattan, New York

La disponibilité de données sur la hauteur ou le nombre d'étages des bâtiments est variable, mais un exemple remarquable est montré sur la figure 1. Elle montre clairement que l'exploitation de l'information sur le nombre des étages est incontournable pour des lieux tels Manhattan. Leur utilisation permet une évaluation plus précise de la surface associée à l'usage résidentiel de chaque bâtiment et améliore ainsi grandement l'estimation de la population par bâtiment.



(a) (b)

Fig. 2: Manhattan, New York. (a) Bâtiments classifiés selon leur usage de sol : résidentiel, activité et mixte (bleu, rouge et vert respectivement). (b) Grille grossière de population en surimposition aux bâtiments. La couleur jaune indique la densité de population la plus élevée.

La figure 2 montre tous les bâtiments retrouvés dans la base OSM, différenciés selon leur usage de sol classifié. La partie droite de la figure montre la surimposition de la grille grossière de population.

4.2 France

Plusieurs villes Françaises ont été traitées, allant de grandes villes telles Paris et Lyon, via des villes moyennes telle Toulouse, à des villes de la taille de Grenoble. La figure 4 montre tous les bâtiments récupérés de la base OSM, différenciés selon leur usage de sol classifié. La partie droite de la figure montre la surimposition de la grille fine de population.

Le tableau 2 montre les médianes des erreurs relatives et absolues, obtenues pour la méthode de désagrégation. Les erreurs relatives sont similaires à travers les différentes villes, même lorsque les densités de population sont différentes. La figure 3 montre les histogrammes des erreurs absolues pour chacune des villes. On peut observer que Paris a la médiane des erreurs absolues la plus élevée même si l'erreur relative médiane y est la plus faible ; ceci est dû à sa densité de population, plus élevée que pour les autres villes testées.

5 Conclusions et perspectives

Nous avons présenté une méthode pour l'estimation désagrégée de la population d'une ville, au niveau des bâtiments individuels. Elle n'utilise que des données ouvertes : la surface dédiée à l'usage résidentiel d'un bâtiment est estimée à partir de données

	Erreur relative médiane	Erreur absolue médiane
Grenoble	42.1%	12.08
Lyon	47.3%	36.92
Paris	41.7%	215.17
Toulouse	47.3%	35.78

Tab. 2: Médianes des erreurs relatives et absolues.

OSM, et des données sur la population disponibles sur une grille sont utilisées afin de produire une estimation de la population par bâtiment.

Une validation est proposée qui procède en simulant les données de population à une résolution disponible au monde entier à partir de données plus fine disponible en France, puis en y appliquant la méthode de désagrégation et en comparant le résultat aux données connues. Les médianes des erreurs absolues et relatives sont calculées pour quatre villes Françaises.

Nos prochains travaux visent une amélioration de la méthode par l'exploitation de caractéristiques urbaines disponibles dans OSM, ce qui permettra de lever l'hypothèse d'une consommation constante de la surface habitable par habitant.

Références

Bakillah, M., S. Liang, A. Mobasheri, J. Jokar Arsanjani, et A. Zipf (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. International Journal of Geographical Information Science 28(9), 1940–1963.

Barrington-Leigh, C. et A. Millard-Ball (2017). The world's user-generated road map is more than 80% complete. PloS ONE 12(8), e0180698.

Boeing, G. (2017). OSMnx : New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems 65(Supplement C), 126 – 139.

Danoedoro, P. (2006). Extracting land-use information related to socio-economic function from quickbird imagery : A case study of Semarang area, Indonesia. Map Asia.

Doxsey-Whitfield, E., K. MacManus, S. B. Adamo, L. Pistolesi, J. Squires, O. Borkovska, et S. R. Baptista (2015). Taking advantage of the improved availability of census data : a first look at the gridded population of the world, version 4. Papers in Applied Geography 1(3), 226–234.

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.

Gervasoni, L., M. Bosch, S. Fenet, et P. Sturm (2016). A framework for evaluating urban land use mix from crowd-sourcing data. In IEEE International Conference on Big Data, pp. 2147–2156. IEEE.

