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/AGROCAMPUS-OUEST)
Christian Sallaberry (LIUPPA, Université de Pau et des Pays de l'Adour)

http://gt-gast.irisa.fr/gast-2018/

Mardi 23 janvier 2018, Paris

PRÉFACE

Le quatrième atelier « Gestion et Analyse des données Spatiales et Temporelles » (GAST) est associé à EGC'2018. 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 quatre soumissions présentées à l'atelier GAST'2018 :

- Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle de Farah Amina Zemri, Karine Zeitouni et Djamila Hamdadou ;
- Extraction de Localisations dans les MicroBlogs de Thi-Bich-Ngoc Hoang et Josiane Mothe ;
- Fouille de Motifs Graduels Fermés Fréquents Sous Contrainte de la Temporalité de Jerry Lonlac, Benjamin Negrevergne, Yannick Miras, Aude Beauger et Engelbert Mephu Nguifo ;
- Analyse du comportement des contributeurs dans l'Information Géographique Volontaire via la construction de réseaux sociaux de Quy Thy Truong, Cyril De Runz et Guillaume Touya.

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 Géraldine Del Mondo (LITIS) pour son intervention en tant que conférencier invitée à la journée GAST'2018. En espérant que ces articles vous apporteront de nouvelles perspectives autour de la Gestion et l'Analyse des données Spatiales et Temporelles, nous vous souhaitons une bonne lecture.

Eric Kergosien
Université Lille/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, Rennes
Marta Severo – DICEN-idf, Paris Nanterre
Géraldine Del Mondo – LITIS, Rouen
Frédéric Flouvat – PPME, Nouméa
Florence Le Ber – ENGEES, Strasbourg
Simon Malinowski – IRISA, Rennes
Nicolas Meger – LISTIC, Annecy
Sandro Bimonte – IRSTEA, Clermont-Ferrand
Mathieu Roche – CIRAD, Montpellier
Fatiha Saïs – LRI, Paris
Serge Guillaume – IRSTEA, Montpellier
Maguelonne Teisseire – IRSTEA, Montpellier
Karine Zeitouni – PRISM, Versailles St Quentin
Christine Plumejeaud – LIENs, La Rochelle
Sylvain Bouveret – LIG, Grenoble
Perret Julien – IGN/COGIT, Paris
Laurent Etienne – LI, Tours
Ana-Maria Raymond Olteanu – IGN/COGIT, Paris
Ludovic Moncla, IRENav/École Navale de Brest

TABLE DES MATIÈRES

Présentation invitée

Articles de l'atelier

Modélisations spatio-temporelles et graphes

Géraldine Del Mondo*

* INSA Rouen / LITIS

References

Géraldine Del Mondo, G., J. G. Stell, C. Claramunt, and R. my Thibaud (2010). A graph model for spatio-temporal evolution. *Journal of Universal Computer Science 16*(11), 1452–1477.

Résumé

La complexité inhérente aux phénomènes spatio-temporels de part la nature des problèmes sous-jacents (e.g. épidémiologie, analyse de l'évolution de l'espace agricole, gestion de crise) et leur structure multi-dimentionnelle entre autres éléments rend leur modélisation tout aussi complexe. De nombreuses propositions sont faîtes utilisant des outils très différents (e.g. SIG, ontologies, graphes) afin de représenter les dimensions spatiale, temporelle et sémantique du phénomène à étudier. Les objectifs de telles modélisations sont de pouvoir analyser, extraire de l'information du phénomène, voire de proposer une visualisation alternative de celui-ci.

Cet exposé se focalisera sur les propositions de modélisation de phénomènes spatio-temporels fondées sur la théorie des graphes. Une première partie présentera une rétrospective subjective d'un certain nombre de modèles spatio-temporels fondés sur cette théorie, il s'agit de montrer ce que l'on entend par graphe spatial ou graphe spatio-temporel sur quelques exemples. L'idée n'est pas d'être exhaustif ni même de faire une étude comparative rigoureuse mais plutôt de montrer leur diversité en termes de domaines concernés, d'objectifs à atteindre et de structure. Par la suite un focus sera fait sur le modèle GST développé dans Géraldine Del Mondo et al. (2010). Enfin, la présentation s'achèvera par une présentation succincte de travaux en cours en lien avec les graphes spatio-temporels.

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle Application dans la surveillance épidémiologique

Farah AminaZemri* ,Karine Zeitouni**
Djamila Hamdadou*

*Laboratoire LIO-Université d'Oran
BP 1524 El M' Naouer 31000 Oran, Algérie
{zemri_farah,dzhamdadoud}@yahoo.fr

**Laboratoire DAVID-Université de Versailles-Saint-Quentin
45, avenue des Etats-Unis 78035 Versailles, France
karine.zeitouni@uvsq.fr

Résumé. Nous présentons, dans cet article, une méthodologie générique pour la construction automatique d'une sectorisation guidée par les données de la région d'étude. Ce découpage spatial se base sur les valeurs des indicateurs d'autocorrélation spatialeglobale (dont indices de Moran et de Geary) et locale (dont LISA d'Anselin et Local Gi* de Getis-Ord). Ces indicateurs révèlent des structures spatiales, telles que des tendances ou des points chauds et permettent d'assister l'analyste pour comprendre le phénomène analysé et détecter des anomalies. Nous proposons dans cet article une démarche systémique, puis nous nous focalisons sur la phase exploratoire en exploitant l'autocorrélation spatiale et en proposant un indice d'autocorrélation spatio-temporelle. Le besoin d'analyse à la fois spatial et spatio-temporel est récurrent en épidémiologie, domaine d'application de nos travaux.

1 Introduction

L'analyse spatiale combine des approches statistiques exploratoires et la fouille de données suit une démarche d'abord descriptive (en appliquant des méthodes de statistiques descriptives) puis exploratoire (en analysant les données monothématiques) et en fin explicative ou prédictive (sur des données multithématiques) dans un but d'aide à la décision. L'analyse spatiale exploratoire s'articule autour de trois composantes qui sont i) la visualisation qui permet de représenter les données spatiales à travers un système d'information géographique, ii) l'analyse spatiale exploratoire de données qui permet d'explorer et de synthétiser les données afin de révéler des configurations spatiales particulières (patrons, relations spatiales) et iii) la modélisation spatiale qui tente d'expliquer les configurations par la spécification d'un modèle statistique et l'estimation des paramètres. Nous constatons que les étapes d'analyse spatiale exploratoire et celles de la fouille de données spatiales (FDS) se chevauchent pour conduire à une démarche systémique et décisionnelle. L'exemple le plus connu d'analyse spatiale exploratoire est celui du Dr Snow, lequel a découvert la cause de la maladie du choléra à Londres en 1854, grâce à une cartographie et une analyse visuelle de la relation entre les cas de décès et les puits d'eau. À travers cet exemple (Zeitouni, 2006) distinguent la différence entre l'analyse spatiale et la fouille de données spatiales.

Analyse cartographique	Fouille de données spatiales
découverte visuelle	découverte automatique
inapplicable sur de gros volumes de données	opère sur de gros volumes de données
Confirmatoire	Exploratoire (génère des hypothèses)

TAB.1–Comparaison entre l'analyse spatiale et la FDS.

La fouille de données spatiales se distingue alors de l'analyse spatiale par le fait qu'elle peut générer automatiquement à partir de gros volumes de données, des hypothèses par des règles caractéristiques ou bien des règles de classement ou encore des règles d'association.

Dans cet article nous proposons une démarche systémique pour analyser des données spatiales par combinaison des différentes approches dans des étapes successives d'analyse spatiale. Nous nous focalisons ensuite sur la phase exploratoire basée sur les indices d'autocorrélation spatiale. Surtout, nous proposons une extension de l'indice local d'autocorrélation spatiale LISA, au contextespatio-temporel – baptisé LISTA -. Tout au long de cet article, nous illustrons nos propos par une application réelle en épidémiologie.

Cet article est organisé comme suit. La section 2 décrit la méthodologie proposée. La section 3résume les travaux de l'état de l'art en relation avec l'autocorrélation spatiale et spatio-temporelle. La section 4 présente précisément les mesures utilisées pour discerner l'autocorrélation spatiale globale et locale. A la fin de cette section, nous introduisons l'autocorrélation spatio-temporelle locale et son application sur le cas d'étude. Nous concluons notre propos par une conclusion et une discussion des perspectives de ce travail.

2Description de la méthodologie

La démarche systémique que nous avons adoptée est largement inspirée des travaux dans (Zeitouni et al, 1999). La figure 1 illustre cette démarche.

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle.

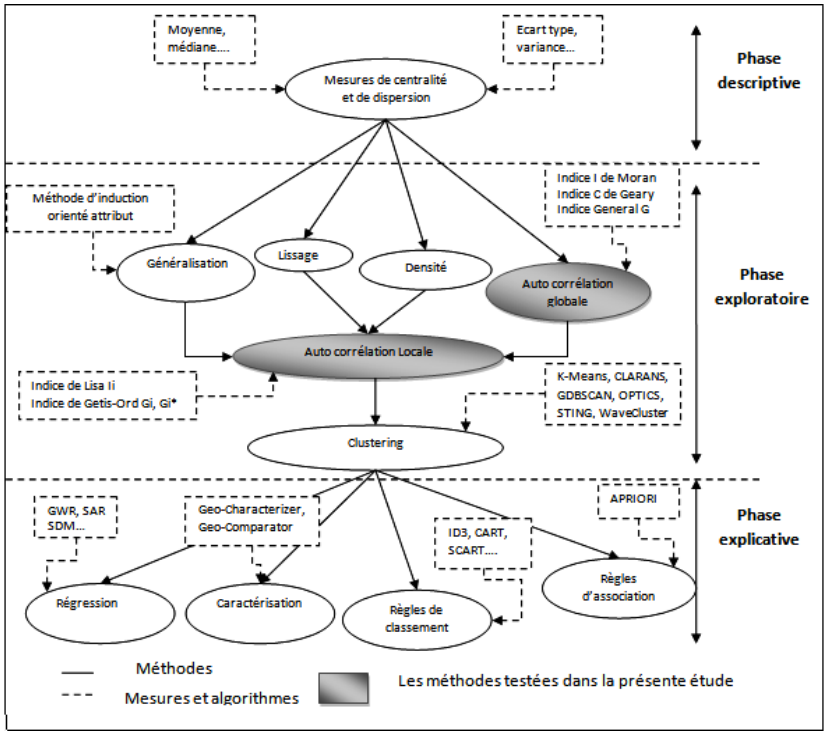


FIG.1–Démarche systémique de la fouille de données spatiales

La figure 1 représente une méthodologie générique pour toute étude spatiale qui vise à analyser l'information géographique dans différents domaines de recherche fondamentale et appliquée. La démarche est composée de trois phases principales qui sont : une phase descriptive, une autre exploratoire et la dernière explicative, voir prédictive.

2.1 Phase descriptive

Cette phase utilise les mesures de statistiques descriptives usuelles qui sont la moyenne et la variance. Les mesures de centralité (centre moyen, centre de gravité) et de dispersion (écart type, écart type de la distance, ellipse de dispersion) permettent de comparer rapidement des nuages de points propres à des segments clés de la population. L'effet discriminant ou non de l'espace peut être donc observé.

2.2 Phase exploratoire

Cette phase opère sur des données monothèmes et se décompose en trois sous étapes :

Analyse globale : donne une description synthétique pour faire apparaitre une tendance générale de la distribution spatiale en appliquant l'autocorrélation globale qui peut être mesurée par : l'indice de Moran I (Moran, 1950), l'indice de Geary C (Geary, 1954), l'indice de Getis-Ord General G (Getis et al, 1992) ou par analyse basée sur la densité, analyse multidimensionnelle lissée ou encore par généralisation.
Analyse locale : découvre les spécificités locales des sous ensembles d'objets par les mesures de l'autocorrélation locales (Indice de LISA (Anselin, 1995) ou indice de Getis-Ord Gi(Getis et al, 1992), Gi*(Getis et al, 1995)) ou analyse factorielle.

Clustering: permet de rassembler des agrégats homogènes partageant des caractéristiques similaires (K-Means (Anil, 2010), K-medoids (Kaufman, 1987) CLARANS (Han, 1994), GDBSCAN (Sander el al, 1998), OPTICS (Ankerts, 1999), STING (Wang, 1997), WaveCluster (Gholamhosein,1998)).

2.3 Phase explicative

Cette phase opère sur plusieurs couches thématiques et permet d'expliquer la particularité des agrégats découverts dans la phase exploratoire par des règles caractéristiques (Ester et al, 1998), des règles de classement (Ester et al, 1997) ou règles d'association (Koperski et Han, 1995) qui permettent éventuellement la prédiction de nouvelles données et la prise de décision.

On se focalise dans la présente étude sur la phase exploratoire et plus particulièrement sur l'analyse de l'autocorrélation spatiale globale et locale.

3Travaux connexes

Nous distinguons dans cette section, deux catégories de travaux connexes : l'analyse d'autocorrélation spatiale et l'analyse temporelle et spatio-temporelle.

3.1 Autocorrélation spatiale

Les méthodes d'autocorrélation spatiale et spatio-temporelle sont largement utilisées pour trouver des relations entre phénomènes spatiaux et découvrir des modèles cachés de l'œil de l'analyste à première vue de la représentation géographique des données. Dans (Tsai et al, 2009) les auteurs ont exploité l'indice d'autocorrélation globale I de Moran et l'indice d'autocorrélation locale de Getis-Ord pour décrire et cartographier les agrégats spatiaux et les régions ou sont situés, les 20 principales causes de décès à Taiwan en 2006. Ils ont pu découvrir grâce à cette analyse que la prévalence de la tuberculose correspond à la localisation des cantons autochtones. (Chaikaew et al, 2009) ont pu définir des points chauds de l'épidémie de la diarrhée dans la province de Chiang Mai située dans le nord de la Thaïlande entre 2001 et 2006. Les chercheurs ont exploité les indicateurs locaux de l'association spatiale (LISA) puis ils ont appliqué l'interpolation de distribution en utilisant la méthode de la densité pour créer une surface continue représentant la densité de l'indice de morbidité de la diarrhée DMBR (*Diarrhea MorBidity Rate*). Ils ont pu enfin découvrir que les villages dans les régions du milieu et du nord ont révélé des incidences élevés. Dans (Wang et al, 2012) Les chercheurs ont utilisé l'indice I de Moran pour discerner l'autocorrélation globale et détecter le modèle de la répartition spatiale de la tuberculose dans la ville de Linyi, en Chine entre 2005 et 2010 puis ils ont fait recours à l'indice de Getis-Ord Gi* pour identifier la localisation des points chauds. Récemment, dans (Runzi et al, 2017) des analyses d'autocorrélation spatiale et spatio-temporelle ont été utilisées pour détecter des agrégats spatio-temporels des oreillons dans la province de Shandong en Chine entre 2005 et 2014. Le test de I de Moran était significatif et l'analyse des indicateurs locaux de l'autocorrélation spatiale (LISA) a révélé des agrégats spatiaux importants avec une incidence élevée. Une autre étude comparative des trois méthodes, scan statistique, L'indicateur local d'autocorrélation spatiale (LISA) et la statistique locale Gi a été faite sur les données du cancer des poumons dans le nord-est de la région d'Ohio entre janvier 1994 et Mai 2006(Sikder et al, 2007). Afin de détecter le rôle de la population migrante dans la transmission de la tuberculose, les auteurs dans (Jia et al, 2008), ont utilisé l'autocorrélation spatiale sur la population permanente et la population migrante entre 2000 et 2006 à Pékin en Chine.

3.2Analyse temporelle et spatio-temporelle

Dans la littérature, nous trouvons de nombreuses études intégrant l'aspect temporel de l'étude dans l'analyse spatiale. Les séries temporelles sont couramment utilisées pour modéliser la variabilité

temporelle, surtout quant il s'agit du suivi sur de longues périodes. C'est le cas dans (Reis et Mandl, 2003) pour réaliser des prévisions sur l'utilisation des services d'urgence. (Andrienko et al, 2004) ont proposé CommonGis qui a été utilisé dans le domaine de la foresterie pour suivre le développement de 2600 arbres dans une période de 100 ans avec un intervalle de 5 ans. Certains travaux visent à supporter les séries temporelles longues ou multi-variées. D'autre études ont intégré l'aspect spatial est temporel pour avoir des études plus complètes. Le *Clustering scan statistic* proposé par (Kulldorff, 2005), implémenté dansl'outil SaTScan, forme des clusters en explorant des fenêtres cylindriques avec une base géographique circulaire et dont la hauteur correspond au temps. Cette technique a été exploitée par plusieurs travaux tels que (Runzi et al, 2017), (Wang, 2012) et (Sikder et al, 2017). Une extension de l'autocorrélation spatiale aux données temporelles a également été proposée dans (Hardisty et Klippel, 2010) et implémentée dans l'outil LISTA-Vis. Le prototype a été utilisé pour surveiller la pandémie du virus H1N1 en 2009 aux Etats-Unis. Dans (Cheng et al, 2011), une proposition de l'extension de l'indice de Pearson pour intégrer l'aspect temporel a été proposée et appliquée à la surveillance du flux routier dans la ville de Londres. Récemment, dans (Jay &Shengwen, 2017), une tentative pour étendre l'indice de Moran tenant compte du temps a été proposée après plusieurs autres propositions similaires. Bien que ce dernier reflète une mesure globale et locale de corrélation tenant compte du temps, aucune étendue temporelle n'a été définie pour mesurer l'intervalle entre périodes considérées. Dans cet article, nous introduisons l'indice de LISTA, une extension de l'indice de LISA telle que la matrice de voisinage tient compte d'un intervalle de voisinage temporel plus ou moins large, donné en paramètre.

4 Mesures utilisées

Dans cette section, nous analysons en détails, les différentes mesures utilisées dans les différentes étapes de l'approche proposée ci-dessus permettant de mener une étude dans le domaine spatial. Une attention particulière est donnée aux mesures d'autocorrélation spatiale et spatio-temporelle, globale et locale.

4.1 Les mesures des statistiques descriptives spatiales

Pour mieux comprendre les données, plusieurs mesures sont appliquées afin d'identifier la tendance centrale, les variations et la dispersion des données géographiques. Nous distinguons les mesures de centralité et de dispersion (Morency, 2006).

