Extraction de co-localisations sous contrainte de la structure spatiale

Rodrigue Govan[1], Nazha Selmaoui-Folcher[1], Aristotelis Giannakos[2], Philippe Fournier-Viger[3]

{rodrigue.govan,nazha.selmaoui}@unc.nc – 4 Juillet 2023

[1] Université de la Nouvelle-Calédonie – ISEA
[2] Université de Picardie Jules Verne – EPROAD
[3] Shenzhen University – Big Data Institute (CSSE)

CNIA'23 – Strasbourg (France)

Introduction

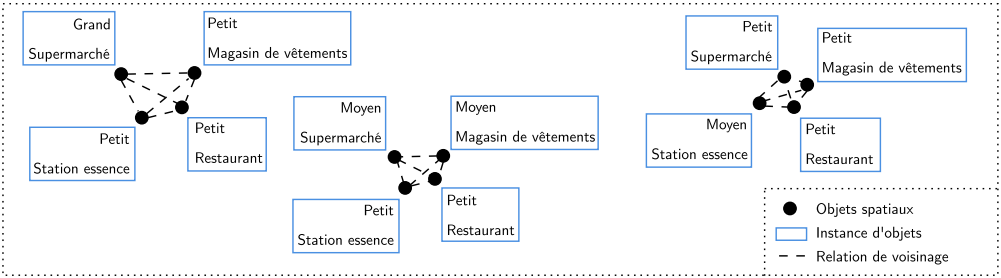
Contexte

Dans le domaine de la fouille de données, l'extraction de co-localisations est une des méthodes permettant d'extraire des connaissances (prenant en compte la dimension spatiale des données) [HSX04 ; KH95 ; SH01].

Introduction

Dans le domaine de la fouille de données, l'extraction de co-localisations est une des méthodes permettant d'extraire des connaissances (prenant en compte la dimension spatiale des données) [HSX04 ; KH95 ; SH01].

- Co-localisation : Sous-ensemble de caractéristiques associées à des objets qui sont géographiquement proches les uns des autres.



Exemple d'une co-localisation avec ses objets spatiaux et leurs instances.

Introduction

En théorie des graphes, les sous-ensembles d'objets géographiquement proches peuvent s'obtenir par l'extraction des cliques maximales [KKK11 ; Yao+16 ; BW19 ; Tra+21].

Introduction

En théorie des graphes, les sous-ensembles d'objets géographiquement proches peuvent s'obtenir par l'extraction des cliques maximales [KKK11 ; Yao+16 ; BW19 ; Tra+21].

Soit $G = (V, E)$, un graphe non orienté avec $V = \{v_1, v_2, \ldots, v_n\}$ l'ensemble des sommets et $E = \{(v_i, v_j) \in V^2 \mid \forall i, j \in \{1, \ldots, n\} \text{ et } i \neq j\}$. Si deux sommets v_i et v_j sont liés par une arête i.e., $(v_i, v_j) \in E$, alors on dit que v_i et v_j sont adjacents.

- Graphe complet : Un graphe est dit complet si et seulement si chaque paire de sommets du graphe est adjacent.

Introduction

En théorie des graphes, les sous-ensembles d'objets géographiquement proches peuvent s'obtenir par l'extraction des cliques maximales [KKK11 ; Yao+16 ; BW19 ; Tra+21].

Soit $G = (V, E)$, un graphe non orienté avec $V = \{v_1, v_2, \ldots, v_n\}$ l'ensemble des sommets et $E = \{(v_i, v_j) \in V^2 \mid \forall i, j \in \{1, \ldots, n\} \text{ et } i \neq j\}$. Si deux sommets v_i et v_j sont liés par une arête i.e., $(v_i, v_j) \in E$, alors on dit que v_i et v_j sont adjacents.

- Graphe complet : Un graphe est dit complet si et seulement si chaque paire de sommets du graphe est adjacent.

Soit $g = (V_g, E_g)$ un sous-graphe de G tel que $V_g \subseteq V$ et $E_g \subseteq \{(v_{g,i}, v_{g,j}) \in E \mid v_{g,i} \in V_g \wedge v_{g,j} \in V_g\}$.

