

# LE VANDALISME DE L'INFORMATION GÉOGRAPHIQUE VOLONTAIRE<sup>1</sup>

par Quy Thy Truong

Institut national de l'information géographique et forestière

73 avenue de Paris 94160 Saint-Mandé

pre.nom@ign.fr

Depuis ces dernières années, le développement des technologies de l'information et de la communication a eu un fort impact sur la production et le partage des données numériques. En effet, celui-ci a notamment entraîné l'émergence des sciences citoyennes et du crowdsourcing, qui exploitent la capacité de la foule à produire des données en quantité massive pour répondre à des problématiques sociétales, environnementales ou scientifiques.

Dans le domaine des sciences de l'information géographique, l'information géographique volontaire désigne tout type de données spatiales collectées par tout type de contributeurs, qu'ils soient professionnels ou pas. Par ailleurs, les plateformes de saisie d'information géographique volontaire connaissent une certaine popularité : par exemple, le projet OpenStreetMap (OSM) compte chaque mois plus de 40 000 contributeurs pour mettre à jour la base cartographique du monde entier.

Toutefois, la qualité de l'information géographique est discutable. En effet, celle-ci peut être inexacte et sujette à des erreurs, voire même à des dégradations volontairement commises par ses contributeurs. Dans ce dernier cas, nous pouvons parler de vandalisme cartographique, ou de carto-vandalisme. Le risque de carto-vandalisme constitue l'un des inconvénients majeurs du crowdsourcing, et malgré le faible nombre d'incidents de carto-vandalisme détectés à notre connaissance, ce phénomène peut dissuader de l'utilisation des données provenant de ce mode de collecte.

Jusqu'à récemment, les instituts nationaux de cartographie avaient le monopole de la production de données géographiques. Etant donnée l'émergence de l'information géographique volontaire, certaines agences de cartographie ont commencé à s'intéresser aux approches de collecte collaborative pour acquérir les données spatiales. Par exemple, en France,

l'Institut National de l'Information Géographique et Forestière a mis en place des projets collaboratifs auxquels peuvent participer des contributeurs non-professionnels pour collecter des données (IGN Rando, LandSense, Espace Collaboratif). Ces instituts ayant pour vocation de fournir l'information géographique officielle, ils se doivent de produire des données de qualité optimale. En conséquence, l'enjeu est d'exploiter le potentiel offert par l'information géographique volontaire – par exemple, pour mettre à jour des bases de données cartographiques officielles – en évitant l'intégration de données de piètre qualité.

Dans ce contexte, cette thèse de doctorat se focalise sur l'évaluation de la qualité de l'information géographique volontaire. En particulier, la recherche porte sur la détection des contributions issues de carto-vandalisme, car celle-ci peut offrir une garantie minimale à l'utilisation des données spatiales collaboratives. Les objectifs visés dans cette thèse sont : tout d'abord de formuler clairement une définition du carto-vandalisme ; puis de qualifier les contributeurs d'information géographique volontaire dans le but d'identifier ceux qui, intentionnellement, détériorent les cartes collaboratives ; enfin, de trouver des méthodes appropriées pour détecter les données de carto-vandalisme.

## Définition du carto-vandalisme

La première partie de la thèse a consisté à définir le terme de « carto-vandalisme ». Pour cela, il a fallu remonter aux origines historiques du vandalisme tel qu'il est communément compris, et d'analyser comment il est défini selon la législation française. Progressivement, en appliquant cette définition au domaine numérique puis cartographique, nous avons pu définir théoriquement le carto-vandalisme comme une dégradation de l'espace cartographique collaboratif. De manière plus pratique, nous avons constitué un échantillon d'incidents réels relevés

<sup>1</sup> Thérèse Quy Thy Truong, Le vandalisme de l'information géographique volontaire : analyse exploratoire et proposition d'une méthodologie de détection automatique, Thèse dirigée par Guillaume Touya et Cyril de Runz, soutenue à l'Université Paris-Est le 8 janvier 2020.