4.1.1 Mesures de centralité

- **Centre moyen** : c'est l'équivalent de la moyenne, en deux dimensions

$$x_{cm} = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; \quad y_{cm} = \frac{\sum_{i=1}^{n} y_i}{n}$$

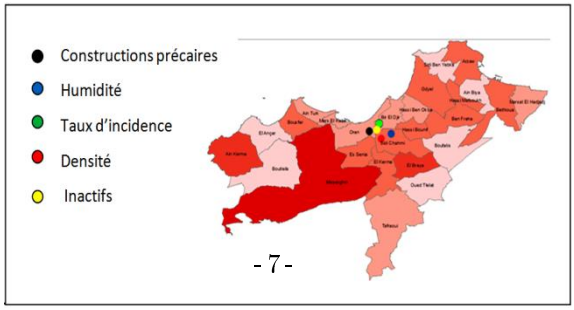


FIG.2–Centre moyen des facteurs de la tuberculose à Oran en 2009

– **Centre de gravité** : c'est le centre moyen avec une intensité observée en un point particulier de l'espace (le poids)

$$x_{cg} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad ; \qquad y_{cg} = \frac{\sum_{i=0}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

4.1.2 Mesures de dispersion

– **Ecart type** :la première mesure de dispersion consiste à mesurer l'écart type (Standard Deviation) dans chacun des axes (x,y), de façon distincte :

$$\boldsymbol{x_{cm}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)} \quad ; \qquad \boldsymbol{y_{cm}} = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{(n-1)}$$

– **Ecart type de la distance** : il s'agit de l'écart type de la distance de chaque point par rapport au centre moyen. C'est donc l'équivalent en deux dimensions de l'écart type classique d'une distribution.

$$\boldsymbol{S_{xy}} = \sqrt{\sum_{i=1}^{n} \frac{(d_{icm})^2}{n-2}}$$

$\boldsymbol{d_{icm}}$la distance entre chaque point et le centre moyen.

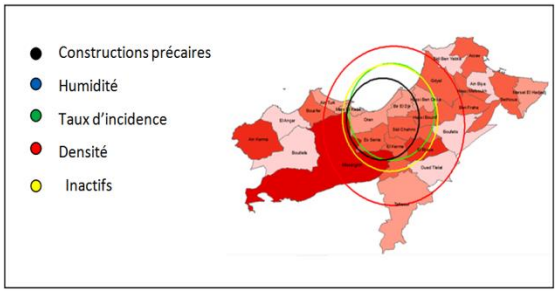


FIG.3–Ecart type de la distance des facteurs de la tuberculose à Oran en 2009

– **L'ellipse de dispersion:** permet de tenir compte de la variance d'état selon la direction considérée (anisotropie) par la définition d'une ellipse. L'orientation de l'ellipse permet d'identifier les axes de dispersion maximale et minimale.

  – L'axe Y est pivoté d'un angle téta

$$\theta = \frac{ARCTAN\{[\sum_{i=1}^{n}(xi-\bar{x})^2 - \sum_{i=1}^{n}(yi-\bar{y})^2] + [(\sum_{i=1}^{n}(xi-\bar{x})^2 - \sum_{i=1}^{n}(yi-\bar{y})^2]^2 + 4(\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y}))^2]^{1/2}\}}{2\sum_{i=1}^{n}(xi-\bar{x})(y_i-\bar{y})}$$

  – Les deux écarts-types sont calculés.

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle.

$$Sx = \sqrt{\frac{\sum_{i=1}^{n}((xi-\bar{x})\cos\theta-(yi-\bar{y})\sin\theta)^2}{n-2}};\ Sy = \sqrt{\frac{\sum_{i=1}^{n}((xi-\bar{x})\sin\theta-(yi-\bar{y})\cos\theta)^2}{n-2}}$$

– La longueur de chaque axe de l'ellipse correspond à deux fois l'écarttype

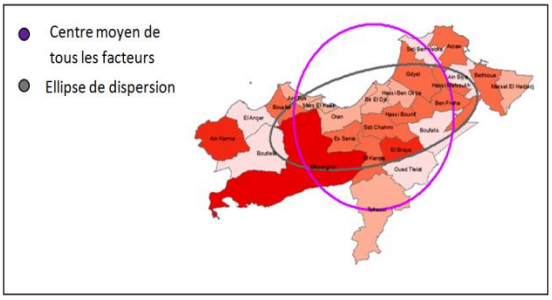


FIG.4–Ellipse de dispersion des facteurs de la tuberculose à Oran en 2009

4.2 Les mesures d'autocorrélation spatiale

L'autocorrélation spatiale est définit comme « la corrélation d'une variable avec elle-même (autocorrélation) attribuable à l'organisation géographique des données (spatiale) » (Griffith, 1991). Les mesures d'autocorrélation spatiale évaluent le degré de similarité entre observations en fonction de leur éloignement (voisinage). On reprend ici la première loi en géographie définie par (Tobler, 1970) : "*Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés*".

4.2.1 Les mesures d'autocorrélation spatiale globale

L'autocorrélation spatiale permet de mesurer la ressemblance entre voisins, en utilisant la notion de graphe de voisinage. Elle est souvent utilisée comme technique exploratoire pour indiquer si une modélisation spatiale est nécessaire. Elle s'applique à des données quantitatives rattachées à des objets polygonaux formant un découpage. Elle se décline en deux méthodes complémentaires : l'indice de Moran (Moran, 1950) et l'indice de Geary (Geary, 1954).

– **Indice de Moran (I)** : L'indice de Moran était présenté la première fois dans (Moran 1950) puis il était popularisé par (Cliff et Ord, 1981). L'indice de Moran (Moran, 1950) mesure l'écart de la valeur d'un point et celle de ses voisins à celle de la moyenne. Il se présente comme le rapport de la covariance sur la variance.

$$I_{Glabal} = \frac{n}{m} \times \frac{\sum_i \sum_j wij(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2}$$

- zi= valeur de la variable au point "i" et de moyenne $\bar{z}$
- i = unité géographique de référence
- j = unités voisines du point "i", définies par la matrice
- n = nombre total d'unités géographiques
- m=nombre total de paires de voisins$\left(\sum_i \sum_j w_{ij}\right)$
- Wij = matrice de pondération

- **Indice de Geary (C)** : L'indice de Geary teste si la variabilité d'une variable entre voisins (donnée par la notion de variance locale) est significativement différente de celle attendue d'un modèle aléatoire (donnée par la variance). L'indice de Geary est défini comme le rapport C de la variance locale et de la Variance globale.

$$C = \frac{{}^{1}/_{2m}\ \sum_i \sum_j w_{ij}\,(xi - xj)^2}{{}^{1}/_{2n(n-1)}\ \sum_i \sum_j (xi - xj)^2}$$

Que soit l'indice I de Moran ou l'indice C de Geary, les deux se basent sur l'étude de similarité et dissimilarité des valeurs voisines pour définir si le modèle spatiale est clustérisé (ressemblance des valeurs voisines), dispersé (des valeurs voisines opposées) ou alors aléatoire (absence d'autocorrélation spatiale) mais n'indiquent pas si les valeurs sont faibles ou élevées. Pour cela, nous avons eu recours à une autre statistique de l'autocorrélation spatiale, l'indice général G (Getis et al, 1992).

- **La statistique générale G :** est une mesure d'autocorrélation spatiale capable de détecter si les points sont chauds ou froids. La statistique G utilise également le produit croisé pour mesurer l'association spatiale, similaire à celle de Moran.

$$G(d) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\,(d) x_i x_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j}$$

- Si G(d) <0 alors Les valeurs groupées sont faibles
- Si G(d)>0 alors les valeurs groupées sont élevées
- Si G(d)=0 alors les valeurs sont proches de la valeur nulle

La statistique G (d) de Getis-Ord donne plus de renseignement sur l'autocorrélation spatiale donnée par l'indice I de Moran. Les statistiques G devraient être utilisés conjointement avec I afin d'identifier les caractéristiques des modèles non révélés par la statistique I toute seule. Les constats faits à l'aide d'indices globaux sont une invitation à explorer davantage les données afin de mieux cerner les différentes configurations spatiales locales (Morency, 2006).

4.2.2 Les mesures d'autocorrélation spatiale locale

Les mesures d'autocorrélation spatiale, appliquées localement, permettent de cerner les secteurs similaires et dissimilaires. Le caractère distinctif ou non d'un secteur en regard de son voisinage peut faciliter la formulation de modèles, surtout dans le choix d'un découpage spatial. L'identification, a priori, des secteurs différents peut faciliter le travail d'analyse corrélative.

- **Indice Getis-Ord Gi et Gi*** : la statistique locale Getis-Ord Gi et Gi* est la version locale de la statistique générale G. Elle indique comment la valeur de chaque unité est associée aux unités voisines à une distance d. A partir d'un ensemble d'entités pondérées, elle identifie les points chauds et les points froids statistiquement significatifs. Cela permet de renseigner si les unités spatiales sont positivement ou négativement associées. La statistique locale Gi peut être écrite comme suit (Getis, 1992, Getis, 1995):

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle.

$$G_i(d) = \left(\frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}\right) \text{ Avec } i \neq j \; ; \qquad G_i^*(d) = \left(\frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}\right) \text{ avec i=j}$$

Avec Wij(d) est le poids avec la distance d

- Si Gi (d) est faible (ou z-score <0) alors il indique des valeurs faibles et inférieures à la moyenne
- Si Gi (d) est élevé (ou z-score >0) alors il indique une association spatiale de valeurs élevées similaires
- Si Gi(d) =0 alors les valeurs s'approchent de la valeur nulle

Les statistiques Gi et Gi* nous permettent de détecter les agrégats (les clusters) locaux de dépendances qui peuvent ne pas apparaître lors de l'utilisation des statistiques globales (I, C ou G).

FIG.5–Analyse de points chauds par autocorrélation locale par indice de Getis-Ord Gi*

Getis et Ord (1992)avaient cherché à mettre en place des indicateurs locaux d'association spatiale, mais sans lien avec les indicateurs globaux existants.Le travail d'Anselin se place dans la continuité de celui de Getis et Ord (1992) en proposant des indicateurs locaux d'Association Spatiale (Local Indice for Spatial Association) permettant le passage d'un indicateur global de la structure spatiale (indice d'autocorrélation spatiale globale) à une mesure locale des ressemblances LISA (Anselin, 1995). Les statistiques locales de Moran (LISA) donnent une mesure de l'autocorrélation spatiale pour chaque localisation.

- **Indice de LISA I_i :** permettent de mesurer le degré de ressemblance d'une unité spatiale avec ses voisines. On peut ainsi révéler les tendances régionales tout en conservant les valeurs locales, c'est-à-dire préserver l'information relative à l'hétérogénéité interne de ces zones.

$$\boldsymbol{I_i} = \frac{\sum_j wij(z_i - \bar{z})\left(z_j - \bar{z}\right)}{\sum_i (z_i - \bar{z})^2}$$

FIG.6–Analyse des clusters par autocorrélation locale par indice de LISA

Selon (Oliveau, 2004), la moyenne des indices locaux est alors égale à l'indice global. Ce lien de proportionnalité entre l'indice global de Moran et les indices locaux des LISA permet d'obtenir pour chaque sous région géographique une estimation de sa ressemblance avec les sous régions voisines par rapport à sa ressemblance à l'ensemble des sous régions. On distingue alors quatre cas de figures:

- **Situation High–High:**Autocorrélation spatiale positive et valeur de l'indice élevé englobe les régions avec un indice fort dans un voisinage qui lui ressemble. (représenté par la couleur noire)
- **Situation High–Low:**Autocorrélation spatiale négative et valeur de l'indice élevé englobe les régions avec un indice fort dans un voisinage qui ne lui ressemble pas (représenté par la couleur orange)
- **Situation Low–High:**Autocorrélation spatiale positive et valeur de l'indice faible englobe les régions avec un indice faible dans un voisinage qui lui ressemble (représenté par la couleur bleue)
- **Situation Low–Low :**Autocorrélation spatiale négative et valeur de l'indice faible englobe les régions avec un indice faible dans un voisinage qui ne lui ressemble pas (représenté par la couleur blanche)

Une fois qu'un niveau de signification a été défini, les valeurs pourraient être tracées sur une carte pour afficher les emplacements des points chauds: sites avec des valeurs élevées avec des voisins similaires (High-High). Les régions identifies en couleur noir sur la carte sont des points chauds (*Hotspots*).

On s'intéresse à l'identification des zones HH qui représentent une anomalie ou un danger dans l'étude de cas proposée. Une analyse plus fine au niveau mois et trimestre pour la surveillance de la tuberculose à Oran utilisant l'indice de LISA est représentée sur la figure 7.

Pour montrer la complémentarité entre les indices de LISA de Anselin et l'indice Gi de Getis-Ord, nous analysons l'exemple suivant :

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle.

	Janvier 2009	Février 2009	Mars 2009	Trimestre1
LISA				
Gi*				
	Avril 2009	**Mai 2009**	**Juin 2009**	**Trimestre 2**
LISA				
Gi*				
	Juillet 2009	**Aout 2009**	**Septembre 2009**	**Trimestre 3**
LISA				
Gi*				
	Octobre 2009	**Novembre 2009**	**Decembre 2009**	**Trimestre 4**
LISA				
Gi*				

FIG.7–Autocorrélation spatiale locale (indice de LISA vs indice de Gi*) de la tuberculose à Oran par mois et par trimestre pour l'année 2009.

Toutes les analyses qui ont été faites ci-dessus, bien que différenciées par périodes (années/trimestres/mois) sont des analyses purement spatiales qui ont été utilisées par application des indices d'autocorrélation locale (LISA et Gi, Gi*) aux niveaux des périodes concernées. Dans ce qui suit, nous introduisons un nouvel indice, LISTA qui mesure l'autocorrélation spatio-temporelle locale et l'utilisons. Nous proposons pour cela, la matrice des indices d'autocorrélation spatio-temporelle.

4.3 Mesures d'autocorrélation spatio-temporelle

La prise en compte de la variabilité temporelle est absente ou peu satisfaisante dans les systèmes précédents. Les objectifs de l'analyse spatiale et l'analyse spatio-temporelle sont différents (Wang et al, 2012). La première vise à identifier les zones présentant une incidence élevée tout au long de la période d'étude, alors que la dernière analyse vise à identifier les zones avec augmentation de l'incidence dans des périodes spécifiques. Pour montrer cette différence (Sharov, 1996) définit les notions de séries temporelle, séries spatiales et séries mixtes et montrent les différences entre elles.

- **Séries temporelles :** séquence de mesures au même point de l'espace prises à différentes périodes temporelles.
- **Séries spatiales :** ensemble de mesures faites simultanément en différents points de l'espace.
- **Séries mixtes :** ensemble de séries temporelles obtenues en différents points dans l'espace

4.3.1Indice d'autocorrélation spatio-temporelle (LISTA)

L'autocorrélation temporelle mesure la ressemblance entre une observation relative à une unité spatiale avec celle antérieure par un intervalle k. Elle est donnée par l'indice de ITA (Indice of Temporal Autocorrelation) de la zone i :

$$ITA(i,k) = \frac{\sum_{w=2}^{l}\left(z_i^{t_w}-\bar{z}^{t_w}\right)\left(z_i^{t_{w-k}}-\bar{z}^{t_{w-k}}\right)}{\sum_{w=1}^{l}\left(z_i^{t_w}-\bar{z}^{t_w}\right)^2} \quad \text{avec } 1 \le \text{w-k} \le \text{w} \le \text{l}$$

L'analyse de l'autocorrélation temporelle est par conséquent analogue à celle de l'autocorrélation spatiale qui mesure la ressemblance entre observations voisines. L'autocorrélation spatio-temporelle entre paires de voisins est donnée par l'indice croisé de corrélation temporelle CITC (Cross Indice of Temporal Correlation) qui est donné par la formule suivante :

$$CITC(i,j,k) = \frac{\sum_{w=2}^{l} W_{ij}\left(z_i^{t_w}-\bar{z}^{t_w}\right)\left(z_j^{t_{w-k}}-\bar{z}^{t_{w-k}}\right)}{\sum_{w=1}^{l}\left(z_i^{t_w}-\bar{z}^{t_w}\right)^2} \text{ avec } 1 \le \text{w-k} \le \text{w} \le \text{l}$$

L'autocorrélation spatio-temporelle peut être donc mesurée par l'indice de LISTA (Local Indice of Spatial Temporal Association) que nous proposons comme une extension de l'indice de LISA (Local Indice of Spatial Association) vu précédemment. LISTA est défini par la formule suivante :

$$LISTA(i,k) = \frac{\sum_j \sum_{w=2}^{l} W_{ij}\left(z_i^{t_w}-\bar{z}^{t_w}\right)\left(z_j^{t_{w-k}}-\bar{z}^{t_{w-k}}\right)}{\sum_i \sum_{w=1}^{l}\left(z_i^{t_w}-\bar{z}^{t_w}\right)^2} \quad \text{avec } 1 \le \text{w-k} \le \text{w} \le \text{l}$$

- k = intervalle de temps entre périodes considérées (k=1 entre t_1 et t_2 et k=2 entre t_1 et t_3)
- t_w= période considérée
- i=unité spatiale
- j=unité voisine

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle.

- l= nombre des périodes considérées (si l'unité est le mois, l=4 correspond à l'étude trimestrielle et l= 12 à l'étude mensuelle)
- $z_i^{t_w}$= mesure dans la zone considérée au temps t_w
- $z_j^{t_w}$= mesure dans une zone voisine à zi au temps t_w
- Wij = la matrice de voisinage pondérée

Trimestre 1	Trimestre 2
	
Trimestre 3	**Trimestre 4**
	

FIG.8–Représentation cartographique des données de la tuberculose à Oran par trimestre

4.3.2 Matrice des indices de l'autocorrélation spatio-temporelle

Nous définissons la matrice des indices de l'autocorrélation spatio-temporelle par une matrice des indices de l'autocorrelation spatiales mutipliée par le nombre des périodes considérées par l'étude temporelle. La figure 9 représente une matrice des indices de l'autocorrélation spatio-temporelle locale étendue sur quatres périodes successesives des trimestre (T1,T2,T3,T4). Une cellule dans la matrice des indices de l'autocorrélation spatio-temporelle représente la mesure de l'autocorrélation spatio-temporelle entre la région i dans $t_w$et la région j dans t_{w-k}.