Introduction

En théorie des graphes, les sous-ensembles d'objets géographiquement proches peuvent s'obtenir par l'extraction des cliques maximales [KKK11 ; Yao+16 ; BW19 ; Tra+21].

Soit $G = (V, E)$, un graphe non orienté avec $V = \{v_1, v_2, \ldots, v_n\}$ l'ensemble des sommets et $E = \{(v_i, v_j) \in V^2 \mid \forall i, j \in \{1, \ldots, n\} \text{ et } i \neq j\}$. Si deux sommets v_i et v_j sont liés par une arête i.e., $(v_i, v_j) \in E$, alors on dit que v_i et v_j sont adjacents.

- Graphe complet : Un graphe est dit complet si et seulement si chaque paire de sommets du graphe est adjacent.

Soit $g = (V_g, E_g)$ un sous-graphe de G tel que $V_g \subseteq V$ et $E_g \subseteq \{(v_{g,i}, v_{g,j}) \in E \mid v_{g,i} \in V_g \wedge v_{g,j} \in V_g\}$.

- Clique : Une clique de G est un sous-graphe $g \subseteq G$ tel que g est complet ;

Introduction

En théorie des graphes, les sous-ensembles d'objets géographiquement proches peuvent s'obtenir par l'extraction des cliques maximales [KKK11 ; Yao+16 ; BW19 ; Tra+21].

Soit $G = (V, E)$, un graphe non orienté avec $V = \{v_1, v_2, \ldots, v_n\}$ l'ensemble des sommets et $E = \{(v_i, v_j) \in V^2 \mid \forall i, j \in \{1, \ldots, n\} \text{ et } i \neq j\}$. Si deux sommets v_i et v_j sont liés par une arête i.e., $(v_i, v_j) \in E$, alors on dit que v_i et v_j sont adjacents.

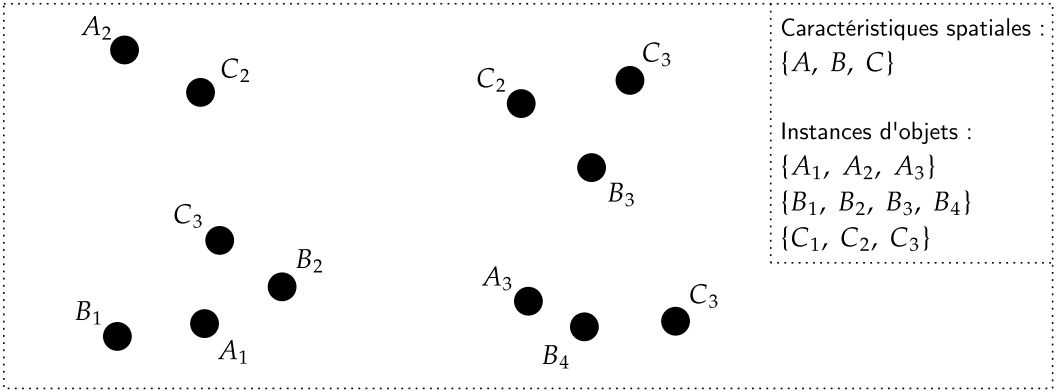
- Graphe complet : Un graphe est dit complet si et seulement si chaque paire de sommets du graphe est adjacent.

Soit $g = (V_g, E_g)$ un sous-graphe de G tel que $V_g \subseteq V$ et $E_g \subseteq \{(v_{g,i}, v_{g,j}) \in E \mid v_{g,i} \in V_g \wedge v_{g,j} \in V_g\}$.

- Clique : Une clique de G est un sous-graphe $g \subseteq G$ tel que g est complet ;
- Clique maximale : La clique g est dite maximale si et seulement s'il n'existe pas de clique g' telle que $g \subset g' \subseteq G$.