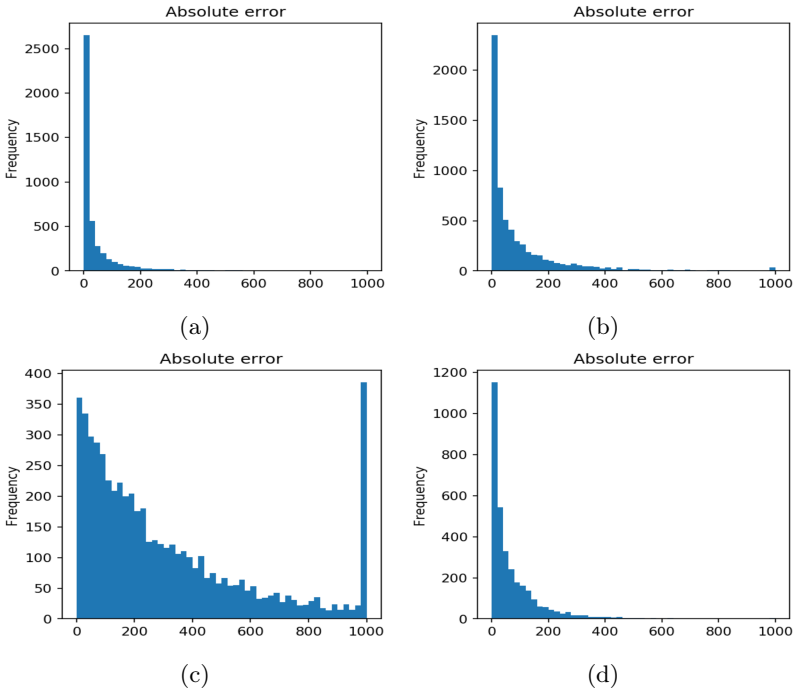


Fig. 3: Histogrammes des erreurs absolues pour (a) Grenoble (b) Lyon (c) Paris et (d) Toulouse.

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.

Harvey, J. (2002). Estimating census district populations from satellite imagery : some approaches and limitations. International Journal of Remote Sensing 23(10), 2071–2095.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. Geographical Analysis 45(3), 324–344.

Liu, X., P. C. Kyriakidis, et M. F. Goodchild (2008). Population-density estimation using regression and area-to-point residual kriging. International Journal of Geographical Information Science 22(4), 431–447.

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.

Lu, Z., J. Im, L. Quackenbush, et K. Halligan (2010). Population estimation based on multi-sensor data fusion. International Journal of Remote Sensing 31(21), 5587–5604.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. The Professional Geographer 55(1), 31–42.

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.

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.

Reibel, M. et A. Agrawal (2007). Areal interpolation of population counts using pre-classified land cover data. Population Research and Policy Review 26(5-6), 619–633.

Rodrigues, F., A. Alves, E. Polisciuc, S. Jiang, J. Ferreira, et F. Pereira (2013). Estimating disaggregated employment size from points-of-interest and census data : From mining the web to model implementation and visualization. International Journal on Advances in Intelligent Systems 6(1), 41–52.

Sridharan, H. et F. Qiu (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. Geographical Analysis 45(3), 238–258.

Touya, G., V. Antoniou, A.-M. Olteanu-Raimond, et M.-D. Van Damme (2017). Assessing crowdsourced POI quality : Combining methods based on reference data, history, and spatial relations. ISPRS International Journal of Geo-Information 6(3), 80.

Ural, S., E. Hussain, et J. Shan (2011). Building population mapping with aerial imagery and GIS data. International Journal of Applied Earth Observation and Geoinformation 13(6), 841–852.

Wu, S.-S., X. Qiu, et L. Wang (2005). Population estimation methods in GIS and remote sensing : a review. GIScience & Remote Sensing 42(1), 80–96.

