Prototype SOLAP appliqué sur des champs continus en mode raster

Analyse de hot spots de criminalité

Jean-Paul Kasprzyk[1]

1. Unité de Géomatique, Université de Liège
17 Allée du 6 Août, B4000 Liège, Belgique
jp.kasprzyk@ulg.ac.be

RESUME. *Les outils SOLAP (Spatial OnLine Analytical Processing) sont des serveurs permettant l'analyse rapide de données archivées dans un entrepôt de données à des fins décisionnelles. Actuellement, la plupart des solutions SOLAP ne gèrent les données spatiales qu'à travers le mode vecteur. Cependant, certaines techniques de visualisation transforment les données vectorielles en champs continus raster par interpolation. C'est notamment le cas en cartographie de la criminalité où un algorithme d'estimation de la densité par la méthode du noyau (« Kernel Density Estimation », ou KDE) est utilisé sur des délits ponctuels pour produire des cartes de hot spots. Ces cartes offrent ainsi une vision spatialement continue de la criminalité. Le but de cette recherche est de combiner les techniques du SOLAP et des cartes de hot spots. Un modèle de données multidimensionnel raster adapté aux champs continus est d'abord établi. Ensuite, ce modèle est appliqué aux KDE pour servir de base au prototype présenté ici avec des données de criminalité londoniennes.*

ABSTRACT. *SOLAP (Spatial OnLine Analytical Processing) is a server which allows decision makers to quickly analyze archived data from a spatial datawarehouse. Until now, most of SOLAP tools only manage spatial data through the vector mode. However, some visualization techniques use interpolation to transform them into continuous fields in raster mode. Crime hotspot mapping is one of them: data are modeled in raster with a KDE (Kernel Density Estimation) algorithm to offer a spatially continuous visualization of crimes. This research aims at combining hotspot mapping and SOLAP. First, we adapt SOLAP to continuous fields with a raster multidimensional data model. Then the raster model is adapted to KDE. The data model is validated with a prototype including London crime data.*

MOTS-CLES : *Estimation de la densité par la méthode du noyau, entrepôt de données, SIG, cartographie criminelle, mode maillé.*

KEYWORDS: *Kernel Denisty Estimation, data warehouse, GIS, crime mapping, continuous field*

1. Introduction

1.1. Question de recherche

La quantité de données spatiales est de plus en plus importante compte tenu de l'évolution des techniques d'acquisition et de stockage de ces dernières. Par conséquent, retirer efficacement de l'information hors de ces données à des fins décisionnelles ne peut se faire sans outils adaptés. Les outils SOLAP (« Spatial OnLine Analytical Processing ») répondent à cette demande.

Des données archivées dans un entrepôt de données peuvent ainsi être rapidement agrégées et visualisées à l'aide d'une interface tabulaire, graphique ou cartographique. La structure multidimensionnelle d'un entrepôt de données est constituée de mesures (exemple : chiffre d'affaire) associées à des dimensions (exemple : espace, temps, type de produit). Une dimension peut contenir plusieurs niveaux de hiérarchie constitués de membres (exemple : les communes de Belgique, les provinces de Belgique). Une mesure associée à des membres de dimension constitue un fait (exemple : chiffre d'affaire des produits vestimentaires de juillet 2014 en province de Liège).

Le SOLAP a d'abord été conçu pour la gestion des données spatialement discrètes. L'exploitation des données spatialement continues (pollution, climatologie, océanographie, *etc*) constituent un domaine de recherche dans lequel s'inscrit ce travail. Alors que la plupart des solutions SOLAP actuelles exploitent des données spatiales vectorielles (adaptées aux données discrètes), le prototype présenté ici se base sur des données maillées (ou raster). Comme nous le verrons lors de la démonstration, cette modélisation autorise non seulement l'analyse de phénomènes spatialement continus, mais offre également une vision continue des données naturellement discrètes (comme des délits criminels) grâce à des techniques d'interpolations bien connues du mode raster.