Pour calculer les valeurs de la matrice des indices d'autocorrélation spatio-temporelle, nous intégrons les considérations suivantes :

- Calculer l'indice de LISTA entre trimestres pour les mêmes régions revient à calculer l'autocorrélation temporelle entre une région zi avec ses voisines zj. Ces valeurs représentent les valeurs des sous matrices (T1-T2, T2-T3, T3-T4, T1-T3, T2-T4, T1-T4)
- Calculer l'indice de LISTA entre régions dans le même trimestre revient à calculer l'indice de LISA tout simplement (LISTA équivaut à LISA dans ce cas). Ces valeurs représentent les valeurs des sous matrices (T1-T1, T2-T2, T3-T3, T4-T4).
- La diagonale de toute la matrice des indices de l'autocorrélation spatio-temporelle représente les valeurs de LISA des régions zi avec elles mêmes (ces valeurs sont ignorées et remplacées par la valeur nulle).

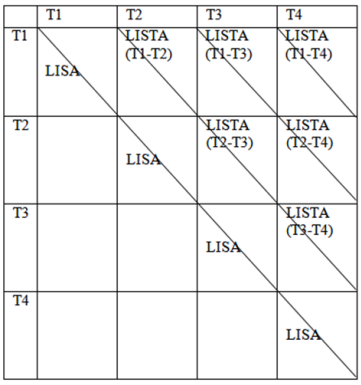


FIG.9–Illustration de LISTA définie sur la matrice des indices d'autocorrélation spatio-temporelle

- Calculer l'indice de LISTA entre trimestres pour les mêmes régions revient à calculer l'autocorrélation temporelle entre une région zi avec elles mêmes. Ces valeurs représentent les lignes qui marquent les diagonales des sous matrices (T1-T2, T2-T3, T3-T4, T1-T3, T2-T4, T1-T4) et représentent les indices d'autocorrélation temporelles LISTA (1, tw), LISTA (2, tw) et LISTA (3, tw).
- Le reste des cellules est calculé par l'indice de LISTA proposé.

Les représentations cartographiques des indices d'autocorrélation spatio-temporelles calculés par indice de LISTA sont représentées sur la figure 10.

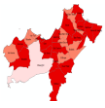
	T1	T2	T3	T4
T1				
T2				
T3				
T4				

FIG.10–Représentation cartographique de l'autocorrélation spatio-temporelle de la tuberculose à Oran par trimestre calculée par indice de LISTA

Sectorisation guidée par l'autocorrélation spatiale et spatio-temporelle.

Nous pouvons distinguer clairement qu'il existe une forte autocorrélation spatio-temporelle entre les trimestres voisins (T1-T2, T2-T3, T3-T4) et une faible autocorrélation spatio-temporelle entre les trimestres non consécutifs (T1-T3 et T2-T4). Ce résultat confirme d'une part que le phénomène de propagation de l'épidémie dépend bien du temps et d'autre part, indique que sa propagation est bien visible d'un trimestre au trimestre suivant, mais pas lorsque ces trimestres sont séparés d'un ou plusieurs trimestres (valeur de k). Plus l'intervalle k est petit plus l'autocorrélation spatiotemporelle est significative.

Afin de montrer l'utilité de l'indice de LISTA par rapport à l'indice de LISA, nous prenons les exemples de quelques valeurs representées sur le tableau 2. Ces valeurs sont extraites des sous matrices de la matrice des indices de l'autocorrélation spatio-temporelle de la figure 10. Ces valeurs sont calculées par la formule de l'indice CITC ci-dessus.

Region zi	**Region zj**	**CITC (i,j,k)**
Elbraya(T1)	Boufatis(T2)	1,1043427
El ancor(T1)	Misserguine(T2)	1,0717802
Misserguine(T1)	Bousfer(T3)	1,0296758
Misserguine(T2)	Ain Turck(T3)	1,0066348
Boufatis(T3)	ElBraya(T4)	0,5335526
Misserguine(T2)	AinTurk(T4)	0,9613501
Misserguine(T2)	Boutlelis(T4)	0,9677061
Misserguine(T2)	ElKerma(T4)	0,7492245
Misserguine(T2)	EsSenia(T4)	0,6356142

TAB.2–Exemples de valeurs de CITC (i, j, k)

Nous constatons à travers les résultats de l'étude de l'autocorrelation spatio-temporelle entre les régions voisines de la zone d'étude qu'il y avait de fortes autocorrelations entre les regions zi dans les trimestres (T1,T2,T3) et les régions zj dans les trimestres (T2,T3,T4). Par exemple la région Elbraya dans T1 est fortement autocorrélée avec la région Boufatis dans T2. La méme chose entre Elancor dans T1 avec Missserguine dans T2 avec les indices de CITC respectifs 1,1 et 1,07. De plus,Misserguine dans T2 est fortement autocorrélée avec sa zone voisine Ain Turck dans T3. Celasignifie que l'épidemie s'estpropagée de Misserguine vers Ain Turk mais cette dernière n'est pas autocorrélée avec ses zones voisines (Bousfer, M El Kebir et Misserguine) dans T4 avec les indices d'autocorrélation de CITC respectifs 0,23 ;0,24 ; 0,30. Cela signifie que la propagation de l'épidemie a étélimitéedans le temps au trimestre T3 et ne s'est pas étendue aux zones voisines pendant le trimestre suivant T4. Par contre, la région de Misserguine en trimestre T2 était fortement autocorélée avec ses voisines (Ain Turck, Boutlelis, El Kerma, et Es Senia) avec des valeurs significatives de CITC, respectivement : 0,9613501 ; 0,9677061 ; 0,7492245 ; et 0,6356142. Ceci signifie que l'épidemie s'est propagée rapidement entre lestrimestresT3 etT4, de la région de Misserguine à ses régions voisines. On observe aussi que la région de Misserguine a été contaminée par la région d'El Ancor au cours du trimestre T2, puis elle a été la cause de la contamination de toutes ses zones voisines plus tard lors du trimestreT4.

Les scénarios qui ont été révélés grace au calcul de l'indice de LISTA n'ont pas été détectés auparavant par une analyse d'autocorrelation spatiale seule. L'indice de LISTA est capable de détecter des tendances à travers le temps car il opére sur plusieurs couches temporelles. Une région dans T1 peut être autocorrélée avec une région voisine dans T2 même si ces deux régions ne sont pas forcément autocorrélées spatialement dans le même trimestre T1 ou T2. Une analyse plus fine au niveau mensuel mérite d'être explorée afin de dicerner l'autocorrélation spatio-temporelle entre les mois et détecter d'autres scénarios pour savoir à quel moment la propagation a été declanchée et à quel endroit.

5Conclusion et perspectives

Dans cet article, une extension de l'indice de LISA pour intégrer l'aspect temporel a été proposée et appliquée à la surveillance de la tuberculose à Oran. Cette proposition a été faite en proposant la matrice des indices de l'autocorrélation spatio-temporelle permettant de considérer le temps et l'espace en même temps. Nous nous sommes limités dans ce travail au niveau trimestre pour tester l'applicabilité de l'indice de LISTA proposé. Nous pouvons également proposer une analyse des séries spatiotemporelles en se basant sur l'indice de LISTA. Dans nos futurs travaux, nous proposerons une analyse plus fine en descendant dans la granularité temporelle aux niveaux mois, semaine, jour pour révéler les agrégats spatio-temporels non détectés dans l'étude trimestrielle. L'échelle d'observation aura certainement un impact sur la compréhension du phénomène étudié. Nous avons par ailleurs proposé une démarche qui va de l'analyse descriptive à l'analyse explicative, voir prédictive par divers outils et approches. Il reste à intégrer véritablement de tels outils.

Summary

In this paper, we present a generic methodology for the automatic construction of data-driven sectorization of the study area. This spatial division is based on the values of global spatial autocorrelation indicators (including Moran's and Geary's indices) and local indicators (including LISA of Anselin and Local Gi * of Getis-Ord). These indicators reveal spatial structures, such as trends or hotspots, and help the analyst to understand the phenomenon being analyzed and to detect anomalies. We propose in this paper a systemic approach, then we focus on the exploratory phase exploiting the spatial autocorrelation and proposing a spatio-temporal autocorrelation index. The need for both spatial and spatio-temporal analysis is recurrent in epidemiology, the field of application of our work.

Références

Anil K.J (2010). Data clustering : 50 years beyond k-means. Pattern Recognition Letters, pages 651 666.

Ankerst.M, Marku. M.B, Kriegel H.P et Sander.J (1999). OPTICS : ordering points t identify the clustering structure. In ACM Sigmod Record, volume 28, pages 49 60.ACM,

Anselin, L (1995). Local indicators of spatial association - LISA. Geographical Analysis, 27, 2, pp. 93-115

Ben Y.R and Mandl.K.D (2003). Time Series modeling for syndromic surveillance. BMC Medical Informatics and Decision Making, 3:2

Chaikaew. N, Tripathi. N.K, Souris M (2009). Exploring spatial patterns and hotspots of diarrhea in Chiang Mai. Thailand. Int J Health Geogr 2009, 8:36.

Cliff, A. D., and J. K. Ord (1981). Spatial Promsm: Models and Applications. London: Pion Press.

Ester, M., Kriegel, H.P., Sander, J.(1997). Spatial Data Mining: A Database Approach, Proceedings of the 5th Symposium on Spatial Databases, Berlin, Germany.

Geary, R.C., (1954). "The contiguity ratio and statistical mapping".The incorporated Statistician, vol.5, n°3,pp.115–145.

Andrienko.G, Andreinko.N and Gatalsky.P (2004). Visual Mining of Spatial Time Series Data. Knowledge Discovery in Databases: kdd 2004, Proceedings, 3202, pp. 524-52.

Getis, A., Ord, K.J., (1992)."The analysis of spatial association by use of distance statistics", Geographical analysis, Vol. 24, No. 3, pp. 189-206.

Gholamhosein.S, Surojit C.J and Aidong. Z (1998). Wavecluster : A multi-resolution clustering approach for very large spatial databases. In VLDB, volume 98, pages 428 439.

Griffith DA, Arnrhein CG (1991). Statistical Analysis for Geographers Engle-Wood Cliffs, NJ : Prentice Hall.

Hardisty F, Klippel A (2010). Analysingspatio-temporal autocorrelation with LISTA-Viz. Int J Geogr Inf. Sci 24(10):1515–1526.

Jay. L & Shengwen.L. (2017). Extending Moran's Index for Measuring Spatiotemporal Clustering of Geographic Events published in journal of Geographical analysis. Volume 49, Issue 1January 2017 .Pages 36–57

Jia ZW, Jia XW, Liu YX, Dye C, Chen F, Chen CS, Zhang WY, Li XW, Cao WC, Liu HL(2008). Spatial analysis of tuberculosis cases in migrants and permanent residents, Beijing, 2000–2006. Emerg Infect Dis, 14(9):1413–1419.

Kaufman. L and Rousseeuw. P.J (1990). Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons.

Koperski K. and Han J., (1995). "Discovery of Spatial Association Rules in Geographic Information Databases", In Advances in Spatial Databases (SSD'95), p. 47-66, Portland, ME, August.

Kulldorff. M, (2005). «SaTScanTMv6.0 : Software for the spatial and space-time scan statistics, Information Management Services, INc.,

Moran, P.A.P., (1950). "Notes on continuous stochastic phenomena", Biometrika, Vol. 37, n°1/2, pp. 17-23.

Morency. C (2006). Etude de méthodes d'analyse spatiale et illustration à l'aide de micro données urbaines de la grande région de Montreal. les Cahiers Scientifiques du TransportN° 49/2006 - Pages 77-102

Ng R., Han J., (1994). ″Efficient and effective clustering method for spatial data mining″, in proceeding of international conference on very large database, Santiago, Chile, September, p-144-155.

Oliveau.S (2004). Modernisation villageoise et distance à la ville en Inde du Sud, Thèse de doctorat, UMR 8504 Géographie-cités, Université Paris 1 PanthéonSorbonne

Ord JK, Getis A (1995). Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis,27:286-306.

Runzi L , Shenghui C, Cheng L , Shannon R, Jin C, Qinqin X, Xiaodong L, Yanxun L, Fuzhong X, Qing X & Xiujun L(2017). Epidemiological Characteristics and Spatial-Temporal Clusters of Mumps in Shandong Province, China, 2005–2014.

Sander J., Ester M., Kriegel H.-P., Xu X.(1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in Data Mining and Knowledge Discovery, An International Journal, Kluwer Academic Publishers, Vol. 2, No. 2

Sharov A. A. (1996). Quantitative Population Ecology. Virginia Tech, Online. lecture (http://www.ento.vt.edu/~sharov/PopEcol/)

Sikder I & Woodside J. (2007). Detection of Space-Time Cluster.International Conference on Information and Communication Technology. ICICT 2007, 7-9 March 2007, Dhaka, Bangladesh.

Cheng Tao, Haworth J, Wang J (2011). Spatio-temporal autocorrelation of road network data. J GeogrSyst14 : 389-413. DOI 10.1007/s10109-011-0149-5

Tsai PJ, Lin ML, Chu CM, Perng CH (2009). Spatial autocorrelation analysis of health care hotspots in Taiwan in 2006. BMC Publ Health, 9:464.

Tukey. J (1977). Exploratory Data Analysis.Addison-Wesley, 691 p.

Wang, T., Xue, F., Chen, Y., Ma, Y., & Liu, Y. (2012). The spatial epidemiology of tuberculosis in Linyi City, China, 2005–2010. BMC public health, 12(1), 885

Wang W, Yang J, Muntz R (1997). Sting : A statistical information grid approach to spatial data mining. Technical report CSD- 97006, computer science department, university of California, Los angeles.

Zeitouni K et Yeh L (1999). Le Data Mining Spatial et les bases de données spatiales. Revue internationale de géomatique. Volume 9 – n° 4/1999, pages 389 à 423

Zeitouni K. (2006). Analyse et extraction de connaissances des bases de donées spatio-temporelle.Human-Computer Interaction.Université de Versailles-Saint Quentin en Yvelines, 2006.

Extraction de Localisations dans les MicroBlogs

Thi-Bich-Ngoc Hoang*,**, Josiane Mothe*

*Université de Toulouse et IRIT, UMR5505 CNRS, France
Prénom.Nom@irit.fr
**University of Economics, the University of Danang, Vietnam

Résumé. La circulation de l'information est de plus en plus rapide. Les applications comme WhatsApp ou Twitter permettent d'échanger des informations sur des évènements de façon quasi instantanée. Il s'agit de ressources précieuses desquelles peuvent être extraites des informations sur des événements (temps, localisation ou entité concernée). Nous nous centrons ici sur l'aspect localisation qui a de nombreuses applications aussi bien dans le cadre d'outils géo-spatialisés que pour des recommandations personnalisées. Dans le contexte de microblogs, les outils développés en traitement du langage naturel ne sont pas suffisants compte tenu de la forme des messages; par exemple les tweets ne sont pas linguistiquement corrects. Par ailleurs, le nombre important de messages à traiter est également un challenge. Dans ce article, nous présentons un modèle pour prédire si un microblog (tweet) contient une localisation ou non et nous montrons que cette prédiction améliore l'efficacité de l'extraction de localisations des tweets.

1 Introduction

De nombreux travaux actuels s'intéressent aux microblogs et à leur exploitation. Par exemple, SanJuan et al. (2012) ont introduit une tâche d'évaluation à CLEF [1] concernant la contextualisation de tweets pour aider à leur compréhension. Dans TREC [2], la tâche vise à proposer des recommandations contextuelles aux utilisateurs (Ounis et al., 2011). La tâche CLEF a évolué récemment pour prendre en compte différents besoins qui peuvent être utiles aux utilisateurs dans le cadre d'événements comme les festivals (Goeuriot et al., 2016; Ermakova et al., 2017; Goeuriot et al., 2018).

Un événement possède trois composants essentiels ((Sundheim, 1996)) : (a) une localisation qui indique *où* l'événement se passe ; (b) une temporalité qui indique *quand* l'événement se passe ; (c) une information sur l'entité concernée qui indique *sur quoi* ou *sur qui* porte l'événement.

Cet article, dont une version étendue a été publiée dans le journal IPM (Hoang et Mothe, 2018), est centré sur la dimension de localisation qui est vitale pour les applications géo-spatiales (Munro, 2011). Par exemple, l'une des premières informations transmises aux sys-

1. CLEF est un programme de recherche européen centré sur l'évaluation de tâches de recherche d'information `http://www.clef-initiative.eu/`
2. TREC est un autre programme d'évaluation en RI patronné par le NIST USA `trec.nist.gov`

tèmes de secours en cas de catastrophe est l'endroit où la catastrophe s'est produite (Lingad et al., 2013). Les informations de localisation sont parfois présentes dans les microblogs quasi simultanément aux événements eux-mêmes. Par exemple, les utilisateurs de messagerie instantanée comme Twitter sont susceptibles de transmettre des mises à jours très régulières et les utilisateurs eux-mêmes trouvent la localisation de l'information très importante (Vieweg et al., 2010).

Au cours des dernières années, plusieurs systèmes de reconnaissance d'entités nommées (EN) traitent du problème de l'extraction de localisations spécifiées dans les documents (Bontcheva et al., 2013; Etzioni et al., 2005) ; mais ces systèmes ne fonctionnent pas bien sur des textes informels. En effet, les analyseurs de texte utilisent des fonctionnalités telles que le type de mot, les lettres en majuscules et le contexte agrégé, qui ne sont souvent pas exacts dans des microblogs bruités, non structurés et courts (Huang et al., 2015).

L'identification ou extraction de localisations repose principalement sur : 1) la recherche et comparaison du texte avec les noms d'entités dans des répertoires, 2) l'utilisation de la structure et du contexte du texte. Le premier type de méthodes est simple mais limite l'extraction à une liste prédéfinie de noms, alors que le second est capable de reconnaître les noms même s'ils ne figurent pas sur la liste (Huang et al., 2015).