Introduction

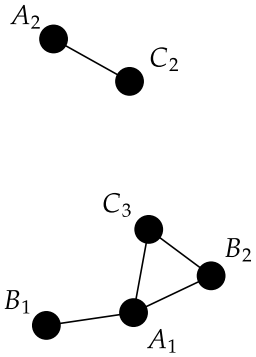
Exemple de co-localisations basées sur un ensemble de cliques de données spatialisées.

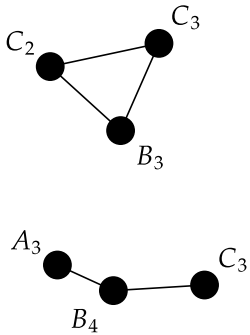


a) Données spatiales avec les caractéristiques et leurs instances.

Introduction

Exemple de co-localisations basées sur un ensemble de cliques de données spatialisées.





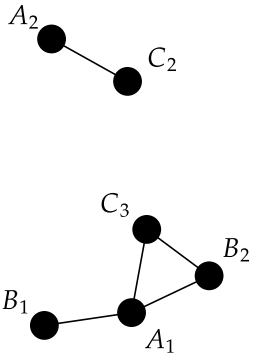
Selon un seuil de distance de voisinage d donné, deux instances d'objets sont adjacentes (voisines) si leur distance $\leqslant d$ (seuil fixé).

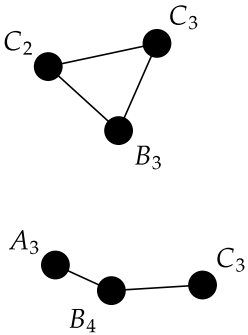
Voisins :
$\{A_1 - C_2\}$, $\{B_2 - C_3\}$, ...
Pas voisins :
$\{A_3 - B_3\}$, $\{B_1 - C_3\}$, ...

b) Représentation du graphe selon un seuil de distance d.

Introduction

Exemple de co-localisations basées sur un ensemble de cliques de données spatialisées.





Cliques :
$\{A_2, C_2\}$, $\{B_3, C_2, C_3\}$, $\{A_3, B_4\}$

Pas une clique :
$\{B_1, B_2\}$, $\{A_3, B_4, C_3\}$, $\{A_1, B_1, C_3\}$, ...

Note : tous les objets spatiaux dans une clique sont voisins ($\leqslant d$).

c) Ensemble de cliques à partir de la représentation du graphe.

Introduction

Exemple de co-localisations basées sur un ensemble de cliques de données spatialisées.

Instances/cliques de la co-localisation $\{A,\ B\}$: $\{A_1,\ B_1\}$, $\{A_1,\ B_2\}$, $\{A_3,\ B_4\}$

Instances/cliques de la co-localisation $\{B,\ C\}$: $\{B_3,\ C_2\}$, $\{B_3,\ C_3\}$, $\{B_2,\ C_3\}$, $\{B_4,\ C_3\}$

Instances/cliques de la co-localisation $\{A,\ C\}$: $\{A_2,\ C_2\}$, $\{A_1,\ C_3\}$

Instances/cliques de la co-localisation $\{A,\ B,\ C\}$: $\{A_1,\ B_2,\ C_3\}$

d) Co-localisations à partir de l'ensemble de cliques.

Introduction

Pour définir si une co-localisation est pertinente, l'indice de participation (basé sur le ratio de participation) est utilisé. L'indice de participation est aussi appelé la prévalence.

Introduction

Pour définir si une co-localisation est pertinente, l'indice de participation (basé sur le ratio de participation) est utilisé. L'indice de participation est aussi appelé la prévalence.

- Ratio de participation : $Pr(f_i, \mathcal{C}) = \frac{|\{ \text{ instances de } f_i \text{ participant à } \mathcal{C})\}|}{|\{ \text{ instances de } f_i \ \}|}$
- Indice de participation : $Pi(\mathcal{C}) = \min_{f_i \in \mathcal{C}} Pr(f_i, \mathcal{C})$

Introduction

Pour définir si une co-localisation est pertinente, l'indice de participation (basé sur le ratio de participation) est utilisé. L'indice de participation est aussi appelé la prévalence.