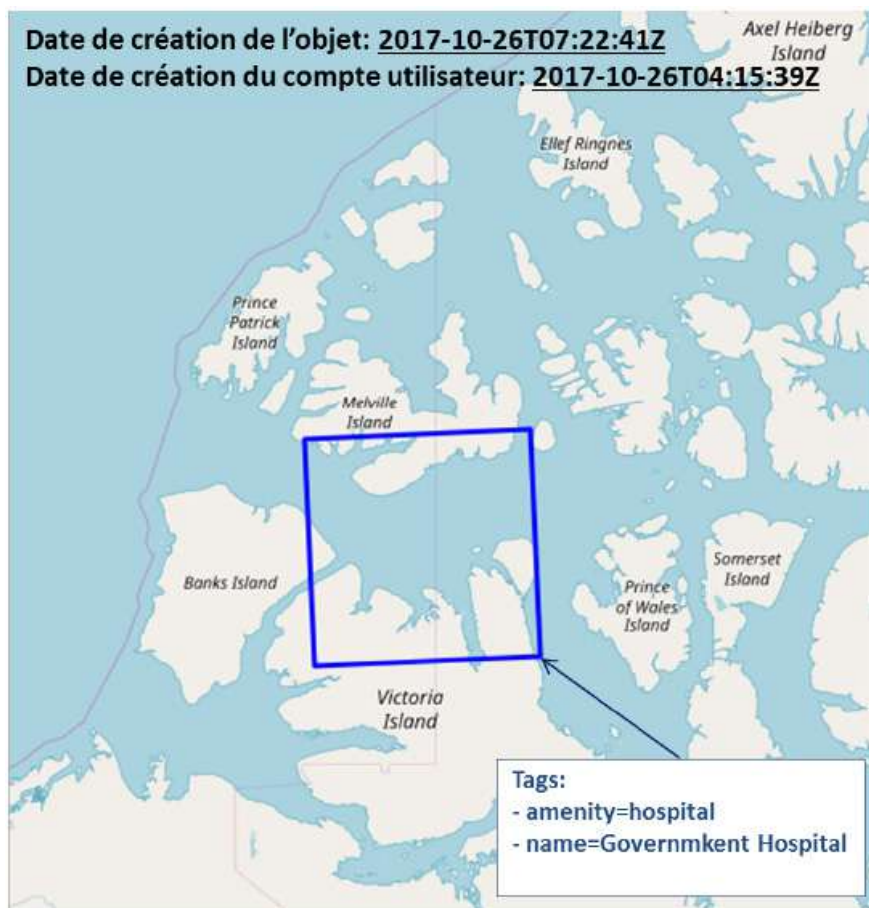


Figure 1 : Hôpital dont la géométrie chevauche plusieurs îles du Nord du Canada dans OpenStreetMap.