1.2. Etat de l'art

Alors que les entrepôts de données (Kimball, 1996) et les SOLAP (Bédard, 1997) émergent au milieu des années 90, les travaux combinant information spatialement continue et SOLAP datent d'une dizaine d'années. Ahmed et Miquel (2005) établissent un modèle multidimensionnel conceptuel exploitant le caractère continu de l'espace et du temps à travers des requêtes interpolant les données provenant d'un cube de données discret. La spatialité des objets reste néanmoins gérée par une structure vectorielle classique : une table des faits contient des mesures numériques reliées à des dimensions spatiales ou non-spatiales. Vaisman et Zimanyi (2009) introduisent le type « champ continu» pour un objet géographique. Celui-ci peut être considéré à la fois comme mesure continue dans la table des faits (liée à une fonction dépendant du temps et de l'espace) ou comme dimension éventuellement associée à une hiérarchie. Une mesure dans l'espace continu peut ainsi être calculée à la volée grâce à diverses méthodes d'interpolation spatiale (TIN, bilinéaire sur une grille, *etc*). Objets géographiques discrets et champs continus peuvent également coexister dans un même modèle (Bimonte et Kang, 2010). Les

mesures « champs continus » sont alors hiérarchisées en fonction d'une dimension spatiale discrète à laquelle s'adapte la résolution du champ. Les différents niveaux de généralisation cartographique ainsi obtenus sont calculés par interpolation et les agrégations sont effectuées au moyen d'opérations issues de l'algèbre de carte (Tomlin, 1983 ; Gomez *et al*, 2012). Certains modèles logiques considèrent la structure maillée dans un entrepôt de données en mode vecteur. Celle-ci peut alors être utilisée pour définir la mesure (McHugh, 2008) ou la dimension (Miquel *et al*, 2002).

Les modèles SOLAP continus proposés dans la littérature se distinguent donc à plusieurs niveaux : considération de la spatialité comme mesure ou comme dimension, interpolation, utilisation de l'algèbre de carte, type d'implémentation d'un champ continu, niveau conceptuel ou logique. Alors que plusieurs prototypes exploitent le mode vectoriel, le mode raster, pourtant bien adapté à l'information spatialement continue par définition, semble assez peu implémenté. Cela résulte probablement du fait que la plupart des solutions existantes pour la programmation SOLAP ne considèrent que le mode vecteur (Intelli3, 2014 ; Spatialytics, 2014).

De plus, les données spatialement continues et le format raster ne se limitent pas qu'aux phénomènes naturellement continus. Dans le domaine de la cartographie de la criminalité, les délits (données ponctuelles) sont fréquemment convertis en champs continus raster pour produire des cartes de chaleur (hot spots). Ces dernières, généralement obtenues par interpolation grâce un algorithme KDE (« Kernel Density Estimation »), s'avèrent très efficaces pour la visualisation et la prédiction de la criminalité (Chainey *et al*, 2008). Un KDE produit un raster où chaque pixel contient une valeur dépendant du nombre de crimes et de leur proximité par rapport au centre du pixel dans une fenêtre de lissage. Le résultat dépend donc de trois paramètres : la résolution du pixel (1), la taille de la fenêtre de lissage (2) et la fonction KDE (3). Pour une même fonction KDE, une variation de la taille de la fenêtre produit des hot spots plus ou moins étendus et nombreux, donc adaptés à différentes échelles d'analyse (Di Salvo *et al*, 2005). En revanche, il a été démontré que la résolution a peu d'influence sur le caractère prédictif des hot spots de criminalité (Chainey, 2013). Notons que par « criminalité », nous faisons référence au terme anglais qui inclut à la fois les crimes et les délits.

1.3. Contribution

Le prototype proposé ici est un SOLAP exploitant des champs continus en mode raster. Le raster est considéré comme mesure dans la table des faits et est agrégé en fonction du temps et du type de crime au moyen d'opérations issues de l'algèbre de carte. Les données de démonstration, provenant du domaine de la cartographie criminelle, sont des délits ponctuels pré-agrégés et analysés sous forme de KDE (champ continu). L'interface utilisateur est composée de cartes de hot pots associées à des graphiques. Puisque la résolution d'un KDE a peu d'influence sur la pertinence des hot spots, les requêtes sont optimisées par la manipulation de grilles rasters relativement petites (de l'ordre de 150x150 pixels) : basse résolution sur un grand territoire ou haute résolution sur un petit territoire. Cette modélisation raster permet d'inclure plusieurs fonctionnalités spécifiques dans l'analyse multidimensionnelle :

notamment la vision continue de l'espace géographique, la généralisation des données spatiales en fonction de l'échelle d'analyse ou encore l'application de filtres spatiaux créés à la volée. Notons qu'une approche SIG-dominant a été adoptée plutôt qu'une approche OLAP-dominant ou OLAP-SIG intégrés (Bédard, 2010). En effet, les SOLAP des deux autres catégories intègrent la table de pivot[1] qui est inexploitable dans notre cas car les mesures traitées sont exclusivement au format raster et non alphanumérique.