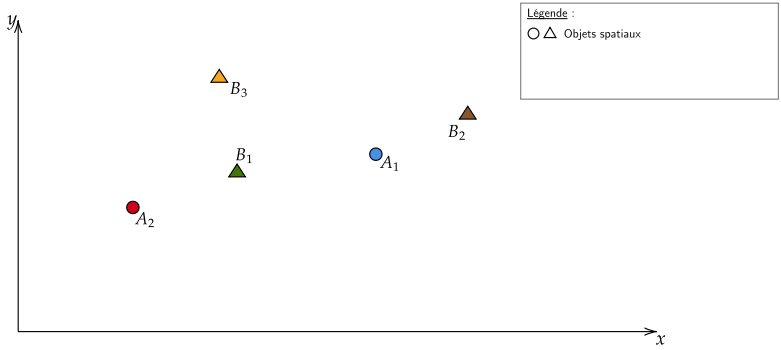
- Ratio de participation : $Pr(f_i, \mathcal{C}) = \frac{|\{ \text{ instances de } f_i \text{ participant à } \mathcal{C})\}|}{|\{ \text{ instances de } f_i \ \}|}$
- Indice de participation : $Pi(\mathcal{C}) = \min_{f_i \in \mathcal{C}} Pr(f_i, \mathcal{C})$

Instances/cliques de la co-localisation $\{A, B\}$: $\{A_1, B_1\}$, $\{A_1, B_2\}$, $\{A_3, B_4\}$

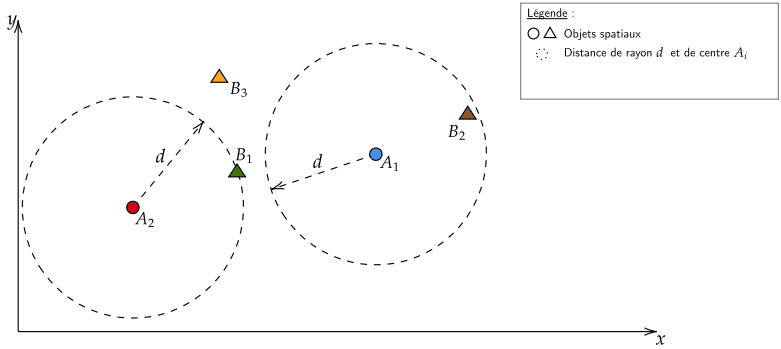
Avec A_1, A_2, A_3 comme instances de A et B_1, B_2, B_3, B_4 comme instances de B de l'exemple précédent, nous avons :

- $Pr(A, \{A, B\}) = \frac{|\{A_1, A_3\}|}{|\{A_1, A_2, A_3\}|} = \frac{2}{3}$ et $Pr(B, \{A, B\}) = \frac{|\{B_1, B_2, B_4\}|}{|\{B_1, B_2, B_3, B_4\}|} = \frac{3}{4}$
- $Pi(\{A, B\}) = \min_{f_i \in \{A,B\}} Pr(f_i, \{A, B\}) = \min(\frac{2}{3}, \frac{3}{4}) = \frac{2}{3}$