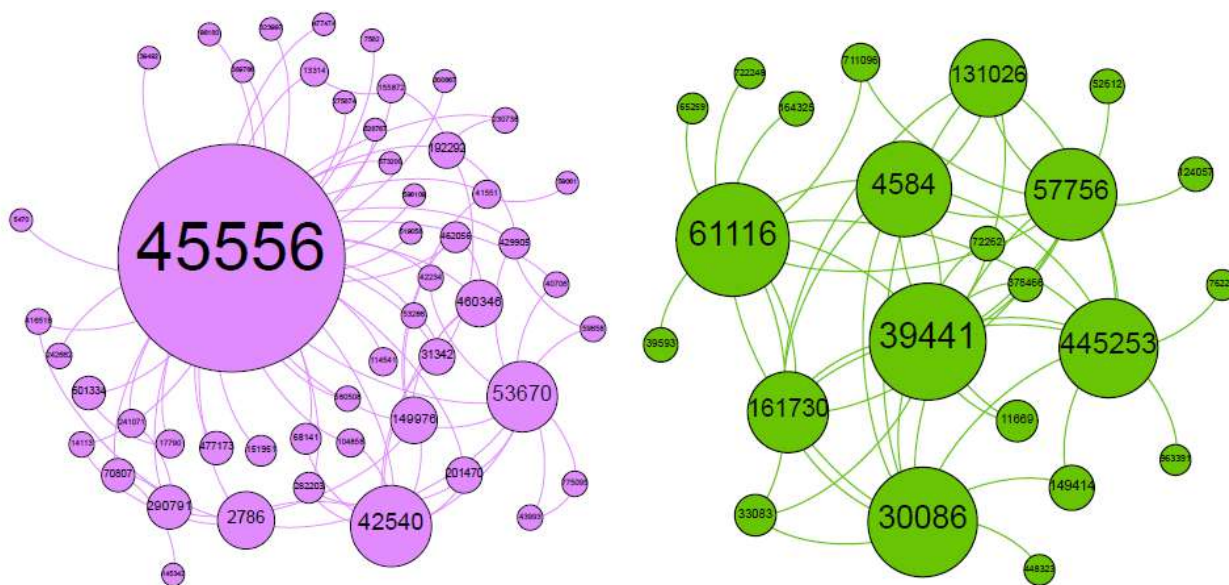


Figure 2 : Deux exemples de communautés extraites de l'analyse des interactions entre contributeurs OSM sur la ville allemande de Stuhr.

par des modérateurs du projet OSM comme étant du vandalisme cartographique (fig. 1). L'analyse de cet échantillon de contributions de carto-vandalisme nous a permis de valider la définition proposée.

## Qualité du contributeur

En considérant que les données spatiales corrompues proviennent de contributeurs malicieux, cette thèse défend que l'évaluation de la qualité du contributeur d'information géographique volontaire permet d'évaluer la qualité de ses contributions. En ce sens, nous avons cherché à étudier le comportement du contributeur d'information géographique volontaire. En particulier, les opérations cartographiques autorisées sur une plateforme de saisie collaborative – tels que l'ajout, la modification, la complétion ou la suppression de données – constituent différents moyens à travers lesquels les contributeurs peuvent interagir et collaborer entre eux. L'analyse des interactions qu'a un contributeur avec le reste de la communauté de contributeurs à partir de son activité cartographique fournit des informations sur sa fiabilité et donc sur la qualité de ses contributions (fig. 2).

La modélisation des interactions entre les contributeurs d'information géographique volontaire sous forme de graphes sociaux est un moyen judicieux car nous avons pu exploiter des méthodes d'analyse issues de la théorie des graphes pour étudier les collaborations entre contributeurs. Par ailleurs, puisque différents types d'interaction illustrent des aspects différents du comportement d'un contributeur, il a fallu construire plusieurs graphes sociaux matérialisant des interactions différentes. Pour étudier ces multiples graphes, nous avons construit un graphe multiplexe, qui est un modèle développé dans le domaine de l'analyse des réseaux sociaux pour représenter les différentes dimensions dans lesquelles de mêmes individus partagent ou non des relations de nature différente (au travail, en famille, entre amis, sur les réseaux sociaux, par exemple). Nous proposons donc d'exploiter un modèle de réseau multiplexe de collaboration des contributeurs qui contient plusieurs graphes d'interactions. Ce modèle permet de représenter autant de formes de collaboration que possible, et donc de représenter de manière plus réaliste le comportement des contributeurs.

Expérimentalement, nous avons implémenté un réseau multiplexe de collaboration de contributeurs d'OSM. En y lançant des calculs de détection de communauté, nous avons pu mettre en évidence des profils contributeurs fiables, de type modérateur

ou pionnier, dont les contributions se sont avérées être de bonne qualité. Dans le but de qualifier globalement tous les contributeurs, les résultats issus de l'analyse multiplexe sont ensuite réutilisés sous la forme d'indicateurs de fiabilité. En particulier, deux méthodes sont testées pour évaluer la fiabilité du contributeur : la première consiste en un score de fiabilité, tandis que la seconde produit un classement des contributeurs selon un algorithme de décision multicritère (PROMETHEE-II). Les expériences menées sur des contributeurs d'OSM ont alors montré que les indicateurs construits à partir des graphes d'interaction permettent de filtrer plus précisément les contributeurs non fiables.