Stanford NER est un système d'extraction d'EN qui s'appuie sur une méthode d'apprentissage automatique (Toutanova et al., 2003) ; il fonctionne bien sur les nouvelles mais mal sur les microblogs. Récemment, Bontcheva et al. (2013) ont adapté leur système GATE d'extraction d'EN pour les tweets. Ils ont également adapté l'analyseur de Stanford pour les collections de tweets. Leurs propositions ont permis d'augmenter la performance en termes de mesure F de 60 % à 80 %, principalement pour l'extraction de personnes, d'organisations et du temps, mais pas en ce qui concerne les lieux. Ritter et al. (2011) a abordé le problème de l'extraction d'EN pour les microblogs en utilisant un modèle probabiliste et une base de données ouverte (Freebase) comme source d'apprentissage. Leurs expériences montrent que leur approche surpasse les outils existants sur les tweets pour les entités de localisation avec une mesure F de 77%. Alors que Gate NLP est plus efficace en termes de rappel, Stanford NER et Ritter sont plus efficaces en termes de précision (Bontcheva et al., 2013; Hoang et al., 2017). Dans cet article, nous introduisons une méthode qui combine ces outils pour cibler des applications orientées vers le rappel ou orientées vers la précision. Nous proposons également une méthode prédisant si les microblogs contiennent une localisation. Nous proposons également de filtrer les localisations extraites à l'aide de DBpedia pour augmenter la précision des outils.

L'article est organisé comme suit : dans la section 2, nous présentons les résultats obtenus par une méthode de fusion de différents outils d'extraction. Dans la section 3, nous présentons nos propositions pour la prédiction de la présence d'une localisation dans un tweet et montrons que cette prédiction améliore significativement l'efficacité de l'extraction. Finalement, nous concluons cet article en Section 4. Ces travaux ont également donné lieu à publication dans la revue Information Processing Management (Hoang et Mothe, 2018).

2 Combinaison des méthodes d'extraction de la littérature

Plusieurs méthodes se sont intéressées à l'extraction de localisation dans des textes comme Ritter tool (Ritter et al., 2011), Gate NLP framework (Bontcheva et al., 2013) et Stanford NER (Finkel et al., 2005). Nous avons étudié la combinaison de ces trois méthodes : nous avons

extrait les localisations identifiées par chacun des trois outils et les avons fusionnés. Nous avons également considéré leur filtrage après extraction en nous appuyant sur la base DBpedia (`http://dbpedia.org/snorql/`). Pour les évaluations, nous avons utilisé deux collec-

TAB. 1 – *Description of data used for training and testing.*

	Ritter's dataset	MSM2013 dataset
Training	142 TCL, 1420 TNL	331 TCL, 1655 TLN
Testing	71 TCL, 761 TNL	165 TCL, 664 TNL

tions standards : la collection Ritter (Ritter et al., 2011) et la collection MSM2013 (Cano Basave et al., 2013). La collection Ritter contient 2 394 tweets dont 213 (soit $8,8\%$) avec localisation et 2 181 sans. MSM2013 contient 2 815 tweets dont 496 (soit $17,6\%$) avec localisation et 2 319 sans. Les localisations dans ces collections sont annotées ; elles contiennent un ensemble d'apprentissage et un ensemble de test. La Table 1 présente les caractéristiques de ces collections.

Les résultats pour le rappel, la précision et la mesure F sont présentés dans la table 2 (nous n'avons pas testé l'ensemble des combinaisons possibles et laissons cela pour des travaux futurs). Nous avons utilisé le T-test avec comme fonction témoin l'extraction par l'outil Ritter (première ligne de la table 2).

TAB. 2 – *Résultats de la combinaison des modèles Ritter, Gate et Stanford et du filtrage avec DBPedia. Rappel - R(%), Précision - P(%), Mesure F - F(%). (*) indique un résultat statistiquement significatif par rapport à la fonction témoin.*

	Données Ritter			**Données MSM2013**		
	R(%)	**P(%)**	**F(%)**	**R(%)**	**P(%)**	**F(%)**
Ritter (témoin)	71	82	77	61	80	69
Ritter +Stanford+DBp	77*	79	**78**	72*	79	**75***
Ritter+Gate+DBp	78*	71	74	74*	77	**75***
Ritter+Stanford	80*	64	72	**78***	72	**75***
Ritter+Gate	**82***	56	66	**78***	64	71
Ritter+DBp	45	**97***	62	48	**88***	62

Comme le montre la table 2, la combinaison de l'outil Ritter et de Stanford-NER filtré par DBpedia donne la meilleure mesure F. Pour MSM2013, la mesure F augmente de 69 % à 75 %. Lorsque l'on s'intéresse à une forte précision, c'est la combinaison de Ritter avec le filtrage BDPedia qui est la plus efficace (dernière ligne) alors que pour le rappel, il s'agit de la combinaison de Ritter avec Gate (avant dernière ligne).

3 Prédiction de la présence de localisation

Prévoir qu'un tweet contient un nom de lieu n'est pas simple car les tweets sont généralement écrits dans un langage pseudo-naturel et peuvent ne pas correspondre à des phrases grammaticalement correctes. Les outils usuels de traitement automatique de la langue rencontrent alors des difficultés. Nous proposons un ensemble de caractéristiques pour représenter

les tweets et nous étudions la pertinence de cette représentation dans un modèle prédictif basé sur un apprentissage automatique. La Table 3 présente ces caractéristiques ; (Hoang et Mothe, 2018) présente plus de détails.

TAB. 3 – *Caractéristiques pour prédire la présence de localisation dans un tweet.*

Nom	**Description**	**Exemple**
1. Geography gazetteer	Contient un terme qui apparaît dans Gate geography gazetteer	Today I got a new job ; tomorow jI will be staying in **Dublin**
2. Prep+PP	Contient une préposition juste avant un nom propre	- RT @RMBWilliams : Here **in Gainesville** ! - Greek Festival **at St Johns** before ASPEN !
3. PP	Nombre de nom propres	going **to** alderwood :). # PP : 1
4.Prep	Contient une des 7 prépositions anglaises de lieu ou de mouvement : *at, in, on, from, to, toward, towards*	- Feeling really good after great week **in** our London offices - @Strigy got mine **in** bbt aintree today
5. Place+PP	Contient un terme spécifiant un lieu (*town, city, state, region, country*) juste avant ou après un nom propre	- The football fever : Ohio head coach Frank Solich says Ohio **state** knows they have a special team and season underway
6. Time	Contient une expression de temps (*today, tomorrow, weekend, tonight...*)	- Headed to da gump **today** alabama here I come - Come check out Costa Lounge **tonight** !
7. DefArt+PP	Contient un article défini juste avant un nom propre	- Beautiful day ! Nice to get away from **the Florida** heat
8. Htag	Contient un hashtag	**#Brazil**
9. Adj	Nombre d'adjectifs	- **Bad** time for leicester fans. # Adj :1
10. Verb	Nombre de verbes	- Willingham **took** a turn. # Verb : 2

Les caractéristiques "PP","Adj","Verb" sont des entiers alors que les autres sont des valeurs booléennes.

Nous avons utilisé les mêmes collections que dans la section 2 pour l'apprentissage de notre modèle pour la détection de la présence d'une localisation. Nous avons utilisé différents algorithmes d'apprentissage : Naive Bayes (NB), Support Vector Machine (SMO) et Random Forest (RF) avec une validation croisée. Lors de l'apprentissage, il est possible d'optimiser différents critères ; nous avons choisi la précision et les vrais positifs comme critères d'optimisation. Nous ne présentons pas ici le détail des résultats mais nous obtenons une mesure F d'environ $0,65$ et une précision (accuracy) de $0,80$ à $0,92$ en fonction des cas. RF permet d'obtenir les meilleurs résultats.

TAB. 4 – *Efficacité de l'algorithme Ritter pour les collections Ritter et MSM2013 en termes de Rappel, Précision, mesure F, sur l'ensemble de tests tel que décrit dans le Tableau 1 et les tweets que nous prédisons comme contenant une localisation (avec RF). Les valeurs statistiquement différentes de la fonction témoin sont indiquées par une étoile (*). Le nombre entre parenthèses est le meilleur résultat des trois tirages.*

	Ritter dataset			**MSM2013 dataset**		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
Baseline	69	85	75	60	80	69
Accuracy	45(51)	**96*(98)**	61(66)	37(40)	**89*(92)**	52(55)
TP	45(51)	**96*(98)**	61(66)	37(40)	**89*(92)**	52(55)

Le modèle appris permet donc de prédire si un nouveau tweet contient une localisation ou non. C'est sur ces seuls tweets que nous extrayons ensuite les localisations. La table 4 présente les résultats que nous avons obtenus lors de l'extraction des emplacements des tweets prédits comme contenant une localisation. Un seul tirage apprentissage/test ne permet pas de conclure sur les résultats, nous avons donc utilisé trois tirages et rapporté les valeurs moyennes. Le nombre entre parenthèses est le meilleur résultat des trois tirages.

4 Conclusion

Nous avons proposé une approche pour l'extraction de localisations et un modèle pour prédire la présence de localisations dans les tweets. Notre approche d'extraction de localisations repose sur la fusion de méthodes d'extraction existantes et améliore de manière significative les performances lorsque nous visons soit des applications axées sur le rappel, soit au contraire sur la précision. Nous avons montré que : (1) la fusion des localisations extraites par les outils Ritter et Stanford puis filtrées par DBpedia augmente la mesure F. (2) la fusion des localisations extraites par Ritter et Gate améliore considérablement le rappel alors que l'utilisation de DBpedia pour filtrer les entités de localisation reconnues par Ritter augmente considérablement la précision.

Nous avons également fait l'hypothèse que nous pourrions augmenter la précision si nous pouvions prédire la présence de localisations dans les tweets. Nous avons donc introduit une méthode pour prédire si un tweet contient une localisation ou non. Nous avons défini de nouvelles caractéristiques pour représenter les tweets et évalué les paramètres d'apprentissage automatique. Les résultats montrent que : (3) Random Forest et Naive Bayes sont les meilleures solutions d'apprentissage pour ce problème(4) la modification des critères d'optimisation (précision ou vrai positif) ne change pas significativement la mesure F alors que ce changement d'optimisation a un vrai impact sur le taux de vrais positifs et de faux positifs. (5) pour l'extraction de localisation, nous avons amélioré la précision en nous concentrant uniquement sur les tweets prédits comme contenant une localisation par notre méthode. (6) le compromis entre l'augmentation de la précision et la diminution du rappel restent à étudier.

Dans les travaux futurs, nous souhaitons utiliser les méthodes d'encapsulation de mots (word embedding) pour la représentation afin d'étudier l'impact à la fois pour la prédiction de la présence de localisations et pour leur extraction dans les microblogs.

Références

Bontcheva, K., L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, et N. Aswani (2013). TwitIE : An open-source information extraction pipeline for microblog text. In *RANLP*, pp. 83–90.

Cano Basave, A. E., A. Varga, M. Rowe, M. Stankovic, et A.-S. Dadzie (2013). Making sense of microposts (# msm2013) concept extraction challenge.

Ermakova, L., L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, et E. SanJuan (2017). CLEF 2017 Microblog Cultural Contextualization Lab Overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 304–314.

Etzioni, O., M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, et A. Yates (2005). Unsupervised named-entity extraction from the web : An experimental study. *Artificial intelligence 165*(1), 91–134.

Finkel, J. R., T. Grenager, et C. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the 43rd annual meeting on association for computational linguistics*, pp. 363–370. ACL.

Goeuriot, L., G. Linares, J. Mothe, P. Mulhem, et E. SanJuan (2018). Building Evaluation datasets for Cultural Microblog Retrieval. In *Language Resources and Evaluation Conference, LREC'18*.

Goeuriot, L., J. Mothe, P. Mulhem, F. Murtagh, et E. SanJuan (2016). Overview of the CLEF 2016 Cultural micro-blog Contextualization Workshop. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 371–378. Springer.

Hoang, T. B. N., V. Moriceau, et J. Mothe (2017). Predicting locations in tweets. In *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Hoang, T. B. N. et J. Mothe (2018). Location extraction from tweets. *Information Processing & Management 54*(2), 129–144.

Huang, Y., Z. Liu, et P. Nguyen (2015). Location-based event search in social texts. In *International Conference on Computing, Networking and Communications (ICNC)*, pp. 668–672. IEEE.

Lingad, J., S. Karimi, et J. Yin (2013). Location extraction from disaster-related microblogs. In *Proc. of the 22nd international conference on world wide web*, pp. 1017–1020. ACM.

Munro, R. (2011). Subword and spatiotemporal models for identifying actionable information in haitian kreyol. In *Proc. of the conference on computational natural language learning*, pp. 68–77. ACL.

Ounis, I., C. Macdonald, J. Lin, et I. Soboroff (2011). Overview of the trec-2011 microblog track. In *Proc. of the 20th Text REtrieval Conference*, Volume 32.

Ritter, A., S. Clark, O. Etzioni, et al. (2011). Named entity recognition in tweets : an experimental study. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 1524–1534. ACL.

SanJuan, E., V. Moriceau, X. Tannier, P. Bellot, et J. Mothe (2012). Overview of the INEX 2012 Tweet Contextualization Track. *Initiative for XML Retrieval INEX*, 148.

Sundheim, B. M. (1996). Overview of results of the MUC-6 evaluation. In *Proc. of a workshop on held at Vienna, Virginia*, pp. 423–442. A.

Toutanova, K., D. Klein, C. D. Manning, et Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the Conference of the ACL on Human Language Technology*, pp. 173–180. ACL.

Vieweg, S., A. L. Hughes, K. Starbird, et L. Palen (2010). Microblogging during two natural hazards events : what twitter may contribute to situational awareness. In *Proc. of the SIGCHI*, pp. 1079–1088. ACM.

Summary

Applications such as WhatsApp or Twitter allows anyone to exchange information about events almost instantly. There are therefore valuable resources from which information about events (time, location or entity) can be extracted. In this paper, we focus on localization, which has many applications in the context of geo-spatialized tools or for personalized recommendations. In the context of microblogs, tools developed in natural language processing are not sufficient; for example, tweets are generally not linguistically correct. Moreover, the large number of messages to be processed is also a challenge to solve. in this paper, we present a model for predicting whether a short text contains a localization or not and we show that this prediction improves localization extraction.

Fouille de Motifs Graduels Fermés Fréquents Sous Contrainte de la Temporalité

Jerry Lonlac[*,***] Benjamin Negrevergne[**] Yannick Miras[****] Aude Beauger[***]
Engelbert Mephu Nguifo[*]

[*]CNRS, UMR 6158, LIMOS, Université Clermont Auvergne, F-63173 Aubière, France
[**]LAMSADE, CNRS UMR 7243, Université Paris Dauphine
{benjamin.negrevergne}@dauphine.fr
[***]CNRS, UMR 6042, GEOLAB, Université Clermont Auvergne, F-63000 Clermont-Ferrand
{jerry.lonlac_konlac, engelbert.mephu_nguifo, aude.beauger}@uca.fr
[****]HNHP, Muséum National d'Histoire Naturelle, UMR 7194
{yannick.miras}@mnhn.fr

Résumé. Le problème de fouille de motifs graduels consiste à découvrir des co-variations simultanées fréquentes d'attributs numériques dans une base de données. Plusieurs algorithmes d'extraction automatique de tels motifs ont été proposés dans la littérature. La principale différence entre ces algorithmes réside dans la sémantique de variation considérée. Dans certains domaines d'application, on trouve des bases de données dont les objets sont munis d'une relation d'ordre temporel. Les approches de l'état de l'art ne prennent pas en compte cette contrainte de temporalité dans le processus de fouille. Dans ce papier, nous proposons une approche pour l'extraction de motifs graduels fréquents dont l'ordre des objets supportant les motifs correspond à l'ordre temporel. Cette approche réduit le nombre de motifs générés quand les objets suivent une relation d'ordre temporel. Les résultats expérimentaux obtenus sur les données paléoécologiques montrent l'efficacité de notre approche, l'interprétation de ces résultats apporte de nouvelles connaissances aux experts paléoécologiques et montre ainsi l'intérêt de ce type de motifs.

1 Introduction

Les motifs graduels qui capturent les corrélations d'ordre de la forme "plus/moins X, plus/moins Y" jouent un rôle important dans plusieurs applications du monde réel où les données numériques doivent être gérées. Les algorithmes de fouille de données sont le plus souvent utilisés pour extraire automatiquement de tels motifs (Di-Jorio et al., 2008; Di-Jorio et al., 2009; Laurent et al., 2009; Do et al., 2015, 2010).

Dans Di-Jorio et al. (2008), étant donné un motif graduel, les auteurs proposent une heuristique permettant d'éliminer les objets qui empêchent le nombre maximum d'objets de la base de données d'être ordonné (objets ayant les plus grands ensembles de conflits). L'ensemble de conflits d'un objet étant constitué de tous les objets qui sont en conflit avec celui-ci. Dans

Di-Jorio et al. (2009), une méthode exacte fondée sur l'utilisation de structures binaires est proposée. Dans cette méthode appelée GRITE pour "GRadual ITemset Extraction", les données sont représentées à travers un graphe dans lequel les nœuds représentent les objets de la base de données, et les arcs expriment les relations de précédences dérivées à partir des attributs considérés.

La plupart des algorithmes proposés pour la fouille de motifs graduels se heurtent au problème de gestion de la quantité très élevée de motifs extraits. Sur certaines données, le nombre de motifs graduels fréquents peut être important, rendant leur interprétation par l'expert quasiment impossible. Une façon de réduire le nombre de motifs extraits est d'utiliser des représentations condensées de motifs. En effet, à partir d'un ensemble de motifs spécifiques, comme les motifs graduels fermés (Ayouni et al., 2010), il est possible de régénérer l'ensemble de tous les motifs graduels. De plus, les motifs fermés évitent d'avoir des informations redondantes. Dans cet ordre d'idée, Ayouni et al. (2010) propose une approche pour réduire le nombre de motifs extraits. Cette approche est juste un post-traitement de l'approche de Di-Jorio et al. (2009).

Do et al. (2015) propose un algorithme appelé GLCM qui réduit le nombre de motifs au cours du processus de fouille. GLCM est fondé sur l'extension de l'idée développée dans l'algorithme LCM (Uno et al., 2004) et permet d'extraire efficacement les itemsets graduels fermés fréquents sur des grandes bases de données du monde réel avec une complexité en temps linéaire en nombre de motifs et une complexité mémoire constante par rapport au nombre de motifs.