Summary

In this article we present a method to perform dissagregated population estimation at building level using open data. Our goal is to estimate the number of people living at the fine level of individual households by using open urban data and coarse-scaled population data. First, a fine scale description of residential land use per building is built using OpenStreetMap. Then, using coarse-scale gridded population data, we perform the down-scaling for each household given their containing area for residential usage. We rely solely on open data in order to ensure replicability, and to be able to apply our method to any city in the world, as long as sufficient data exists. The evaluation is carried out using fine-grained census block data for cities in France as ground-truth.

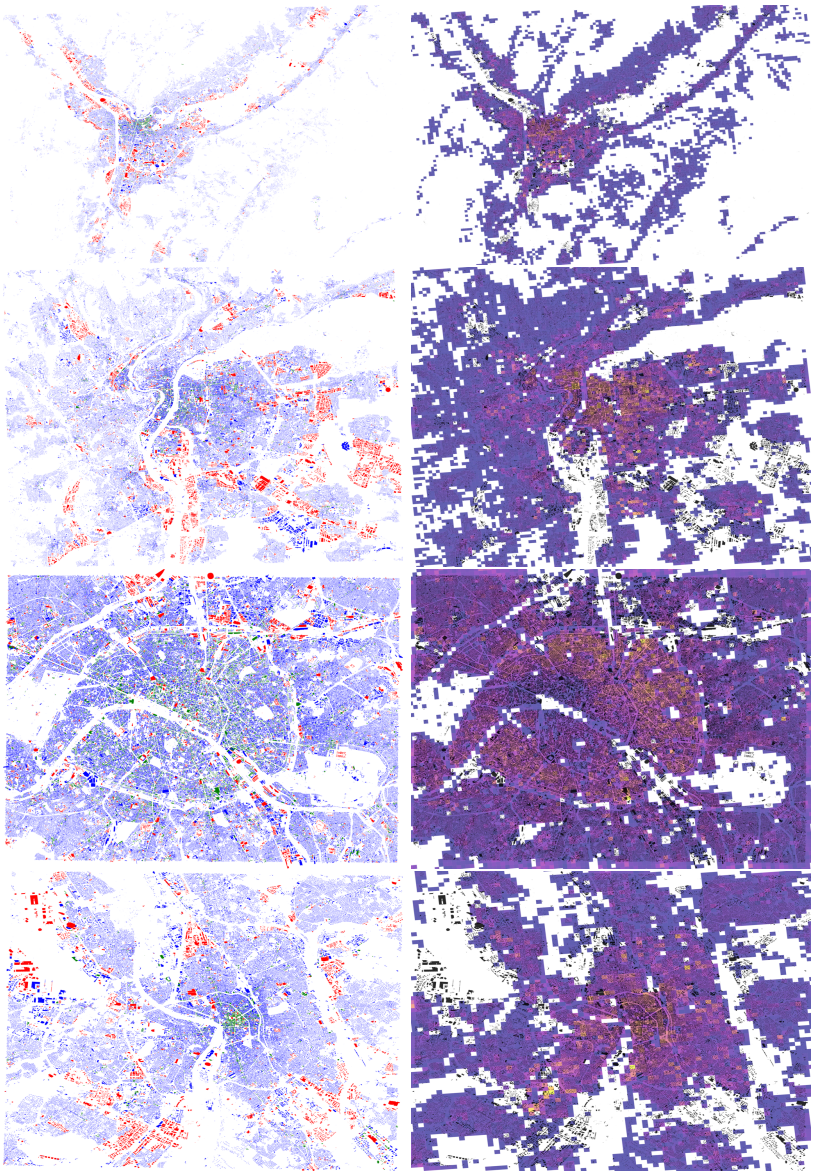


Fig. 4: Lignes : villes de Grenoble, Lyon, Paris et Toulouse. Colonnes : à gauche, les bâtiments et l'usage de sol associé (résidentiel, activité et mixte, en bleu, rouge et vert respectivement). A droite, surimposition de la grille fine de données de population. La couleur jaune indique une densité de population élevée.