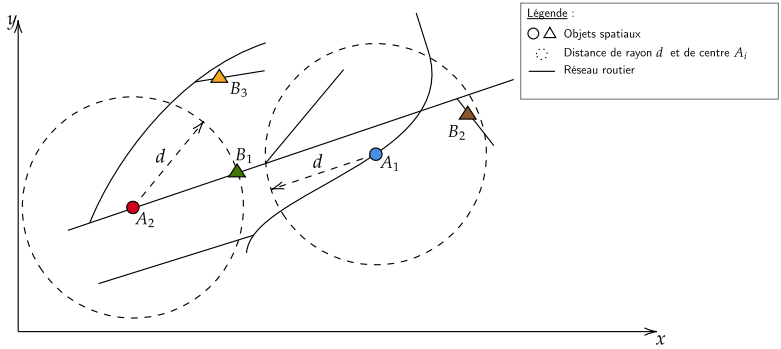
Problématique



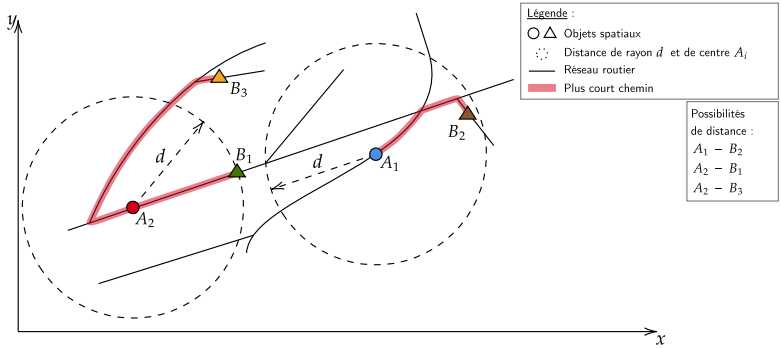
Problématique



Problématique



Problématique



Pour définir la relation de voisinage, la majorité des méthodes se basent sur une distance euclidienne [YS06 ; WBL09 ; Kim+14] ;

- La structure spatiale n'est pas prise en compte !
- Il faut donc considérer une autre mesure de voisinage.

Méthode

Prise en compte de la contrainte

Soit $\mathcal{O}$ un ensemble d'objets spatiaux et $\mathcal{F}$ un ensemble de caractéristiques. On note G_S, un graphe représentant la structure spatiale tels que $G_S = (V_S, E_S)$ où V_S est l'ensemble des sommets et E_S l'ensemble des arêtes.

L'intégration de la contrainte se réalise selon les étapes suivantes :

1. Pour chaque objet $o_i \in \mathcal{O}$, nous l'associons dans la structure spatiale G_S au plus proche objet $o_S \in V_S$ (par la distance euclidienne) ;

Méthode

Prise en compte de la contrainte

Soit $\mathcal{O}$ un ensemble d'objets spatiaux et $\mathcal{F}$ un ensemble de caractéristiques. On note G_S, un graphe représentant la structure spatiale tels que $G_S = (V_S, E_S)$ où V_S est l'ensemble des sommets et E_S l'ensemble des arêtes.

L'intégration de la contrainte se réalise selon les étapes suivantes :

1. Pour chaque objet $o_i \in \mathcal{O}$, nous l'associons dans la structure spatiale G_S au plus proche objet $o_S \in V_S$ (par la distance euclidienne) ;
2. Nous déterminons le plus court chemin pour chaque objet de V_S aux autres objets localisés dans un cercle de rayon d selon la distance euclidienne ;

Méthode

Prise en compte de la contrainte

Soit $\mathcal{O}$ un ensemble d'objets spatiaux et $\mathcal{F}$ un ensemble de caractéristiques. On note G_S, un graphe représentant la structure spatiale tels que $G_S = (V_S, E_S)$ où V_S est l'ensemble des sommets et E_S l'ensemble des arêtes.

L'intégration de la contrainte se réalise selon les étapes suivantes :

1. Pour chaque objet $o_i \in \mathcal{O}$, nous l'associons dans la structure spatiale G_S au plus proche objet $o_S \in V_S$ (par la distance euclidienne) ;
2. Nous déterminons le plus court chemin pour chaque objet de V_S aux autres objets localisés dans un cercle de rayon d selon la distance euclidienne ;
3. Si la longueur du plus court chemin [Dij+59 ; VK11] entre deux objets de V_S est inférieure ou égale au seuil de distance d, alors ces objets sont considérés comme voisins.

Méthode

Construction du graphe

Sous la contrainte de la structure spatiale, le graphe est défini tel que $G = (\mathcal{O}, E_{\mathcal{O}})$ avec :

- $\mathcal{O}$ est l'ensemble des objets spatiaux ;
- $E_{\mathcal{O}} = \{(o_i, o_j) \mid \exists (o_{S,i}, o_{S,j}) \in E_S, D_{sp}(o_{S,i}, o_{S,j}) \leq d, \forall (i,j) \in [\![1, n]\!]^2, i \neq j\}$.