Dans Négrevergne et al. (2014), la fermeture d'un motif graduel s est définie comme étant l'intersection des transactions contenant s, à laquelle sont supprimés tous les attributs avec des variations descendantes placés avant le premier attribut avec une variation ascendante dans le motif. Cette définition n'est plus valide dans le contexte de la temporalité car les motifs graduels qui respectent la contrainte de temporalité ne sont pas symétriques. Ainsi, dans notre approche, la fermeture est vue comme une autre contrainte et est efficacement combinée avec la contrainte temporelle pour extraire les motifs graduels fermés fréquents dont les séquences d'objets respectent l'ordre temporel.

D'autre part, les algorithmes proposés pour l'extraction de motifs graduels effectuent plusieurs opérations très coûteuses en temps au cours du processus d'extraction, ils ne permettent pas de traiter efficacement les bases de données réelles très volumineuses (contenant beaucoup d'attributs). Pour résoudre ce problème, les auteurs de Laurent et al. (2010) utilisent le multi-threading pour exploiter au mieux les multiples cœurs présents dans la plupart des ordinateurs pour améliorer les performances de l'algorithme GRITE. Dans le même ordre d'idée, une version parallèle de l'algorithme GLCM appelée PGLCM est proposée dans Do et al. (2015, 2010).

Dans Négrevergne et al. (2014), ParaMiner, un algorithme générique et parallèle pour l'extraction de motifs graduels fermés fréquents est proposé. ParaMiner est actuellement l'algorithme le plus efficace pour extraire les motifs graduels fermés fréquents des grandes bases de données numériques.

D'autre part, les algorithmes proposés dans la littérature pour la fouille de motifs graduels ne supposent aucune contrainte temporelle sur les données. Cependant, il existe des domaines d'application où les objets ont une signification temporelle, c'est le cas d'une base de données paléoécologiques. Etant donné cette contrainte, on peut uniquement s'intéresser aux motifs dont l'ordre des objets concordants respecte la contrainte temporelle, les autres motifs n'étant

pas pertinents. De ce fait, les algorithmes proposés dans la littérature ne sont pas adaptés pour l'extraction des motifs graduels sous contrainte temporelle sur les transactions.

Nous proposons ainsi dans ce papier, une approche pour extraire les motifs graduels fermés fréquents où l'ordre des objets supportant les motifs correspond à l'ordre temporel. Cette approche exploite l'encodage proposé dans l'algorithme `ParaMiner` et utilise l'algorithme générique qui calcule la fermeture d'un motif en l'augmentant avec les éléments de l'intersection des transactions de son ensemble support. En effet, l'approche proposée dans l'algorithme `ParaMiner` pour le calcul de la fermeture d'un motif graduel dépend du fait que le problème de fouille de motifs graduels soit symétrique. Mais cette symétrie n'est plus satisfaite en tenant compte de la contrainte temporelle, c'est la raison pour laquelle, nous adoptons l'algorithme générique donné pour le calcul de la fermeture d'un motif graduel. Notre approche permet de découvrir de nouvelles formes de motifs lorsqu'elle est appliquée sur les données numériques avec contrainte temporelle sur les transactions. De plus, lorsqu'elle est intégrée à `ParaMiner`, notre approche permet de réduire le nombre de motifs extraits en éliminant les motifs non pertinents au cours du processus de fouille.

Le papier est organisé comme suit : nous présentons en section 2 un contexte de données numériques avec contrainte temporelle. Après avoir introduit la notion de motifs graduels, notre approche d'extraction de motifs graduels fermés fréquents sous contrainte temporelle est décrite en section 4. Avant de conclure, les résultats expérimentaux menés sur des données paléoécologiques et leur interprétation sont présentés. Nous comparons aussi en termes de nombre de motifs extraits, l'algorithme original `ParaMiner` avec celui intégrant la contrainte temporelle.

2 Domaine d'application : la paléoécologie

Les recherches paléoécologiques permettent de reconstruire au cours du temps les dynamiques écologiques et l'évolution de la biodiversité (par exemple l'évolution de la végétation ou du fonctionnement d'un écosystème lacustre) sous l'influence des variations climatiques et des activités humaines (par exemple agriculture et pastoralisme) (Smol et al., 2001). La reconstruction de ces trajectoires écologiques sur 7 millénaires pour le lac d'Aydat, situé dans la Chaîne des Puys en région Auvergne Rhône-Alpes et menacé d'eutrophisation, permet de mieux connaître son état écologique actuel et de concourir à l'établissement de gouvernances viables (Miras et al., 2015). Cette recherche est fondée, sur l'abondance de différents indicateurs paléoécologiques (grains de pollen et spores de végétaux ; micro-fossiles non polliniques : différentes formes de résistance du phytoplancton et du zooplancton ; diatomées) conservés dans l'enregistrement sédimentaire lacustre. Toutes ces données paléoécologiques sont ensuite stockées dans des bases de données numériques.

Les données paléoécologiques sont constituées d'un ensemble d'attributs à valeurs numériques correspondant à la quantité de chaque indicateur paléoécologique contenu dans un enregistrement sédimentaire prélevé, par des opérations de carottage, au sein d'un écosystème lacustre. La séquence sédimentaire obtenue est ensuite datée, échantillonnée, et pour chaque échantillon, à une profondeur donnée, une date est calculée. L'abondance de chaque indicateur est ensuite relevée pour chaque échantillon. Les objets de cette base de données correspondent aux différentes dates obtenues sur l'enregistrement sédimentaire, et les colonnes aux différents indicateurs paléoécologiques relevés.

Plus formellement, soient $\mathcal{D}$ les dates des différents échantillons de la séquence sédimentaire, $\mathcal{I}$ les différents indicateurs paléoécologiques relevés à ces dates $\mathcal{D}$, alors le tableau de données paléoécologiques est défini par $\Delta = \mathcal{D} \times \mathcal{I}$. Les notations suivantes caractérisent les données contenues dans Δ.

Soit $d \in \mathcal{D}$ et $i \in \mathcal{I}$:

— $\Delta(d, i)$ indique le nombre d'instances de l'indicateur i présents à la date d ;

— $\Delta(d, i) = 0$ indique l'absence de l'indicateur i à la date d ;

— $\Delta(\mathcal{D}, i)$ montre l'évolution temporelle de l'indicateur i ;

— $\Delta(d, \mathcal{I})$ caractérise l'état des conditions paléoécologiques à la date d.

Une particularité des données paléoécologiques contenues dans Δ est qu'elles sont évolutives. De plus, le tableau Δ comportent généralement peu de lignes (les dates des échantillons) au regard du nombre de colonnes (nombre d'indicateurs paléoécologiques) et est très peu dense (il contient un grand nombre de valeurs nulles).

Le tableau 1 est un extrait de données paléoécologiques, ces colonnes sont étiquetées par les noms scientifiques des différents indicateurs. Pour des raisons de simplicité, afin d'illustrer notre approche dans la suite, nous présentons ici une base restreinte à 7 indicateurs et 7 dates. Nous considérerons le tableau 1 que nous désignons par Δ, comme une base de transactions contenant des attributs à valeurs numériques. Ainsi, les 7 indicateurs paléoécologiques correspondront aux différents attributs de notre base de données et les 7 dates identifieront les différents objets.

	Poaceae	**Secale.t**	**Rumex.ace**	**Equisetum**	**Plantago.l**	**Filipendula.v**	**Coprofilous.f**
d_1	84	61	7	0	1	2	0
d_2	116	36	4	1	11	2	31
$d3$	90	52	2	3	5	2	13
d_4	124	34	1	5	12	1	36
d_5	102	49	0	6	7	0	17
d_6	135	17	0	1	18	0	62
d_7	106	40	3	1	9	0	18

TAB. 1 – *Exemple de base de données paléoécologiques* Δ.

Il est alors question pour les experts de la recherche paléoécologique, de les analyser afin d'en extraire des connaissances spécifiques telles que la mise au jour des groupements de coévolution d'indicateurs multi-variés paléoécologiques (par exemple des groupements constitués de grains de pollen, de micro-fossiles non polliniques et de diatomées) nécessaires à la compréhension de l'évolution de la biodiversité et du fonctionnement d'un écosystème au cours du temps. Toutes ces informations implicites sont généralement recherchées par les experts en paléoenvironnement en utilisant les méthodes d'analyses statistiques classiques. Celles-ci reposent le plus souvent sur un simple tracé d'un graphique contenant des courbes d'évolution des différents indicateurs paléoécologiques à partir des données paléoécologiques et sur une comparaison empirique de ces courbes afin de relever des groupes de courbes qui évoluent à des périodes identiques.

Il apparaît évident qu'il est très fastidieux pour les experts de la paléoécologie d'identifier de manière empirique les différentes coévolutions de ces indicateurs. De plus le risque est d'exclure un traitement exhaustif des données et de laisser échapper éventuellement la pépite d'information pertinente : les groupes à faible coévolution mais à fort impact sur le changement de la biodiversité. Dans ce contexte, les motifs graduels s'avèrent adaptés pour résoudre le

problème d'extraction automatique des groupements de coévolution d'indicateurs multi-variés paléoécologiques.

Dans ce papier, nous montrons que, sans la prise en compte de la contrainte de temporalité entre les objets de la base de données, les sémantiques proposées dans les algorithmes de fouille de motifs graduels de la littérature ne sont pas adaptés au contexte de la temporalité (exemple des bases de données paléoécologiques). Pour ces données, cette contrainte permet d'éviter au cours du processus de fouille, de générer des motifs graduels qui ne correspondent pas aux coévolutions, ce qui permet des gains de temps de calcul et de mémoire consommée considérable. L'approche proposée apporte de l'information supplémentaire par rapport aux techniques classiques utilisées jusqu'à présent pour l'analyse de données paléoéologiques, qui sont essentiellement des méthodes statistiques et de classification.

3 Les motifs graduels

Dans cette section, nous rappelons la notion d'itemsets graduels aussi appelés motifs graduels en utilisant les définitions proposées dans Di-Jorio et al. (2009) et l'illustrons sur notre jeu de données en utilisant l'algorithme `Paraminer`.

3.1 définitions préliminaires

Considérons une base de données Δ contenant un ensemble d'objets $\mathcal{D} = \{ d_1, ..., d_n\}$ décrit par un ensemble d'attributs $\mathcal{I} = \{ i_1, ..., i_m\}$ à valeurs numériques. $\forall d \in \mathcal{D}$, $d[i]$ désigne la valeur de l'objet d sur l'attribut i.

Les itemsets graduels extraits de Δ sont de la forme "plus/moins i_1, . . ., plus/moins i_k" $(k \leq m)$. Ces motifs sont définis sur un sous-ensemble de $\mathcal{D}$ où les éléments sont associés à un ordre croissant ou décroissant. Dans le contexte de fouille d'itemsets graduels, nous considérons les valeurs d'attributs (aussi appelés items) et les comparaisons sont faites entre les objets.

Chaque attribut sera par la suite considéré deux fois : une fois pour indiquer son augmentation, et une fois pour indiquer sa diminution, en utilisant les opérateurs "$\leq$" et "$\geq$". Ceci conduit à considérer de nouveaux types d'items, appelés items graduels.

Définition 1 (Item graduel) *Soit Δ une base de données définie sur un ensemble d'attributs à valeurs numériques $\mathcal{I}$, un item graduel est défini sous la forme i^*, où i est un attribut de $\mathcal{I}$ et $* \in \{\geq, \leq\}$ un opérateur de comparaison.*

En considérant la base de données du tableau 1, $Poaceae^{\geq}$ (respectivement $Poaceae^{\leq}$) est un item graduel indiquant que la valeur de l'attribut *Poaceae* augmente (respectivement diminue).

Un itemset graduel est défini comme suit :

Définition 2 (Itemset graduel) *Un itemset graduel $s = (i_1^{*_1}, ..., i_k^{*_k})$ est un ensemble non vide d'items graduels.*

Par exemple, $\{Poaceae^{\geq}, Rumex^{\leq}\}$ est un itemset graduel indiquant que *plus les valeurs de l'attribut Poaceae augmentent, plus les valeurs de l'attribut Rumex diminuent*

Un itemset graduel impose une contrainte de variation sur plusieurs attributs simultanément. La longueur d'un itemset graduel est égale au nombre d'items graduels qu'il contient. Un k-itemset graduel est un itemset graduel contenant k items graduels. $(Poaceae^{\geq}, Rumex^{\leq})$ est un 2-itemset graduel.

3.2 Découverte d'itemsets graduels fréquents

Le calcul du support d'un motif graduel dans une base de données revient à mesurer à quel point le motif est présent dans cette base de données. Plusieurs définitions de support ont été proposées dans la littérature (Hüllermeier, 2002; Berzal et al., 2007; Di-Jorio et al., 2009), montrant que les itemsets graduels peuvent suivre différentes sémantiques de variation. Le choix entre ces sémantiques dépend le plus souvent de l'application considérée. La définition de support proposée dans Di-Jorio et al. (2009) est fondée sur la longueur de la plus longue séquence de transactions qui peut être consécutivement ordonnée par rapport au motif graduel. L'intérêt de cette définition est qu'une telle séquence de transactions peut ensuite être facilement présentée à l'analyste, permettant d'isoler et de réordonner une partie des données et de l'étiqueter avec une description en termes de co-variations.

Dans ce papier, nous considérons la sémantique de variation proposée dans Di-Jorio et al. (2009) qui est celle implémentée par l'algorithme `Paraminer` et nous l'adaptons pour prendre en compte la contrainte temporelle au cours du processus de fouille. Cela nous permet de découvrir de nouveaux types d'itemsets graduels appelés ici itemsets graduels temporels qui sont des itemsets graduels dont les plus longues séquences de transactions correspondantes respectent l'ordre temporel. Ces types d'itemsets graduels sont particulièrement intéressants dans le domaine paléoécologique (décrit en section 2) où les experts recherchent à partir de leurs données numériques paléoécologiques les motifs qui capturent des co-évolutions simultanément fréquentes d'attributs. De plus, lorsqu'elle est intégrée à `Paraminer`, notre approche permet de réduire drastiquement la quantité de motifs générée au cours du processus de fouille et de bénéficier de la réduction du temps d'exécution et de la mémoire.

Dans la suite, nous rappelons la définition de la notion d'ordre induit par un itemset graduel proposée dans Do et al. (2015) et qui est nécessaire à la compréhension de la définition du support d'itemset graduel proposée dans Di-Jorio et al. (2009).

Définition 3 (l'Itemset graduel induit l'ordre) *Soit $s = (i_1^{*_1}, ..., i_k^{*_k})$ un itemset graduel, et Δ une base de données numériques. Deux objets d et d' de Δ peuvent être ordonnés par rapport à s si toutes les valeurs des items correspondants de l'itemset graduel peuvent être ordonnées en respectant toutes les variations d'items graduels de s : pour tout $l \in [1, k]$, $d[i_l] \leq d'[i_l]$ si $*_l = '\geq'$ et $d[i_l] \geq d'[i_l]$ si $*_l = '\leq'$. Le fait que d précède d' dans l'ordre induit par s est dénoté $d \lhd_s d'$.*

En considérant la base de données du tableau 1, d_1 et d_2 peuvent être ordonnés par rapport à $s_1 = (Poaceae^{\geq}, Rumex^{\leq})$ comme $d_1[Poaceae] \leq d_2[Poaceae]$ et $d_1[Rumex] \geq d_2[Rumex]$: nous avons $d_1 \lhd_{s_1} d_2$.

Cet ordre est seulement un ordre partiel. Par exemple, considérons les objets d_2 et d_3 du tableau 1 : ils ne peuvent être ordonnés par rapport à s_1. En effet, le motif s_1 n'est pas pertinent pour expliquer les variations entre d_2 et d_3, et plus généralement toute paire d'objets qui ne peut être ordonnée. A l'inverse, un motif graduel est pertinent pour expliquer les variations

apparaissant entre objets qu'il peut ordonner. La définition de support que nous considérons dans ce papier va plus loin et focalise sur la taille de la plus longue séquence d'objets qui peut être ordonnée par rapport à l'itemset graduel considéré. L'intuition étant que de tels motifs seront supportés par de longues variations continues dans les données (longues périodes de coévolution entre indicateurs paléoécologiques dans le cas des données paléoécologiques). De telles variations continues étant particulièrement intéressantes à extraire pour une meilleure compréhension des données.

Définition 4 (Support d'un itemset graduel) *Soit Δ une base de données numériques contenant un ensemble d'objets $\{d_1, ..., d_n\}$ et $L = \langle d_{j_1}, ..., d_{j_s} \rangle$ une séquence d'objets de Δ, avec $\forall k \in [1..s], j_k \in [1..n]$ et $\forall k, k' \in [1..s], k \neq k' \Rightarrow j_k \neq j_{k'}$. Soit s un itemset graduel. L respecte s si $\forall k \in [1..s-1]$ nous avons $d_{j_k} \lhd_s d_{j_{k+1}}$. Soit G_s l'ensemble des séquences d'objets qui respectent s. Alors le support de s est défini par $Support(s) = \frac{max_{L \in G_s}(|L|)}{|\Delta|}$, c'est-à-dire la taille de la plus longue liste d'objets qui respectent s.*

En considérant encore la base de données du tableau 1 et le motif $s_2 = (Poaceae^{\geq}, Secale^{\leq}, Equisetum^{\leq}, Plantago^{\geq}, Coprofilous^{\geq})$, l'ensemble de toutes les listes d'objets respectant s_2 est $G_{s_2} = \{\langle d_3, d_7, d_2, d_6\rangle, \langle d_5, d_4, d_6\rangle\}$. La plus longue liste de G_{s_2} est $\{\langle d_3, d_7, d_2, d_6\rangle\}$ et sa taille est égale à 4. Ainsi, $support(s_2) = \frac{4}{7} = 0.57$, indiquant que 57% des objets peuvent être consécutivement ordonnés par rapport à s_2.

Définition 5 (Itemset graduel complémentaire) *Soit $s = (i_1^{*_1}, ..., i_k^{*_k})$ un itemset graduel, et c une fonction telle que $c(\geq) = '\leq'$ et $c(\leq) = '\geq'$. Alors $c(s) = (i_1^{*_1^c}, ..., i_k^{*_k^c})$ est l'itemset graduel complémentaire (symétrique) de s et est défini comme $\forall j \in [1..k], *_j^c = c(*_j)$.*

L'itemset graduel complémentaire de s_1 est dénoté $c(Poaceae^{\geq}, Rumex^{\leq}) = (Poaceae^{\leq}, Rumex^{\geq})$.