La présentation du prototype SOLAP raster est structurée en deux parties : sa réalisation depuis la modélisation multidimensionnelle raster jusqu'à la description des éléments constituant son architecture, et sa démonstration à travers quelques exemples de requêtes appliquées sur des données de criminalité concernant la ville de Londres. Les avantages de ce type d'outil seront ainsi illustrés. En guise de conclusion, nous nous pencherons sur d'autres applications potentielles du SOLAP raster, ainsi que les bénéfices de son association avec un SOLAP vecteur.

2. Réalisation

2.1. Modélisation

Le modèle multidimensionnel proposé dans le cadre du prototype est un simple schéma en étoile (une table des faits reliée à des dimensions). La seule différence par rapport à un modèle OLAP-type réside dans la définition de la mesure. Le domaine d'un attribut « mesure raster » (dans la table des faits) est constitué de rasters avec une taille fixée (même nombre de rangées et de colonnes) et une géoréférenciation identique. Par conséquent, tous les rasters d'un attribut « mesure raster » partagent la même emprise spatiale et la même résolution.

Puisque les données utilisées sont pré-agrégées par une méthode d'interpolation particulière (KDE), le domaine d'une mesure doit être restreint d'avantage. Ainsi, le domaine d'un attribut « mesure KDE » est un sous-ensemble du domaine d'un attribut « mesure raster » constitué de KDE aux paramètres identiques (résolution, fenêtre de lissage et fonction KDE). Ainsi, la somme de deux KDE A et B par opération locale d'algèbre de carte est égale à un KDE réalisé à partir de l'union des deux nuages de points utilisés respectivement dans le calcul des KDE A et B. Cela se démontre par associativité et commutativité de la somme.

$$K(P_a \cup P_b) = K(P_a) + K(P_b) \tag{1}$$

Lorsque la table des faits comprend plusieurs attributs « mesure raster » portant sur le même fait, mais avec des domaines différents (résolutions / fenêtres de lissage et emprises spatiales), des opérations de forage spatial telles que « spatial roll up » ou « spatial drill down » sont possibles en passant d'une mesure raster à l'autre d'un

[1] tableau interactif permettant la navigation dans un cube de données et l'affichage de mesures alphanumériques

point de vue relationnel, mais en passant d'un niveau spatial à l'autre d'un point de vue multidimensionnel. En effet, puisque la dimension spatiale continue est implicitement incluse dans la mesure, il est logique que ses différents niveaux soient gérés par cette dernière. Cela se traduit par une variation de la taille et du nombre de hot spots lorsqu'il s'agit de mesures KDE.

2.2. Architecture

L'architecture du prototype est constituée d'outils « Open Source ». L'ETL (« Extract, Transform, Load ») GeoKettle intègre les données dans l'entrepôt géré par PostGreSQL et son extension spatiale PostGIS (pour la gestion des données raster). Le cube de données SOLAP est programmé en PHP côté serveur. Il calcule les requêtes de l'utilisateur et les renvoie au serveur de données spatiales MapServer. Enfin, le côté client est constitué de cartes, de graphiques et d'un arbre à dimensions programmés en JavaScript avec les librairies OpenLayers et DHTMLX.

3. Démonstration

3.1. Données

Les données londoniennes utilisées proviennent de la « City of London Police » et de la « Metropolitan Police » (UK Police, 2014). Elles sont limitées à l'année 2012 dans le cadre de la démonstration. Leur précision géométrique est au niveau de la rue et leur précision temporelle au niveau du mois. Elles sont classées selon 11 types de délit. Environ 1 200 000 délits ont ainsi été pré-agrégés sous forme de KDE dans la table des faits. Celle-ci compte donc 132 faits liés à deux dimensions (temps et type de crime) et à une mesure KDE. En effet, 12 mois x 11 types de délit = 132 faits. Une résolution de 300 mètres et une fenêtre de lissage de 1500 mètres ont été utilisées avec une fonction KDE quartique pour la transformation des données en champs continus.

3.2. Exemples de requêtes

Cette section dresse un aperçu du prototype SOLAP raster à travers quelques exemples d'opérations. En accord avec les performances de temps de réponse exigées par la définition d'un SOLAP, toutes les requêtes sont exécutées en moins de cinq secondes via une interface « point & click » (ne nécessitant pas de connaissance en programmation). Le prototype est accessible en ligne via l'url suivante : http://nolap01.ulg.ac.be/demo