## Détection du carto-vandalisme par apprentissage automatique

La dernière partie de cette thèse explore la capacité des méthodes d'apprentissage à détecter le carto-vandalisme. Notre état de l'art sur la détection du vandalisme dans les bases de connaissances ouvertes ont permis d'identifier des indicateurs et des méthodes pertinentes à cette problématique, et qui semblaient utiles pour détecter le vandalisme cartographique. Par ailleurs, pour évaluer la performance de ces méthodes d'apprentissage dans la détection du carto-vandalisme, nous avons construit le tout premier corpus de données labélisées, composées de contributions OSM et de contributions artificielles de carto-vandalisme. Le processus expérimental vise à répondre à trois problématiques : 1) trouver les bonnes variables descriptives qui expliquent le carto-vandalisme ; 2) évaluer la performance des méthodes d'apprentissage automatique ; 3) améliorer le corpus de données de carto-vandalisme.

Les expériences utilisant les méthodes d'apprentissage non-supervisées ont démontré la nécessité des indicateurs sur les contributeurs pour détecter le carto-vandalisme, ce qui appuie la pertinence de notre recherche sur la qualité des contributeurs. De plus, les indicateurs de comparaison avec des données d'autorité se sont révélés utiles pour détecter des contributions erronées ou anormales. Quant aux méthodes d'apprentissage supervisées, les expériences ont montré que les forêts aléatoires prédisent correctement le carto-vandalisme sur une zone, à condition d'avoir au préalable entraîné le système de classification sur la zone à prédire. En revanche, les faibles résultats issus de la détection du carto-vandalisme par des réseaux de neurones de convolution encouragent à poursuivre les recherches pour déterminer si cette méthode peut être exploitée pour réaliser des tâches spécifiques dans un processus plus long de détection du carto-vandalisme.

Enfin, nos analyses ont permis de mettre en évidence les limites du corpus de données labélisées, notamment à cause de la « faible qualité » des contributions artificielles de carto-vandalisme (fig. 3). Nous avons alors proposé des idées et des

suggestions pour récupérer des contributions réelles de carto-vandalisme, telles que la mise en place d'un système d'annotation de contributions réelles, qui peut s'effectuer de manière manuelle ou automatique.



Figure 3 : Exemples d'édérations vandalisées artificielles créées pour ce travail

---

## Bibliographie sommaire

Truong, Q. T., Touya, G. et Runz, C. (2020) "OSMWatchman: Learning How to Detect Vandalized Contributions in OSM Using a Random Forest Classifier. ISPRS Int. J". *Geo-Inf.* 2020, 9, 504.

Truong, Q.-T., de Runz, C. et Touya, G. (2019). « Analysis of collaboration networks in OpenStreetMap through weighted social multigraph mining », *International Journal of Geographical Information Science*, 33(8):1651-1682.

Truong, Q. T., Touya, G. et de Runz, C. (2018). « Le vandalisme dans l'information géographique volontaire : apprendre pour mieux détecter ? », *Actes de la conférence SAGEO 2018*, pages 61-76, Montpellier, France.

Truong, Q.-T., Touya, G. et de Runz, C. (2018). « Towards Vandalism Detection in OpenStreetMap Through a Data Driven Approach ». In Winter, S., Griffin, A. et Sester, M., (eds.): *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 de Leibniz International Proceedings in Informatics (LIPIcs), Dagstuhl, Germany. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Truong, Q. T., De Runz, C. et Touya, G. (2018). « Analyse du comportement des contributeurs dans l'Information Géographique Volontaire via la construction de réseaux sociaux ». In Runz, C de., Kergosien, É., Guyet, T. et Sallaberry, C., (eds) : *18ème Conférence Internationale Sur l'Extraction et La Gestion Des Connaissances (EGC 2018)*, pages 44-54, Paris, France.

Truong, Q.-T., Touya, G. et de Runz, C. (2018). « Building Social Networks in Volunteered Geographic Information Communities: What Contributor Behaviours Reveal About Crowdsourced Data Quality ». In Fogliaroni, P., Balla-tore, A. et Clementini, E. (eds): *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, *Lecture Notes in Geoinformation and Cartography*, pages 125-131. Springer International Publishing.