Proposition 1 (Di-Jorio et al. (2009)) $Support(s) = Support(c(s))$.

La proposition 1 évite des calculs inutiles, comme la génération de la moitié d'itemsets graduels est suffisante pour déduire automatiquement l'autre moitié. Cela signifie que pour chaque itemset graduel, il existe un itemset graduel symétrique ayant le même support.

Un itemset graduel est dit fréquent si son support est plus grand ou égal à un seuil de support minimum fixé par l'utilisateur.

Le problème de fouille d'itemsets graduels fréquents consiste à trouver l'ensemble complet d'itemsets graduels fréquents dans une base de données Δ contenant des attributs (items) à valeurs numériques, par rapport à un seuil de support minimum appelé *minSupp*.

L'approche décrite ci-dessus permet d'extraire automatiquement les motifs graduels fréquents dans une base de données numériques. Cependant, cette approche ne suppose aucune contrainte temporelle entre objets de la base de données, ce qui ne convient pas pour le cas des bases de données dont les objets suivent une relation d'ordre temporel (exemple de la base de données paléoécologiques donnée par le tableau 1).

Dans le contexte de contrainte temporelle, étant donné une base de données numériques Δ contenant un ensemble d'objets, on peut rechercher tous les itemsets graduels fréquents dont les séquences d'objets correspondantes respectent l'ordre temporel. Par exemple, en considérant la base de données de la table 1, l'itemset graduel s_2 n'est pas intéressant dans le contexte

de la temporalité du au fait que les listes d'objets respectant s_2 que sont $L_1 = \langle d_3, d_7, d_2, d_6 \rangle$ et $L_2 = \langle d_5, d_4, d_6 \rangle$ ne respectent pas l'ordre temporel (d_7 précède d_2 dans L_1 et d_5 précède d_4 dans L_2). L'itemset graduel $s_1 = (Poaceae^{\geq}, Rumex^{\leq})$ est un candidat intéressant comme l'une de ses listes d'objets ($\langle d_1, d_2, d_4, d_6 \rangle$) respecte l'ordre temporel. Sa pertinence dépend dans ce cas uniquement de son support par rapport au seuil de support minimum *minSupp* fixé par l'utilisateur.

Dans ce papier, nous proposons d'intégrer la contrainte temporelle dans l'algorithme `Paraminer` afin d'extraire automatiquement les itemsets graduels fréquents à partir d'une base de données numériques Cette approche exploite l'encodage du problème de fouille d'itemsets graduels proposé dans Négrevergne et al. (2014) et utilise le principe des algorithmes classiques de fouille de motifs pour éviter le problème de la non-symétrie des motifs tout en respectant l'ordre temporel. Notre approche est une approche de fouille de motifs sous contraintes différente de celle proposée dans Négrevergne et al. (2014) qui ne permet pas l'utilisation des algorithmes classiques de fouille de motifs.

L'application de notre approche sur les données paléoécologiques permet d'extraire les itemsets graduels fréquents qui correspondent aux groupements d'indicateurs paléoécologiques multi-variés permettant de modéliser l'évolution des écosystèmes considérés.

4 Fouille d'itemsets graduels fermés fréquents sous contrainte temporelle

Cette section présente notre processus permettant de découvrir à partir d'une base de données numériques, les itemsets graduels fermés fréquents dont les séquences d'objets suivent un ordre temporel.

4.1 Gradualité sous la contrainte de temporalité

La sémantique de temporalité que nous proposons de prendre en compte pour notre contexte diffère de la sémantique proposée dans Négrevergne et al. (2014) sur les points suivants :

— le support d'un 1-itemset graduel n'est plus toujours égal à 100% comme dans le contexte classique (Di-Jorio et al., 2009; Négrevergne et al., 2014). En effet, dans le contexte classique, le support d'un item graduel est toujours égal à 100% comme il est toujours possible d'ordonner tous les objets par colonne. Cela n'est pas valide dans le contexte avec contrainte de temporalité. Par exemple, aucun des 7 items graduels du tableau 1 n'a un support égal à 100% sous la contrainte temporelle.

— les itemsets graduels recherchés ici ne sont pas symétriques comme les itemsets graduels classiques. En effet, la proposition 1 donnée dans Di-Jorio et al. (2009) n'est plus valide dans le contexte de temporalité. La génération de la moitié d'itemsets graduels n'est plus suffisante pour déduire automatiquement l'autre moitié. Il est nécessaire de rechercher tous les itemsets graduels avec leur itemset graduel complémentaire correspondant.

— le calcul de la fermeture d'un itemset graduel est défini d'une manière beaucoup plus simple que celle proposée dans Négrevergne et al. (2014). Cela est du au fait que les itemsets graduels ne sont pas symétriques dans le contexte de temporalité.

Dans la suite, nous donnons une définition formelle du problème de fouille de motifs graduels fréquents sous la contrainte de temporalité. Ensuite, nous l'illustrons à travers un exemple en utilisant la base de données numériques Δ précédente, et nous montrons que les sémantiques de gradualité de la littérature sont inadaptées pour ce problème. Plus précisément, nous souhaitons extraire des co-variations simultanées de valeurs d'attributs d'une base de données numériques dont les objets respectent un ordre temporel.

Définition 6 (Itemset graduel respectant l'ordre temporel) *Soit Δ une base de données numériques contenant un ensemble d'objets $\{d_1, ..., d_n\}$ et s un ietmset graduel. Soit $L_s = \langle d_{l_1}, ..., d_{l_j} \rangle$, la plus longue séquence d'objets de Δ respectant s. s respecte l'ordre temporel des objets de Δ si l'inégalité suivante est satisfaite : $d_{l_1} < d_{l_2} < ... < d_{l_j}$.*

Définition 7 (Itemset graduel frère) *Soit $s = \{i_1^{*_1}, ..., i_k^{*_k}\}$ un itemset graduel. $s' = \{i_1^{'*_1}, ..., i_k^{'*_k}\}$ est un itemset graduel frère de s et noté $Sibling(s)$ si et seulement si $s' \neq s$ et $i'_1 = i_1, ..., i'_k = i_k$.*

Proposition 2 *Soit $s = \{i_1^{*_1}, ..., i_k^{*_k}\}$ un itemset graduel. s admet $2^k - 1$ itemsets graduels frères.*

Il est trivial de voir que le complémentaire d'un itemset graduel (l'itemset graduel symétrique) est l'un de ses frères.

Soit $\mathcal{I} = \{ i_1, ..., i_m\}$ un ensemble d'attributs à valeurs numériques et $\mathcal{D} = \{ d_1, ..., d_n\}$ un ensemble d'objets où chaque objet d_j ($j \in [1, n]$) stocke une valeur numérique pour tout attribut dans $\mathcal{I}$. Pour l'extraction d'itemsets graduels sous contrainte temporelle, nous intégrons une contrainte temporelle sur les variations d'attributs dans l'encodage proposé dans Négrevergne et al. (2014). Cet encodage est modifié comme suit :

$\mathcal{A}$ est l'ensemble des variations d'attributs : $\mathcal{A} = \{ i_1^{\geq}, i_1^{\leq}, ..., i_m^{\geq}, i_m^{\leq}\}$.

Dans la nouvelle base de données Δ', il y a autant de transactions que de paires d'objets $(d_j, d_{j'}), d_j, d_{j'} \in \mathcal{D}$, avec $j, j' \in [1, n]$ et $j < j'$. Nous dénotons par $t_{(d_j, d_{j'})}$ la transaction qui contient la variation pour tout attribut dans $\mathcal{A}$ entre les objets d_j et $d_{j'}$: pour tout attribut $i \in \mathcal{I}$, $i^{\geq} \in t_{(d_j, d_{j'})} \iff d_j[i] \leq d_{j'}[i]$, $i^{\leq} \in t_{(d_j, d_{j'})}$ sinon. Le fait que $j < j'$ permet d'imposer la contrainte temporelle sur les variations d'attributs dans $\mathcal{A}$ et est une petite optimisation comparée à l'encodage proposé dans Négrevergne et al. (2014).

L'encodage correspondant du tableau 1 est donné par le tableau 2. Pour des raisons de simplicité, chaque attribut dans le tableau 1 est identifié par la première lettre de son nom excepté l'attribut `Plantago` qui est identifié par `L`.

Avec cet encodage, comme défini dans Négrevergne et al. (2014), le support d'un motif graduel donné $s = (i_1^{*_1}, ..., i_k^{*_k})$ est la taille du plus long chemin d'objets dans l'ensemble support de s. s est dit fréquent si son support est plus grand ou égal à un seuil de support minimum défini par l'utilisateur. Par exemple, le plus long chemin d'objets dans l'ensemble des paires d'objets
$\{\langle d_1, d_2\rangle, \langle d_1, d_3\rangle, \langle d_1, d_4\rangle, \langle d_2, d_3\rangle, \langle d_2, d_4\rangle, \langle d_3, d_4\rangle\}$ est $\{\langle d_1, d_2\rangle, \langle d_2, d_3\rangle, \langle d_3, d_4\rangle\}$ de taille 3.

4.2 Algorithme

L'algorithme 1 calcule la fermeture unique de tout itemset graduel respectant l'ordre temporel.

Extraction de motifs graduels fréquents sous contrainte

$t_{(d_1,d_2)}$	$\{P^{\geq}, S^{\leq}, R^{\leq}, E^{\geq}, L^{\geq}, F^{\leq}, C^{\geq}\}$
...	...
$t_{(d_1,d_7)}$	$\{P^{\geq}, S^{\leq}, R^{\leq}, E^{\geq}, L^{\geq}, F^{\leq}, C^{\geq}\}$
$t_{(d_2,d_3)}$	$\{P^{\leq}, S^{\geq}, R^{\leq}, E^{\geq}, L^{\leq}, F^{\leq}, C^{\leq}\}$
...	...
$t_{(d_2,d_7)}$	$\{P^{\leq}, S^{\geq}, R^{\leq}, E^{\leq}, E^{\geq}, L^{\leq}, F^{\leq}, C^{\leq}\}$
$t_{(d_3,d_4)}$	$\{P^{\geq}, S^{\leq}, R^{\leq}, E^{\geq}, L^{\geq}, F^{\geq}, C^{\geq}\}$
...	...
$t_{(d_3,d_7)}$	$\{P^{\geq}, S^{\leq}, R^{\geq}, E^{\leq}, L^{\geq}, F^{\leq}, C^{\geq}\}$
$t_{(d_4,d_5)}$	$\{P^{\leq}, S^{\geq}, R^{\leq}, E^{\geq}, L^{\leq}, F^{\geq}, C^{\leq}\}$
...	...
$t_{(d_4,d_7)}$	$\{P^{\leq}, S^{\geq}, R^{\geq}, E^{\leq}, L^{\leq}, F^{\leq}, C^{\leq}\}$
$t_{(d_5,d_6}$	$\{P^{\geq}, S^{\leq}, R^{\leq}, R^{\geq}, E^{\leq}, L^{\geq}, F^{\leq}, C^{\geq}\}$
$t_{(d_5,d_7)}$	$\{P^{\leq}, S^{\leq}, R^{\geq}, E^{\leq}, L^{\geq}, F^{\leq}, C^{\geq}\}$
$t_{(d_6,d_7)}$	$\{P^{\leq}, S^{\geq}, R^{\geq}, E^{\leq}, E^{\geq}, L^{\leq}, F^{\leq}, C^{\leq}\}$

TAB. 2 – Δ' *: Encodage de la base de données* Δ

Algorithme 1 : opérateur de fermeture pour le problème de fouille d'itemsets graduels encodant les variations d'attributs dans le contexte de temporalité

Données : un motif graduel s et une base de données Δ' encodant les variations d'attributs de la base de données Δ
Résultat : la fermeture unique de s.
$q_{max} \leftarrow \cap\Delta'[s]$;
retourner q_{max} ;

Corollaire 1 *L'algorithme 1 est correct et complet pour l'extraction d'itemsets graduels fermés fréquents sous contrainte temporelle.*

***Preuve* 1** *La preuve est évidente, en effet l'algorithme 1 est une simplification de l'algorithme classique de calcul de motifs fermés donné dans Boley et al. (2010), qui a été démontré correct et complet pour le calcul de l'ensemble des motifs fermés.*

Lemme 1 *Soit Δ une base de données numériques et s un itemset graduel fréquent respectant la contrainte d'ordre temporel. Soit s' un itemset graduel frère de s et L_s (respectivement $L_{s'}$), la plus longue séquence d'objets de Δ respectant s (respectivement s'). Nous avons $|L_s \cap L_{s'}| \leq 1$.*

***Preuve* 2** *Triviale en utilisant la définition 4.*

Proposition 3 *Soit Δ une base de données numériques et s un itemset graduel fréquent respectant la contrainte d'ordre temporel extrait de Δ par rapport à un seuil de support minimum $minSupp$. Si $minSupp > \frac{|\Delta|}{2}$, alors tous ses itemsets graduels frères ne sont pas fréquents.*

***Preuve* 3** *Soit s un itemset graduel fréquent respectant la contrainte d'ordre temporel, extrait de Δ en respectant un seuil de support minimum $minSupp$. Comme $minSupp > \frac{|\Delta|}{2}$ et s est fréquent, alors $Support(s) > \frac{|\Delta|}{2}$. Soit s' un itemset graduel frère de s, en utilisant le lemme 1, nous obtenons $Support(s') < minSupp$, et donc s' infréquent.*

Proposition 4 *Soit Δ une base de données numériques, C (respectivement C_t) l'ensemble de tous les motifs graduels fréquents classiques (respectivement l'ensemble de tous les motifs graduels fréquents respectant la contrainte temporelle) extraits de Δ. Nous avons $|C| \geq |C_t|$.*

***Preuve* 4** *Soit Δ une base de données numériques et $minSupp$ un seuil de support minimum. Supposons qu'il existe un itemset graduel fréquent s tel que $s \in C_t$ et $s \notin C$. Comme $s \in C_t$, il existe une séquence d'objets $L_s = \langle d_{l_1}, ..., d_{l_j} \rangle$ de Δ respectant s et respectant l'ordre temporel, et tel que $|L_s| \geq minSupp$. Comme tous les objets dans L_s sont des objets de Δ, s est un itemset graduel fréquent, donc $s \in C$. Cela contredire le fait que $s \notin C$. On peut avoir des motifs graduels fréquents qui appartiennent à C et pas à C_t, c'est le cas des motifs graduels dont les plus longues séquences d'objets correspondantes ne respectent pas l'ordre temporel.*

5 Etude expérimentale

Dans cette section, nous présentons notre étude expérimentale sur les bases de données d'indicateurs paléoécologiques provenant du lac d'aydat (Massif Central Français).

Nous comparons premièrement en terme de nombre de motifs graduels fermés fréquents extraits, l'efficacité de l'algorithme `ParaMiner` original à celui intégrant la contrainte temporelle. Nous montrons expérimentalement que le nombre de motifs graduels respectant l'ordre temporel pouvant être extrait d'une base de données est inférieur au nombre total de motifs graduels classiques.

Nous validons ensuite l'intérêt de notre approche à travers quelques motifs intéressants extraits d'une base de données paléoécologiques contenant 57 objets et 267 attributs multivariés (indicateurs paléoécologiques) liés aux conditions paléo-hydrologiques (statut trophique de l'eau) et d'anthropisation (grains de pollens). Ces motifs interprétés et validés par les experts paléoécologiques correspondent aux groupements intéressants d'indicateurs paléoécologiques d'évolution d'hydrosystèmes. Les nouvelles connaissances révélées par ces groupements dans le domaine de la recherche environnementale montrent la valeur de l'approche proposée.

Les résultats donnés par le tableau 3 fut menés sur une base de données paléoécologiques d'indicateurs d'anthropisation, le seuil de support minimum *(minSupp)* fut fixé à 0.25. Pour chaque jeu de données, nous donnons dans la première colonne le nombre d'objets (respectivement le nombre d'attributs) `nbObj` (respectivement `nbAtt`), dans la seconde colonne, le nombre total de motifs graduels extraits sous contrainte temporelle et dans la troisième colonne le nombre total de motifs graduels classiques extraits.

nbObj - nbAtt	Notre approche	Paraminer
$20 - 20$	**55 792**	112 178
$40 - 20$	**5 999**	12 150
$80 - 20$	**871**	1 824
$111 - 20$	**347**	778
$111 - 30$	**259 071**	519 528
$111 - 40$	**10 183 349**	20 372 438

TAB. 3 – *Evaluation comparative de notre approche contre Paraminer original, avec variation du nombre d'objets et du nombre d'attributs.*

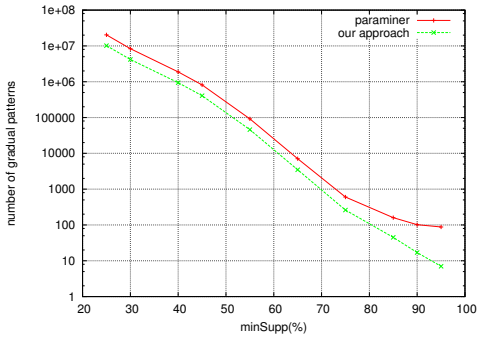


FIG. 1 – *Evaluation comparative de notre approche contre Paraminer original, avec variation de minSupp.*

La figure 1 montre le nombre de motifs graduels fermés fréquents extraits avec variation du support minimum *minSupp* à partir d'une base de données paléoécologiques contenant 111 objets et 40 attributs (indicateurs d'anthropisation paléoécologiques). Sur cette figure, nous focalisons sur la variation du nombre de motifs respectant l'ordre temporel par rapport au nombre de motifs graduels classiques extraits en utilisant `Paraminer` pour une valeur de support minimum *minSupp* donnée. Nous pouvons voir que le nombre de motifs graduels extraits avec notre approche est généralement inférieur au nombre de motifs graduels extraits avec `Paraminer` comme indiqué par la proposition 4.

Afin d'éviter d'extraire à la fois un itemset graduel et ses itemsets graduels frères, pour le reste des expérimentations, le seuil de support minimum *(minSupp)* est fixé à 0.51.