L'extraction des cliques maximales pour obtenir les co-localisations s'effectuera à partir de ce graphe.

Données

Paris et Chicago

Les données ont été créées en collectant des variables via les plateformes *OpenData*[1].

Variable	Attributs	# Objets
Lycées	Type	239
Cinémas	# Sièges (**)	85
(*) Vélos	Capacité (**)	996
Parcs	Type	722
(*) Métros	Ligne	326
Nombre total d'objets spatiaux : 2 968		

Description des données de Paris.

(*) : La variable concerne des stations.

(**) : Les données ont été discrétisées par quantile.

[1] `opendata.paris.fr/`, `data.iledefrance.fr/`, `data.cityofchicago.org/`

Données

Paris et Chicago

Les données ont été créées en collectant des variables via les plateformes *OpenData*[1].

Variable	Attributs	# Objets
Lycées	Type	239
Cinémas	# Sièges (**)	85
(*) Vélos	Capacité (**)	996
Parcs	Type	722
(*) Métros	Ligne	326
Nombre total d'objets spatiaux : 2 968		

Description des données de Paris.

Variable	Attributs	# Objets
Lycées	Type	142
(*) Bus	# Lignes	5 606
(*) Tramway	# Lignes	124
Fast Food		877
(*) Vélos	Capacité (*)	1 402
Parcs	Type	613
Nombre total d'objets spatiaux : 8 764		

Description des données de Chicago.

(*) : La variable concerne des stations.

(**) : Les données ont été discrétisées par quantile.

[1] `opendata.paris.fr/`, `data.iledefrance.fr/`, `data.cityofchicago.org/`

Données

Structure spatiale : réseau routier

Les réseaux routiers ont été récupérées via `OSMnx` [Boe17] :

- Il est possible d'obtenir le réseau accessible à pied ou en voiture ;
- Le réseau est convertible sous forme de graphe.

Données

Structure spatiale : réseau routier

Les réseaux routiers ont été récupérées via `OSMnx` [Boe17] :

- Il est possible d'obtenir le réseau accessible à pied ou en voiture ;
- Le réseau est convertible sous forme de graphe.

Le réseau routier de Paris se compose de 42 870 sommets 241 016 arêtes ;

Réseau routier de Paris intra-muros.

Données

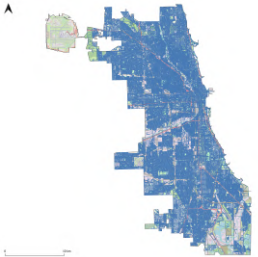
Structure spatiale : réseau routier

Les réseaux routiers ont été récupérées via `OSMnx` [Boe17] :

- Il est possible d'obtenir le réseau accessible à pied ou en voiture ;
- Le réseau est convertible sous forme de graphe.

Le réseau routier de Paris se compose de 42 870 sommets 241 016 arêtes ;
Le réseau routier de Chicago se compose de 184 476 sommets et 1 217 928 arêtes.

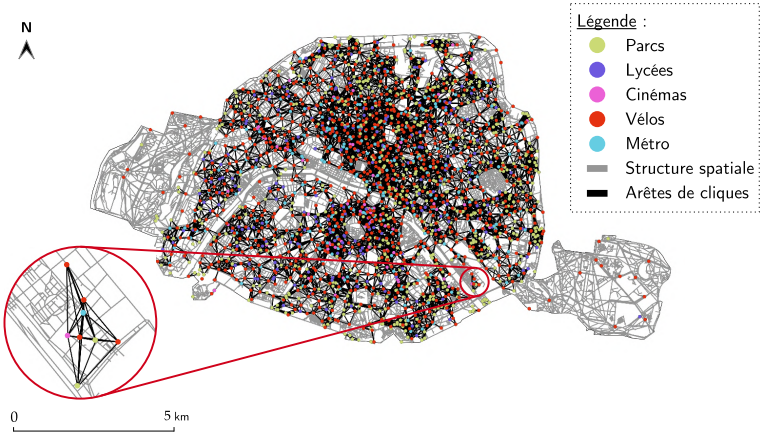
Réseau routier de Paris intra-muros.