L'utilisateur a la possibilité d'agréger les KDE selon plusieurs méthodes (somme, moyenne, écart-type, minimum et maximum) et en fonction de membres choisis dans un arbre à dimensions. La figure 1 montre deux opérations d'agrégation globale du cube selon la somme pour la périphérie (figure 1a) et pour le centre de Londres (figure 1b). Il s'agit d'opérations locales d'algèbre de carte entre les 132 faits qui peuplent l'entrepôt. Le résultat est une carte de hot spots couvrant tous les types de délit pour l'année 2012. L'unique différence entre les deux opérations

réside dans le filtre spatial digitalisé par l'utilisateur : un pour la périphérie et l'autre pour le centre. Après rasterization, le filtre est appliqué sur le résultat de la requête OLAP en le considérant comme un masque binaire. Les légendes sont de simples étirements linéaires basés sur la valeur maximale du raster associé. Elles nous indiquent que la criminalité est globalement plus élevée au centre-ville par rapport à la périphérie. Les graphiques montrent la distribution des types de délit en rapport avec la carte. Ceux-ci sont automatiquement calculés en effectuant la moyenne de tous les pixels du KDE résultant de l'agrégation des 12 mois de l'année pour chaque type de crime. Nous pouvons observer une distribution différente des types de délit à l'extérieur (figure 1a) et à l'intérieur (figure 1b) du centre-ville : prédominance de comportements antisociaux en périphérie et prédominance de vols au centre.

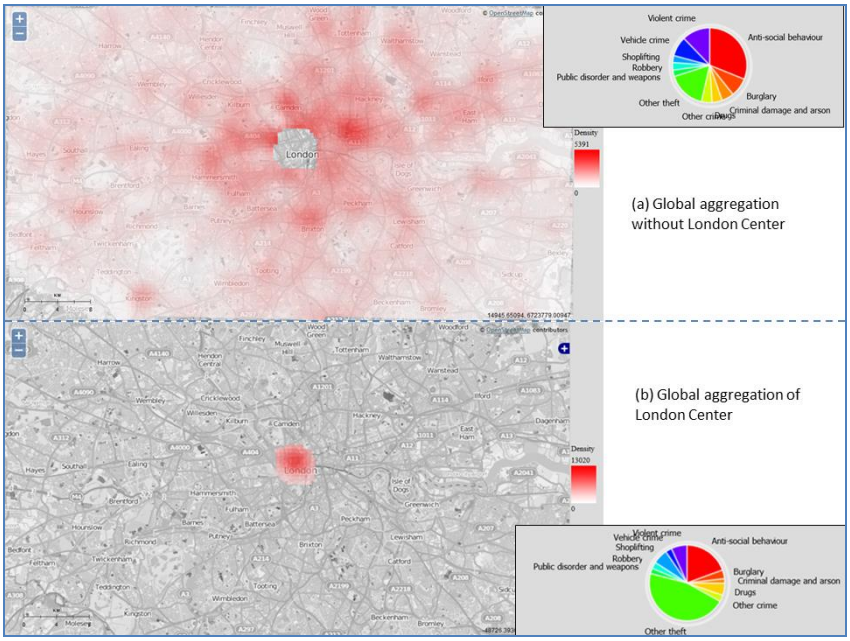

Figure 1. Comparaison entre le centre et la périphérie de Londres

Même si la somme et la moyenne sont certainement les fonctions d'agrégation les plus courantes dans les SOLAP, d'autres fonctions comme l'étendue et l'écart-type peuvent s'avérer très utiles dans notre outil, notamment pour détecter et délimiter des zones géographiques comprenant des changements dans la criminalité. Par exemple, en agrégeant les cambriolages de 2012 par l'étendue, la carte met en évidence les hot spots présentant un taux de cambriolages avec au moins une forte variation mensuelle (différence entre le maximum et le minimum pour chaque pixel). C'est le cas du quartier de Southall où un graphique spatialement filtré sur ce hot spot (agrégation par la moyenne en fonction du mois) montre effectivement une augmentation significative des cambriolages au mois de Juillet.

Bien que non illustrées ici, d'autres opérations sont réalisables grâce à la modélisation raster du SOLAP : production de cartes d'évolution par soustraction de rasters issus de requêtes OLAP (par exemple, évolution des cambriolages entre janvier et février), changement du niveau de généralisation cartographique par sélection d'une mesure différente (voir section 2.1), application d'une classification différente du raster (standardisation, quantiles), *etc.*