Le tableau 4 montre quelques motifs graduels fermés fréquents parmi les plus pertinents extraits de nos bases de données paléoécologiques. Ces motifs correspondent aux groupements d'indicateurs paléoécologiques d'évolution de la biodiversité (floristique et limnologique) au cours du temps.

	Motifs graduels
1	Pediastrum$^{\geq}$,Conochilus.natans.type$^{\geq}$,Anabaena$^{\geq}$
2	Conochilus.natans.type$^{\geq}$,Trichocerca.cylindrica.type$^{\geq}$,Anabaena$^{\geq}$
3	Botryococcus$^{\geq}$,Conochilus.natans.type$^{\geq}$,Anabaena$^{\geq}$
4	Spirogyra$^{\geq}$,Conochilus.hippocrepis.type$^{\geq}$,Turbellarieae$^{\geq}$
5	Conochilus.hippocrepis.type$^{\geq}$,Trichocerca.cylindrica.type$^{\geq}$,Turbellarieae$^{\geq}$
6	AUSU$^{\geq}$,Botryococcus$^{\geq}$,Turbellarieae$^{\geq}$

TAB. 4 – *Motifs graduels intéressants extraits des bases de données paléoécologiques.*

5.1 Interprétation des résultats

Les motifs extraits et présentés dans le tableau 4 sont pertinents car ils sont indicateurs d'enrichissement trophique. Ils contiennent des coévolutions d'indicateurs qui sont cohérents dans la mesure où les motifs d'indicateurs traduisent une coévolution des taxons (indicateurs

paléoécologiques) indiquant un statut trophique élevé des eaux du lac d'Aydat. Ces motifs permettent aussi de renforcer, voire de préciser le potentiel paléoécologique de certains taxons.

Préférences écologiques des assemblages diatomées

Cette étude, au régard des motifs présentés dans le tableau 4 fourni des informations sur les préférences écologiques des taxons. Par exemple, l'indicateur *AUSU (Aulacoseira subarctica)*, lié au motif 6 est associé avec des taxons liés à la haute teneur en éléments nutritifs (*Botryococcus, Turbellarieae*), c'est donc une diatomée méso-eutrophique. Cependant, cette espèce est classiquement considérée comme *oligo-mesotrophique* par les diatomistes (Dam et al. (1994)) et donc associé à une faible teneur en éléments nutritifs. *Dans les études paléoenvironnement européennes, Rioual (2000) et Sienkiewicz et Gasiorowski (2014) indiquent que cette espèce exige des concentrations élevées de phosphore et d'azote et est liée aux conditions mésotrophes, nos résultats se conforment ainsi à leurs observations.*

Ce type d'étude à l'aide des motifs graduels permet ainsi d'avoir une meilleure connaissance des préférences écologiques et de considérer ces taxons comme préférant des eaux mésotrophes et/ou eutrophes dans le contexte paléoenvironnemental.

Validation du potentiel paléoécologique à partir de nouveaux indicateurs

Ces motifs extraits confirment par ailleurs l'utilité de certains micro-fossiles non polliniques dans la caractérisation d'un enrichissement trophique d'un lac (Miras et al., 2015). C'est particulièrement le cas des œufs de rotifères tels *Conochilus.natans.type, Trichocerca.cylindrica* dans les motifs 1 et 2. Cette analyse permet aussi de mieux caractériser la valeur du potentiel écologique des œufs de rotifères au repos. Traditionnellement associés aux conditions oligotrophes (Schöll (2002)), *Conochilus hippocrepis* est plutôt lié aux conditions d'eau très riche en éléments nutritifs dans notre étude, exprimant une écologie plus ubiquiste de ce taxon.

6 Conclusion

Dans ce papier, nous avons proposé une approche pour l'extraction automatique des motifs graduels fermés fréquents quand l'ordre des objets supportant les motifs suit un ordre temporel. Nous avons présenté un domaine d'application (la Paléoécologie) où la fouille de motifs graduels sous contrainte temporelle est d'intérêt. Nous montrons que, dans ce contexte, prendre en compte la contrainte temporelle durant le processus de fouille permet de réduire significativement la quantité de motifs extraits en éliminant les motifs dont les séquences d'objets correspondants ne respectent pas l'ordre temporel. Un algorithme dédié à l'extraction des motifs graduels fermés fréquents dans ce contexte de temporalité a été proposé et a été implémenté. Les expérimentations menés sur les données paléoécologiques permet d'apprendre des groupements fonctionnels de coévolutions d'indicateurs paléoécologiques qui modélisent l'évolution de la biodiversité au cours du temps. Les nouvelles connaissances révélées par ces groupements montrent la valeur ajoutée de l'approche pour la Paléoécologie.

Références

Ayouni, S., A. Laurent, S. B. Yahia, et P. Poncelet (2010). Mining closed gradual patterns. In *Artificial Intelligence and Soft Computing, 10th International Conference, ICAISC, Zakopane, Poland, June 13-17, Part I*, pp. 267–274.

Berzal, F., J. C. Cubero, D. Sánchez, M. A. V. Miranda, et J. Serrano (2007). An alternative approach to discover gradual dependencies. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15*(5), 559–570.

Boley, M., T. Horváth, A. Poigné, et S. Wrobel (2010). Listing closed sets of strongly accessible set systems with applications to data mining. *Theor. Comput. Sci. 411*(3), 691–700.

Dam, H. V., A. Mertens, et J. Sinkeldam (1994). A coded checklist and ecological indicator values of freshwater diatoms from the netherlands. *Netherlands Journal of Aquatic Ecology. 28*(1), 117–133.

Di-Jorio, L., A. Laurent, et M. Teisseire (2008). Fast extraction of gradual association rules : a heuristic based method. In *CSTST, Cergy-Pontoise, France, October 28-31*, pp. 205–210.

Di-Jorio, L., A. Laurent, et M. Teisseire (2009). Mining frequent gradual itemsets from large databases. In *IDA, Lyon, France, August 31 - September 2*, pp. 297–308.

Do, T. D. T., A. Laurent, et A. Termier (2010). PGLCM : efficient parallel mining of closed frequent gradual itemsets. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pp. 138–147.

Do, T. D. T., A. Termier, A. Laurent, B. Négrevergne, B. O. Tehrani, et S. Amer-Yahia (2015). PGLCM : efficient parallel mining of closed frequent gradual itemsets. *Knowl. Inf. Syst. 43*(3), 497–527.

Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. In *Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002, Helsinki, Finland, August 19-23, 2002, Proceedings*, pp. 200–211.

Laurent, A., M. Lesot, et M. Rifqi (2009). GRAANK : exploiting rank correlations for extracting gradual itemsets. In *FQAS, Roskilde, Denmark, October 26-28*, pp. 382–393.

Laurent, A., B. Négrevergne, N. Sicard, et A. Termier (2010). Pgp-mc : Towards a multicore parallel approach for mining gradual patterns. In *Database Systems for Advanced Applications, 15th International Conference, DASFAA 2010, Tsukuba, Japan, April 1-4, Proceedings, Part I*, pp. 78–84.

Miras, Y., A. Beauger, M. Lavrieux, V. Berthon, K. Serieyssol, V. Andrieu-Ponel, et P. Ledger (2015). Tracking long-term human impacts on landscape, vegetal biodiversity and water quality in the lake aydat (auvergne, france) using pollen, non-pollen palynomorphs and diatom assemblages. In *Palaeogeography, Palaeoclimatology, Palaeoecology*, pp. 76–90.

Négrevergne, B., A. Termier, M. Rousset, et J. Méhaut (2014). Paraminer : a generic pattern mining algorithm for multi-core architectures. *Data Min. Knowl. Discov. 28*(3), 593–633.

Rioual, P. (2000). Diatom assemblages and water chemistry of lakes in the french massif central : a methodology for reconstruction of past limnological and climate ïňĆuctuations during the eemian period. In *Unpublished Ph.D. Thesis, University College London, London 509 pp*.

Schöll, K. (2002). Seasonal changes in rotifera assemblages of shallow lake in the fertö-hanság national park, hungary. *Oppuscula Zoologica, Budapest. 34*, 85–94.

Sienkiewicz, E. et M. Gasiorowski (2014). Changes in the trophic status of three mountain lakes natural or anthropogenic process ? *Pol. J. Environ. Stud. 23*(3), 875–892.

Smol, H., H. Birks, et W. Last (2001). Tracking environmental change using lake sediments. In *Terrestrial, Algal and Silicaceaous Indicators III. Kluwer Academic Publishers, Dordrecht*, pp. 319–349.

Uno, T., M. Kiyomi, et H. Arimura (2004). LCM ver. 2 : Efficient mining algorithms for frequent/closed/maximal itemsets. In *FIMI, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1*.

Summary

The problem of mining frequent gradual pattern consists in mining attribute frequent simultaneous co-variations in numerical databases. Several algorithms for automatically extracting such patterns have been proposed in the literature. The main difference between these algorithms resides in the considered variation semantics. In certain application domains of gradual patterns, we have the databases whose the objects follow a temporal order relation. The state of art approaches do not take into account this temporality constraint within the mining process. In this paper we propose an approach for extracting frequent gradual patterns whose ordering of supporting objects matches the temporal order. This approach reduces the number of mined patterns when the objects follow a temporal order relation. The experimental results obtained on the paleoecological data shows the efficiency of our approach and the interpretation of results bring new knowledge to paleoecological experts and thereby shows the interest of this type of patterns.

Analyse du comportement des contributeurs dans l'Information Géographique Volontaire via la construction de réseaux sociaux

Quy Thy Truong*,**, Cyril De Runz*,
Guillaume Touya**

*Modeco, CReSTIC, University of Reims Champagne-Ardenne,
CS 30012, f-51687, Reims cedex 2, France
quy-thy.truong@ign.fr
**Univ. Paris-Est, LASTIG COGIT, IGN, ENSG,
F-94160 Saint-Mande, France

Résumé. Modéliser les interactions sociales au sein de projets de cartographie volontaire et citoyenne nécessite de définir ce qui relie les contributeurs entre eux dans le temps et l'espace. Dans un souci de réalisme, plutôt que d'étudier un seul type de relation, nous choisissons de construire un réseau social multi-couche contenant différents types d'interactions. L'analyse d'un tel multigraphe devrait permettre de détecter des communautés entre les collaborateurs spatio-temporels et de définir des profils de contributeurs typiques. Un cas d'étude sur OpenStreetMap illustre les déductions pouvant être faites sur les contributions à partir de leurs auteurs.

1 Introduction

Les contributions issues de la collecte collaborative d'information géographique sont des données spatio-temporelles. En effet, l'Information Géographique Volontaire (VGI) est une représentation spatiale du monde réel faite par un contributeur à un instant donné. Ce mode d'acquisition collaboratif offre la possibilité aux contributeurs de se compléter et se corriger mutuellement, permettant ainsi l'enrichissement et la mise à jour rapide de la base de données. Bien que cet aspect du fonctionnement collaboratif semble favoriser la qualité des données, la liberté accordée à tout contributeur, du plus expert au plus malintentionné, nuance cette affirmation. La qualité des données VGI reste donc une problématique encore étudiée dans la recherche actuelle.

Pour assurer la qualité des données VGI, Goodchild et Li (2012) ont défini une approche qui "repose sur une hiérarchie d'individus de confiance agissant en tant que modérateurs". Cette approche suppose d'analyser les actions de contribution permettant d'identifier les contributeurs fiables. De plus, elle exige de qualifier les réactions du contributeur face aux contributions d'autrui. Ce papier aborde la problématique suivante : comment modéliser les interactions entre contributeurs à partir de leurs activités spatio-temporelles dans le contexte de qualification des données VGI ? Le sujet de

cet article se positionne comme un cas d'utilisation des données spatio-temporelles. Les interactions entre contributeurs sont obtenues à partir des contributions appartenant à une même fenêtre spatio-temporelle et la collaboration est modélisée sous forme d'un réseau social multiplexe.

De premiers résultats sont présentés à partir d'un cas d'étude sur OpenStreetMap (OSM), montrant que, malgré l'hétérogénéité caractéristique des contributeurs VGI et de leurs contributions, la communauté accorde sa confiance à certains contributeurs et par conséquent à leurs contributions.

2 Travaux connexes

Considérer les communautés de VGI comme des réseaux sociaux n'est pas une nouveauté. Mooney et Corcoran (2014) et Stein et al. (2015) ont proposé des modèles de graphes sociaux pour représenter les collaborations entre les contributeurs d'OSM. En général, les communautés de crowdsourcing fonctionnent grâce aux réactions des contributeurs sur les contributions des uns et des autres. Cependant, un seul aspect à la fois était considéré pour modéliser la collaboration.

Selon Kivelä et al. (2014), "réduire un système social à un réseau dans lequel les acteurs sont connectés (...) par un seul type de relation est souvent une approximation extrêmement grossière de la réalité. C'est pourquoi, depuis des décennies, les sociologues reconnaissent qu'il est crucial d'étudier les systèmes sociaux par la construction de plusieurs réseaux sociaux en utilisant différents types de liens entre le même ensemble d'individus". Il s'agit donc de relier deux contributeurs par plusieurs relations s'il partagent des liens sociaux d'ordres différents, par exemple un lien familial, amical, professionnel, etc.

Jankowski-Lorek et al. (2016) ont repris cette idée pour étudier l'impact du travail d'équipe dans la qualité des données collaboratives à partir d'un réseau social multidimensionnel de la communauté de Wikipédia. Stein et al. (2015) ont aussi confronté deux types de collaboration (largeur et profondeur de la collaboration) dans le but d'identifier des profils de contributeurs. En suivant la même idée, nous proposons de modéliser les interactions entre contributeurs par un réseau social multiplexe. Un tel réseau permet de relier les même individus par plusieurs relations de types différents. Ainsi, ce graphe donne une vision plus complète de la collaboration entre les contributeurs d'une communauté.

La modélisation proposée dans ce papier n'apporte rien de nouveau par rappport au concept même des graphes multiplexes. Toutefois, nous proposons de tirer profit des ces travaux existants sur les réseaux sociaux pour les mettre au service de l'étude des contributeurs et de la qualité des contributions VGI. En effet, la modélisation multiplexe de graphes sociaux est, à notre connaissance, une solution innovante dans le domaine de l'Information Géographique Volontaire. L'intérêt de cet article est de montrer les différentes collaborations qui ont lieu au sein d'une même fenêtre spatio-temporelle de contributions, et en quoi ces interactions permettent de qualifier le rôle des contributeurs dans leur communauté, au sens de collaborateurs spatio-temporels.

3 Modélisation des contributeurs VGI dans un réseau social multiplexe

Nous construisons un type de réseau multi-couche particulier, appelé réseau multiplexe, qui est une séquence de graphes $\{G_\alpha\}_{\alpha=1}^{b} = \{(V_\alpha, E_\alpha)\}$ où E_α est l'ensemble des arcs, V_α est l'ensemble des sommets tel que $V_\alpha = V_\beta = V$ pour tout α, β, et b est le nombre de graphes (Kivelä et al., 2014). Dans notre cas, V constitue l'ensemble des contributeurs et représente la seule connexion entre tous ces graphes initialement indépendants. Selon les graphes considérés, certains sommets de V peuvent être isolés. L'enjeu ici est de définir l'ensemble des arcs E_α. Les propositions qui suivent reprennent ou s'inspirent des modèles d'interactions sociales de la littérature.

Dans la littérature portant sur l'Information Géographique Volontaire, les collaborations ont surtout été considérées selon le point de vue des éditions d'objets. On peut définir $E_{co-édition} = \{(A, B, e)\}$ où le contributeur A a créé la version $v+1$ de la version v du même objet produite par le contributeur B, et e l'intensité de la co-édition *i.e.* le nombre de fois où A a directement édité une nouvelle version d'une contribution précédemment éditée par B (Mooney et Corcoran, 2014). Par ailleurs, l'étude de la co-édition peut s'étendre à celle d'un entrelacement de collaboration (interlocking collaboration), permettant alors de caractériser l'intensité de la collaboration (Stein et al., 2015). Dans ce cas, l'entrelacement de collaboration n'est pas restreinte aux éditions successives entre deux contributeurs. Ainsi, la largeur de collaboration est définie par $E_{largeur} = \{(A, B, w)\}$ où w est le nombre d'objets que les contributeurs A et B ont édité en commun ; la profondeur de collaboration est définie par l'ensemble $E_{profondeur} = \{(A, B, d)\}$ où d est le nombre maximum de fois où le contributeur A a répondu (directement ou non) au contributeur B sur le même objet. Au-delà des différences subtiles qui existent entre ces graphes, ces derniers permettent d'estimer si un contributeur tombe d'accord avec la communauté.

Dans (Heaberlin et DeDeo, 2016), un réseau de normes Wikipédia a été défini comme un graphe dont les sommets sont des pages Wikipédia. Deux pages sont connectées lorsque l'une fait référence à l'autre via un lien hypertexte. Réutiliser la contribution d'autrui pour faire sa propre contribution constitue donc un autre type d'interaction. L'ensemble $E_{réutilisation} = \{(A, B, d)\}$ estime la relation de confiance entre deux utilisateurs A et B, en particulier dans le cas où $d > 0$, puisqu'un contributeur utilise la contribution d'un autre contributeur si le premier a confiance en le second (c'est du moins notre hypothèse). Contrairement aux graphes précédents, le graphe de réutilisation est une nouvelle proposition puisqu'il provient d'une adaptation du réseau de normes Wikipédia.

Enfin, connecter les utilisateurs ayant contribué dans la même zone permet de révéler des collaborations existantes dans la communauté, surtout s'ils contribuent en même temps. Ainsi, on peut définir l'ensemble $E_{co-location} = \{(A, B, l)\}$ qui connecte les contributeurs selon le chevauchement l de leurs zones de contribution respectives (Neis et Zipf, 2012; Zielstra et al., 2014), et l'ensemble $E_{co-temporalité} = \{(A, B, t)\}$ qui décrit le lien temporel séparant un couple de contributeurs. L'agrégation des graphes de co-location et de co-temporalité correspond au modèle spatio-temporel de co-occurence introduit dans (Crandall et al., 2010). Ce modèle avait été défini pour identifier des liens

sociaux, par conséquent l'analyse simultanée de ces deux graphes devrait permettre d'étudier ce même problème. Notre proposition consiste à diviser en deux graphes distincts le graphe de co-occurrence.