Réseau routier de Chicago.

Résultats

Distribution



Données de Paris avec les cliques maximales extraites.

Résultats

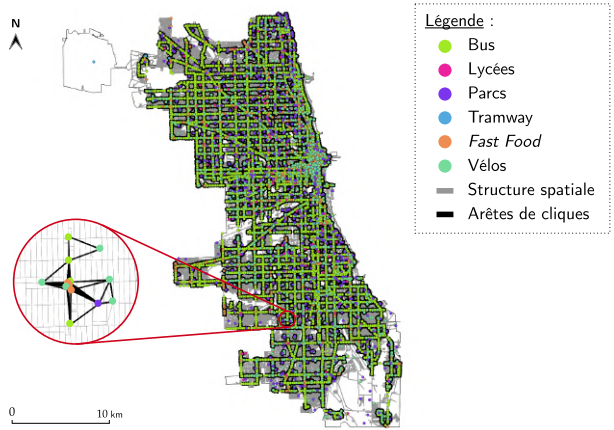
Motifs spatiaux

Co-localisation	Prévalence sous contrainte	Prévalence sans contrainte
{Parcs, Lycées, Vélos}	0.89	0.89
{Lycées, Vélos}	0.86	0.86
{Lycées, Cinémas, Vélos}	0.71	0.71
{Lycées, Vélos, Métros}	0.71	0.71
{Lycées, Cinémas, Vélos, Métros}	0.71	0.71
{Parcs, Lycées, Vélos, Cinémas}	0.56	0.44
{Parcs, Lycées, Cinémas, Métros}	0	0.14
...	...	...

Prévalences des co-localisations de Paris.

Résultats

Distribution



Données de Chicago avec les cliques maximales extraites.

Résultats

Motifs spatiaux

Co-localisation	Prévalence sous contrainte	Prévalence sans contrainte
{Bus, *Fast Food*, Lycées, Vélos}	0.58	0.5
{Bus, *Fast Food*, Lycées, Tramway, Vélos}	0.38	0.38
{Bus, *Fast Food*, Lycées}	0.33	0.17
{Bus, *Fast Food*, Lycées, Tramway}	0.3	0.3
{Bus, *Fast Food*, Lycées, Parcs}	0.17	0.17
{Bus, *Fast Food*, Lycées, Tramway, Vélos, Parcs}	0.15	0.15
...	...	...

Prévalences des co-localisations de Chicago.

Conclusion

- CSS-Miner : Une approche d'extraction de co-localisations prenant en compte la structure spatiale de la zone étudiée ;

Conclusion

- CSS-Miner : Une approche d'extraction de co-localisations prenant en compte la structure spatiale de la zone étudiée ;
- Utilisation du réseau routier comme structure spatiale et de la longueur du plus court chemin comme relation de voisinage ;

Conclusion

- CSS-Miner : Une approche d'extraction de co-localisations prenant en compte la structure spatiale de la zone étudiée ;
- Utilisation du réseau routier comme structure spatiale et de la longueur du plus court chemin comme relation de voisinage ;
- Application sur des données réelles.

Conclusion

- CSS-Miner : Une approche d'extraction de co-localisations prenant en compte la structure spatiale de la zone étudiée ;
- Utilisation du réseau routier comme structure spatiale et de la longueur du plus court chemin comme relation de voisinage ;
- Application sur des données réelles.

Perspectives

- Prendre en compte d'autres contraintes (par exemple l'altitude/pente, le parcours du chemin en voiture/vélo) ;

Conclusion

- CSS-Miner : Une approche d'extraction de co-localisations prenant en compte la structure spatiale de la zone étudiée ;
- Utilisation du réseau routier comme structure spatiale et de la longueur du plus court chemin comme relation de voisinage ;
- Application sur des données réelles.