4. Conclusion

En dressant l'état de l'art portant sur l'intégration de l'information spatialement continue dans les systèmes SOLAP, nous avons pu constater que pratiquement aucun prototype n'avait été implémenté en utilisant le mode raster, pourtant nativement adapté à la modélisation de ce type de donnée. Nous avons également souligné le fait que le raster était aussi utilisé pour des traitements d'interpolation de données spatialement discrètes tels que la génération de cartes de hot spots de criminalité par KDE. En utilisant des concepts théoriques présents dans la littérature SOLAP continu (agrégation par algèbre de carte, champ continu considéré comme mesure dans la table des faits), nous avons établi un modèle multidimensionnel raster pour la gestion des champs continus. Il a ensuite été adapté au cas particulier des KDE pour le traitement des données de criminalité. Ainsi, le modèle KDE SOLAP a servi de base à l'implémentation d'un prototype (architecture « Open Source ») auquel nous avons intégré des données de criminalité londoniennes. Ces données pré-agrégées sous forme de champs continus grâce à l'interpolation KDE ont permis d'illustrer les avantages du SOLAP raster dans l'analyse de hot spots de criminalité : génération simple et rapide de cartes de hot spots via une approche multidimensionnelle, intégration de filtres spatiaux créés à la volée, génération de graphiques apportant des informations supplémentaires sur les hot spots, détection de changements dans la criminalité, *etc*.

Même si l'outil semble prouver son utilité dans le cadre de cette démonstration, d'autres tests sur différents jeux de données doivent encore être réalisés. Des données de criminalité avec des dimensions temporelles et « type de crime » plus précises permettraient d'enrichir l'analyse, mais aussi de tester les limites du système en agrégeant un plus grand nombre de rasters (le nombre de faits croît exponentiellement avec le nombre de dimensions). Le SOLAP raster devrait également être testé sur des données relevant de différents domaines (écologie, pollution, *etc*) avec d'autres méthodes d'interpolation que KDE.

Finalement, un outil SOLAP combinant données vecteur et raster pourrait tirer profit des avantages des deux types de modélisation : le raster pour l'analyse spatialement continue de grand jeux de données pré-agrégées, le vecteur pour l'analyse discrète à travers un grand nombre de dimensions, l'accès aux données individuelles et la navigation à travers des tables de pivot. Le SOLAP raster serait alors considéré comme une analyse complémentaire au vecteur dans une approche OLAP-SIG intégrés.

Bibliographie

Ahmed, T. O., Miquel, M. (2005). Multidimensional structures dedicated to continuous spatiotemporal phenomena. *Proceedings of the 22nd British National conference on Databases: enterprise, Skills and Innovation*, Springer-Verlag, Sunderland, UK**,** p. 29-40.

Bédard, Y. (1997). Spatial OLAP. Géomatique VI: Un monde accessible. *2ème Forum annuel sur la R-D*, Montréal, Canada.

Bédard, Y. (2010). Le géodécisionnel: origine, évolution, état de l'art, enjeux, R&D, École Nationale Supérieure des Mines de Paris – Centre de recherche sur les Risques et les Crises, Sophia-Antipolis, France

Bimonte, S., Kang, M.A. (2010). Towards a model for the multidimensional analysis of field data. *Proceedings of the 14th east European conference on Advances in databases and information systems (ADBIS'10)*, Catania B., Ivanovic M., Thalheim B. (Eds.). Springer-Verlag, Berlin, Allemagne, p. 58-72.

Chainey, S. P., Tompson, L., Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, vol. 21, p. 4-28.

Chainey, S. (2013). Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bulletin de la société géographique de Liège***,** vol. 60.

Di Salvo, M., Gadais, M., Roche-Woillez, G. (2005). L'estimation de la densité par la méthode du noyau: méthodes et outils, Certu, Lyon, France.

Gomez, L.I., Gomez, S.A., Vaisman, A. (2012). A generic data model and query language for spatiotemporal OLAP cube analysis. *Proceedings of the 15th International Conference on Extending Database Technology*, ACM, Berlin, Allemagne**,** p. 300-311.

Intelli3 (2014). Map4Decision, http://www.intelli3.com/map4decision

Kimball, R. (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, New York, USA.

McHugh, R. (2008). *Intégration de la structure matricielle dans les cubes spatiaux*. Mémoire en sciences géomatiques. Université Laval.

Miquel, M., Bédard, Y., Brisebois, A. (2002). Conception d'entrepôts de données géospatiales à partir de sources hétérogènes. Exemple d'application en foresterie. *Ingénierie des Systèmes d'Information*, vol. 7, n°3, p. 89-111.

Spatialytics (2014). http://www.spatialytics.com/

Tomlin, C. D. (1983). A Map Algebra. *Proceedings Harvard Computer Graphics Conference*, Cambridge, USA.

UK Police, 2014. Crime Map, http://www.police.uk

Vaisman, A., Zimanyi, E. (2009). A multidimensional model representing continuous fields in spatial data warehouses. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, USA. ACM: p. 168-177.