Par rapport aux différents travaux cités, notre modèle reprend et réadapte les graphes proposés dans un même réseau. En représentant ces différentes relations dans même un système, des contributeurs remarquables pourront être identifiés, en particulier après détection de communauté (clustering).

4 Cas d'application : analyse des contributeurs d'OpenStreetMap

Pour démontrer l'utilité d'un tel graphe multiplexe, un graphe de trois couches a été réalisé sur un cas d'étude OpenStreetMap (OSM). Ces couches correspondent aux interactions de largeur de collaboration, de profondeur de collaboration et de réutilisation. Ces graphes ont été construits à partir du chargement des contributions sur une fenêtre spatio-temporelle qui s'étend de 2010 à fin 2014 et sur le centre de Paris (île de la Cité et ses alentours). Les graphes de largeur et de profondeur de collaboration sont produits à partir des contributions OSM de type ponctuels (node) tandis que le graphe de réutilisation a été produit à partir des contributions OSM de type linéaires (way), c'est-à-dire que l'on considérait uniquement les objets ways composés de nodes édités par un auteur différent. Pour limiter l'affichage aux interactions les plus fortes, les figures suivantes présentent les sommets dont les arcs (entrant ou sortants) sont de poids supérieur à 2. Un filtrage supplémentaire sur les sommets restreint l'affichage des sommets de degré supérieur à 2. Ces graphes sont affichés sur les FIG. 1, 2, 3, et TAB. 1 explicite les paramètres d'affichage des graphes.

Paramètre	Définition
Couleur du sommet	Proportionnel au degré entrant, en suivant les couleurs suivantes : bleu = faible ; blanc = intermédiaire ; rouge = élevé
Taille du sommet	Proportionnel au degré sortant
Couleur de l'arc	Identique à la couleur du sommet initial
Taille de l'arc	Proportionnel au poids *i.e.* l'intensité de la collaboration

TAB. 1 – *Paramètres de visualisation des graphes. D'après la définition d'un réseau multiplexe, chaque graphe est composé du même ensemble de sommets, mais pour des raisons de lisibilité seuls les arcs de poids supérieur à 2 et les sommets de degré supérieur à 2 sont affichés.*

Dans les graphes de largeur et de profondeur de contribution, on remarque trois contributeurs très actifs (représentés par des gros sommets, c'est-à-dire caractérisé par un fort degré sortant) : #17397, #18855 et #33745. Or, le contributeur #17397 présente un degré entrant beaucoup plus faible que les deux autres : celui-ci a donc été moins édité par le reste de la communauté. De plus, le graphe de réutilisation montre que ses contributions ont été largement réutilisées par les autres contributeurs.

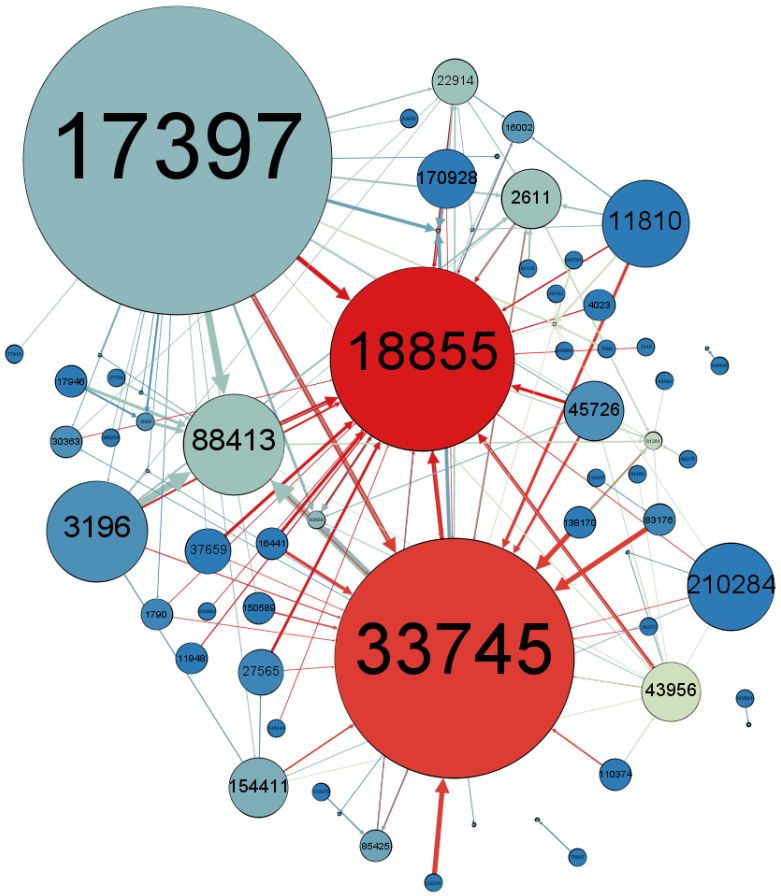


FIG. 1 – *Graphe de largeur de collaboration des contributeurs d'OSM sur Paris (les identifiants sont anonymisés). Un arc dirigé d'un contributeur A vers un contributeur B signifie que A a modifié un objet qui a été précédemment édité par B. L'épaisseur de l'arc correspond au nombre d'objets uniques que le contributeur A a édité après l'utilisateur B.*

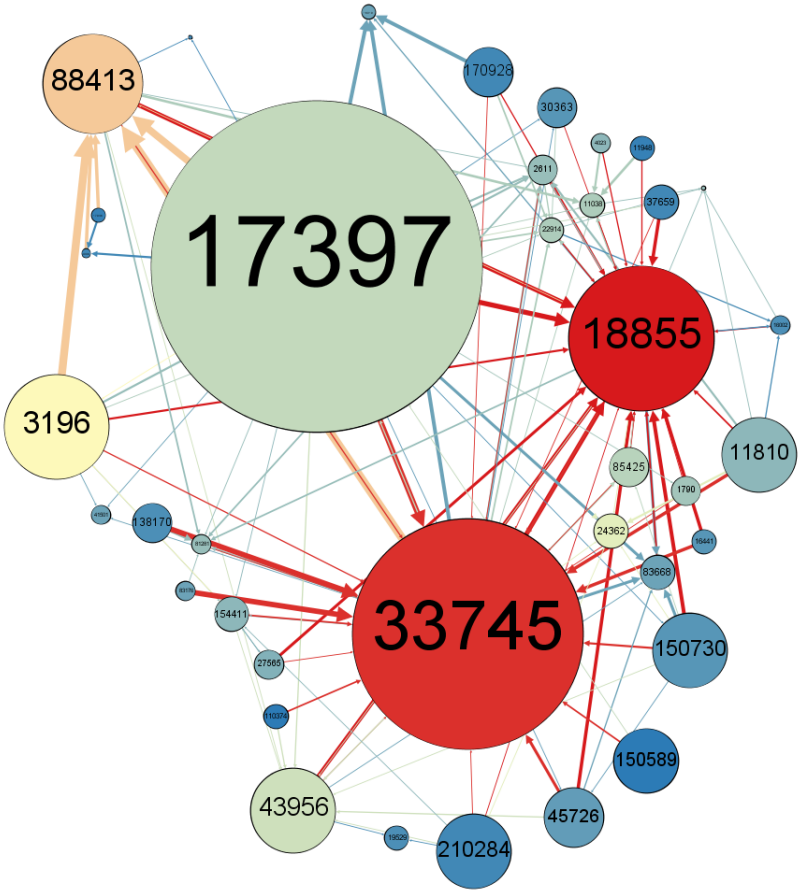


FIG. 2 – *Graphe de profondeur de collaboration des contributeurs OSM sur Paris (les identifiants sont anonymisés). Comme le graphe de largeur de collaboration, un arc dirigé d'un contributeur A vers un contributeur B signifie que A a édité un objet précédemment édité par B. L'épaisseur de l'arc correspond au nombre maximal de fois où A a édité B sur un même objet (mais pas nécessairement de manière directe, d'autres contributeurs ont pu éditer l'objet entre A et B).*

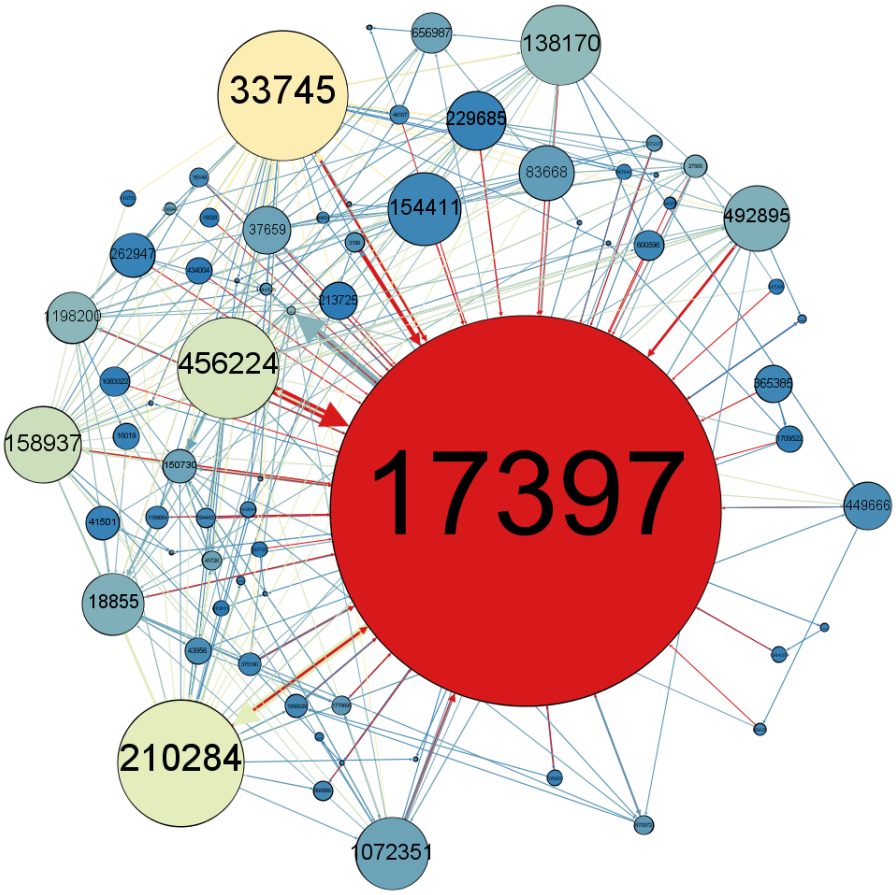


FIG. 3 – *Graphe de réutilisation OSM sur Paris (les identifiants des contributeurs sont anonymisés). Un arc dirigé d'un contributeur A vers un contribtueur B signifie que A a créé un nouvel objet composé d'une contribution de B. En particulier, dans ce cas d'étude on se restreint à l'utilisation d'objets ponctuels composant des objets linéaires.*

La réutilisation de ces données donne du poids à la fiabilité de ces données : en effet, on peut supposer qu'un contributeur A va réutiliser les données d'un autre contributeur B si A considère que les contributions de B sont correctes. Par conséquent, #17397 a un profil proche de celui d'un modérateur, puisqu'il édite beaucoup en étant lui-même peu édité et ses propres contributions sont considérées comme fiables par les autres contributeurs.

Les contributeurs #18855 et #33745 sont édités par de nombreux contributeurs. Les arcs rouges dirigés vers ces deux sommets ont un poids important dans les graphes de largeur et de profondeur de collaboration. Concernant le graphe de largeur de collaboration, cela signifie que ces deux contributeurs ont été édités sur plusieurs objets différents ; par conséquent la largeur de collaboration entre ces deux contributeurs avec le reste de la communauté correspond à une édition qui s'étend spatialement. Quant au graphe de profondeur de collaboration, l'épaisseur d'un arc traduisant le nombre d'interactions maximal avec un contributeur sur un objet donné, on pourra considérer que la profondeur de la collaboration correspond à une interaction qui s'étend dans le temps. Sans même parler de guerre d'éditions, on constate ici que certains contributeurs vont revenir éditer plusieurs fois un même objet après les passages d'un même contributeur. C'est également le cas du contributeur #88413 dont plusieurs arcs entrants de poids relativement important caractérisent une interaction temporellement étendue avec d'autres utilisateurs. On peut donc se poser la question de la fiabilité de ces contributeurs et par conséquent, de leurs contributions. De plus, bien que des graphes de co-location et de co-temporalité auraient été pertinents ici pour décrire la proximité spatio-temporelle entre les contributeurs, les graphes de collaboration (en largeur et en profondeur) permettent déjà de mettre en avant l'intensité spatio-temporelle de la collaboration.

Dans le graphe de réutilisation, le contributeur #33745 apparaît en jaune, ce qui montre qu'un certain nombre de ses contributions ont été réutilisées par d'autres utilisateurs. Cela nuance notre analyse et démontre que les contributions de cet utilisateur ne sont pas forcément de mauvaise qualité. Cela démontre aussi que les graphes de largeur et de profondeur de collaboration ne suffisent pas pour qualifier les contributeurs, ce qui justifie ainsi la nécessité d'une analyse conjointe des différents graphes du réseau multiplexe. Le contributeur #18855 apparaît également sur le graphe de réutilisation, mais de manière moins visible. Précisons toutefois que cela ne signifie pas forcément qu'il est moins fiable que le contributeur #33745 : peut-être que ses contributions ne sont simplement pas faites pour être réutilisées dans des objets linéaires. En effet, tous les objets ponctuels n'ont pas vocation à composer des objets linéaires. Par exemple, les routes ne peuvent pas être construites à partir de points d'intérêt.

5 Conclusion et perspectives

De la même façon que certains travaux menés dans le domaine de l'analyse des réseaux sociaux, ce papier propose de modéliser le comportement des contributeurs VGI au moyen d'un graphe social multiplexe. Les interactions entre contributeurs sont alors récupérées à partir des différents modes de contributions, permettant de reconstituer de manière plus détaillée les liens sociaux dans le temps et dans l'espace. Il est vrai

qu'un tel réseau social met en lumière les perceptions des contributeurs sur les uns et les autres, mais l'analyse va plus loin en abordant la problématique de la qualité des données VGI. En effet, l'application de ce modèle sur OpenStreetMap a pu mettre en évidence des pistes permettant de qualifier les contributions OSM.

Sans grande surprise, la diversité des comportements permet d'expliquer l'hétérogénéité des données. Pour autant, les contributeurs ayant été considérés comme des personnes de confiance dans la communauté témoignent de la fiabilité de leurs contributions. L'analyse de graphes supplémentaires de manière automatisée permettra, selon nous, de saisir la plupart des modes d'opération des contributeurs. En effet, l'intérêt d'utiliser un réseau multiplexe est la possibilité d'y ajouter d'autres graphes, à condition que l'ensemble des sommets soit égal à V (rappelons qu'il est possible que des sommets soient isolés dans un graphe donné). Néanmoins, un écueil à éviter serait d'ajouter des graphes sans s'assurer de leur réel apport vis-à-vis des couches présentes dans le réseau multiplexe initial. Par exemple, on peut remarquer l'existence d'un recouvrement entre les graphes de co-édition et les graphes de collaboration (en largeur et en profondeur), puisqu'ils portent tous sur l'activité d'édition des contributions. Il sera judicieux de faire le tri parmi ces graphes.

L'objectif principal de la construction de ces graphes est de faire ressortir une typologie selon laquelle seront classifiés les contributeurs VGI de manière automatique. A cette fin, nos recherches futures porteront d'une part sur la détection de communauté dans les graphes multiplexes y compris des méthodes de partitionnement (*i.e.* méthodes de clustering) (Battiston et al., 2014). D'autre part, la typologie résultante devra être comparée aux enquêtes qualitatives effectuées, comme celles de Coleman et al. (2010) et Duféal et al. (2016).

Références

Battiston, F., V. Nicosia, et V. Latora (2014). Structural measures for multiplex networks. *Phys. Rev. E 89*, 032804.

Coleman, D. J., Y. Geogiadou, et J. Labonte (2010). Volunteered geographic information : The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research 4*.

Crandall, D. J., L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, et J. Kleinberg (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences 107*(52), 22436–22441.

Duféal, M., C. Jonchères, et M. Noucher (2016). ECCE CARTO - DES ESPACES DE LA CONTRIBUTION A LA CONTRIBUTION SUR L'ESPACE - Profils, pratiques et valeurs d'engagement des contributeurs d'OpenStreetMap (OSM). Research report, UMR 5319.

Goodchild, M. F. et L. Li (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*.

Heaberlin, B. et S. DeDeo (2016). The evolution of wikipedia's norm network. *Future Internet 8*(2), 14+.

Jankowski-Lorek, M., S. Jaroszewicz, Ł. Ostrowski, et A. Wierzbicki (2016). Verifying social network models of wikipedia knowledge community. *Information Sciences 339*, 158–174.

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, et M. A. Porter (2014). Multilayer networks. *Journal of Complex Networks 2*(3), 203–271.

Mooney, P. et P. Corcoran (2014). Analysis of interaction and co-editing patterns amongst openstreetmap contributors. *Transactions in GIS 18*(5), 633–659.

Neis, P. et A. Zipf (2012). Analyzing the contributor activity of a volunteered geographic information project — the case of openstreetmap. *International Journal of Geo-Information*, 146–165.

Stein, K., D. Kremer, et C. Schlieder (2015). Spatial collaboration networks of OpenStreetMap. In J. Jokar Arsanjani, A. Zipf, P. Mooney, et M. Helbich (Eds.), *OpenStreetMap in GIScience*, Lecture Notes in Geoinformation and Cartography, pp. 167–186. Springer International Publishing.

Zielstra, D., H. Hochmair, P. Neis, et F. Tonini (2014). Areal delineation of home regions from contribution and editing patterns in openstreetmap. *ISPRS International Journal of Geo-Information 3*(4), 1211–1233.

Summary

Modelling social interactions in volunteered geographic information projects requires defining what binds contributors together in time and space. In order to be as realistic as possible, instead of choosing one social aspect to study, we choose to build a multi-layered social network that contains several types of interaction between VGI contributors. The analysis of such a multigraph should allow the detection of communities and the definition of typical profiles of contributors. A use case on OpenStreetMap illustrates what inferences can be made about contributions based on their authors.

Index