Perspectives

- Prendre en compte d'autres contraintes (par exemple l'altitude/pente, le parcours du chemin en voiture/vélo) ;
- Intégrer la connaissance d'experts-métier [Flo+15] pour cibler/filtrer les plus pertinents motifs spatiaux ;

Conclusion

- CSS-Miner : Une approche d'extraction de co-localisations prenant en compte la structure spatiale de la zone étudiée ;
- Utilisation du réseau routier comme structure spatiale et de la longueur du plus court chemin comme relation de voisinage ;
- Application sur des données réelles.

Perspectives

- Prendre en compte d'autres contraintes (par exemple l'altitude/pente, le parcours du chemin en voiture/vélo) ;
- Intégrer la connaissance d'experts-métier [Flo+15] pour cibler/filtrer les plus pertinents motifs spatiaux ;
- Évaluer CSS-Miner sur de plus gros jeux de données.

Références

[BW19] Xuguang Bao et Lizhen Wang. “A clique-based approach for co-location pattern mining”. In : *Information Sciences* 490 (2019), p. 244-264.

[Boe17] Geoff Boeing. “OSMnx : New methods for acquiring, constructing, analyzing, and visualizing complex street networks”. In : *Computers, Environment and Urban Systems* 65 (2017), p. 126-139.

[Dij+59] Edsger W Dijkstra et al. “A note on two problems in connexion with graphs”. In : *Numerische mathematik* 1.1 (1959), p. 269-271.

[Flo+15] Frédéric Flouvat et al. “Domain-driven co-location mining : Extraction, visualization and integration in a GIS”. In : *Geoinformatica* 19 (2015), p. 147-183.

[HSX04] Yan Huang, Shashi Shekhar et Hui Xiong. “Discovering colocation patterns from spatial data sets : a general approach”. In : *IEEE Transactions on Knowledge and data engineering* 16.12 (2004), p. 1472-1485.

[Kim+14] Seung Kwan Kim et al. “A framework of spatial co-location pattern mining for ubiquitous GIS”. In : *Multimedia tools and applications* 71.1 (2014), p. 199-218.

[KH95] Krzysztof Koperski et Jiawei Han. “Discovery of spatial association rules in geographic information databases”. In : *International Symposium on Spatial Databases*. Springer. 1995, p. 47-66.

[KKK11] Seung Kwan Kim, Younghee Kim et Ungmo Kim. “Maximal cliques generating algorithm for spatial co-location pattern mining”. In : *Secure and Trust Computing, Data Management and Applications : 8th FIRA International Conference, STA 2011, Loutraki, Greece, June 28-30, 2011. Proceedings 8*. Springer. 2011, p. 241-250.

[SH01] Shashi Shekhar et Yan Huang. “Discovering spatial co-location patterns : A summary of results”. In : *International symposium on spatial and temporal databases*. Springer. 2001, p. 236-256.

[Tra+21] Vanha Tran et al. “MCHT : A maximal clique and hash table-based maximal prevalent co-location pattern mining algorithm”. In : *Expert Systems with Applications* 175 (2021), p. 114830.

Références

[VK11] Gintaras Vaira et Olga Kurasova. "Parallel bidirectional Dijkstra's shortest path algorithm". In : *Databases and Information Systems VI, Frontiers in Artificial Intelligence and Applications* 224 (2011), p. 422-435.

[WBL09] Lizhen Wang, Yuzhen Bao et Zhongyu Lu. "Efficient discovery of spatial co-location patterns using the iCPI-tree". In : *The Open Information Systems Journal* 3.1 (2009).

[Yao+16] Xiaojing Yao et al. "A fast space-saving algorithm for maximal co-location pattern mining". In : *Expert Systems with Applications* 63 (2016), p. 310-323.

[YS06] Jin Soung Yoo et Shashi Shekhar. "A joinless approach for mining spatial colocation patterns". In : *IEEE Transactions on Knowledge and Data Engineering* 18.10 (2006), p. 1323-1337.

MERCI DE VOTRE ATTENTION

Rodrigue Govan[1], Nazha Selmaoui-Folcher[1],
Aristotelis Giannakos[2], Philippe Fournier-Viger[3]

CNIA'23 – Strasbourg (France), le 4 Juillet 2023

{rodrigue.govan,nazha.selmaoui}@unc.nc