Quelques exemples d'améliorations des performances de traitement des données environnementales

François Pinet*

*Université Clermont Auvergne,
INRAE, UR TSCF,
Centre Clermont-Auvergne-Rhône-Alpes,
63178 Aubière, France
francois.pinet@inrae.fr,
www.inrae.fr/tscf

Résumé. Cette communication porte sur un résumé de travaux existants autour du traitement de données agricoles et environnementales. Il s'agit ici de montrer quelques exemples d'amélioration de performances que l'on peut obtenir dans les traitements de données géoréférencées. Nous nous concenterons sur des opérations que l'on retrouve dans les entrepôts de données.

1 Des techniques pour améliorer les performances

Dans cette communication, nous rappellerons quelques exemples d'optimisation pour les données environnementales de type raster. L'objectif est :

— de faire une synthèse des travaux de trois de nos articles déjà publiés (Kang et al., 2015; En-Nejjary et al., 2021; En-Nejjary et al., 2021) – il s'agit d'une continuité de travaux ;
— d'étendre la présentation que nous avons faite à l'atelier IoT d'Inforsid 2022, en insistant sur les applications dans le domaine de l'environnement et les entrepôts de données, en partageant ainsi nos résultats sur ce thème au sein de la communauté EDA.

Les rasters sont un mode de représentation courant dans le domaine des systèmes d'information géographique. Ces données sont présentées sous la forme de grilles régulières géoréférencées où chaque cellule est associée à une valeur. Le plus souvent, nous avons à traiter et analyser une séquence temporelle (c'est-à-dire une succession) de rasters portant sur un même territoire d'étude.

Le but de la présentation sera donc, dans un premier temps, de montrer quelques techniques existantes pour améliorer les temps de traitement de ce type de données. Au cours de la présentation, nous nous concentrerons sur les opérations de sélection et d'agrégation de rasters. Il s'agit d'opérations classiques dans le cadre des entrepôts de données par exemple, mais nécessitant des adaptations sachant le format de données ciblés.

Dans une large collection de rasters portant sur un même territoire, l'opération de sélection permet de choisir automatiquement un sous-ensemble de rasters respectant une condition donnée par l'utilisateur. Ceci permet par exemple de rechercher les rasters représentant un phénomène extrême (par exemple, un pic local dans les valeurs). L'opération d'agrégation permet de résumer un sous-ensemble de rasters, afin d'avoir une vue synthétique de ce sous-ensemble.

Il peut s'agir par exemple de résumer un ensemble des données climatiques sur une période donnée ou sur des plages horaires données - ce résumé permet de mettre en évidence une tendance particulière. Nous traiterons aussi le cas particuliers de l'agrégation de séquences temporelles de rasters qui se chevauchent.

Des méthodes seront présentées pour améliorer les temps de calcul des opérations, que cela soit en produisant un résultat exact, qu'en proposant un résultat estimé. Nous montrerons aussi que certaines des méthodes peuvent être appliquées à des données issues de réseaux de capteurs avant la production des rasters (c'est-à-dire avant l'interpolation spatiale des valeurs).

Les applications présentées seront le suivi des informations environnementales notamment sur des parcelles agricoles (telles que l'humidité du sol et la température) à partir d'un réseau de capteurs, ainsi que le suivi de la qualité de l'air. L'agriculture est de plus en plus un champ d'expérimentation pour les nouvelles technologies de l'informatique (Papajorgji et al., 2010). Les systèmes d'information spatialisées sont un champ de recherche important, au sein duquel le traitement des données est essentiel - voir par exemple (Bimonte, 2007; Laurini et Thompson, 1992; Laurini, 2001; Malinowski et ZimÃ¡nyi, 2008; Kang et al., 2004; Boulil et al., 2012; Bédard et al., 2007; Damiani et Spaccapietra).

2 Quelques détails sur les techniques

La figure 1 présente le cadre général des travaux (Kang et al., 2015; En-Nejjary et al., 2021; En-Nejjary et al., 2021). Dans un très large ensemble de rasters, on souhaite sélectionner un sous-ensemble (une séquence), et effectuer des opérations dessus, de type agrégation ou sélection. Les traitements peuvent porter sur une zone d'intérêt au sein des rasters (ou bien sur la totalité de chaque raster).

Prenons le cas de l'agrégation. Supposons que l'on souhaite agréger un ensemble de rasters représentant la mesure d'un phénomène spatial continue sur une région donnée. Il peut s'agir de l'humidité du sol, la température, la qualité de l'air, etc. Chaque raster correspond à une mesure sur la zone à un instant t. On souhaite agréger les rasters et produire un seul raster résumé à partir d'un ensemble de rasters. L'agrégation pourrait par exemple se faire sur une dimension temporelle (agrégation des rasters de chaque matinée, ou soir, de chaque jour, semaine, mois, etc.). Cette agrégation peut être vue comme un calcul matriciel (par ex. des sommes ou moyennes de matrices - chaque raster étant vu comme une matrice). La figure 2 montre une amélioration possible des performances, par une estimation du résultat de l'agrégation (Kang et al., 2015). Il s'agit de faire un pré-traitement dans la base de rasters (une fois pour toute). On va regrouper les rasters en cluster selon leur degré de similarité. Les rasters du même cluster sont considérées comme ayant des valeurs proches. Au moment de l'agrégation, plutôt que calculer l'agrégation sur tous les rasters, on va choisir un (seul) raster par cluster (en pondérant son poids dans le calcul). Ainsi, si on veut agréger les 8 rasters de la Figure 1, on peut n'agréger que les 3 rasters entourés sur la figure, en pondérant les valeurs de chaque cellules des rasters par le nombre total de rasters dans son cluster ; on ne prend qu'un seul raster par cluster. Cette heuristique se justifie par le fait qu'on considère que les rasters qui sont dans le même cluster ont des valeurs proches. Donc, elles peuvent être substituées pour estimer le résultat des calculs d'agrégations.

Au cours de la présentation, d'autres exemples d'heuristiques seront décrits pour l'agrégation et la sélection des rasters.

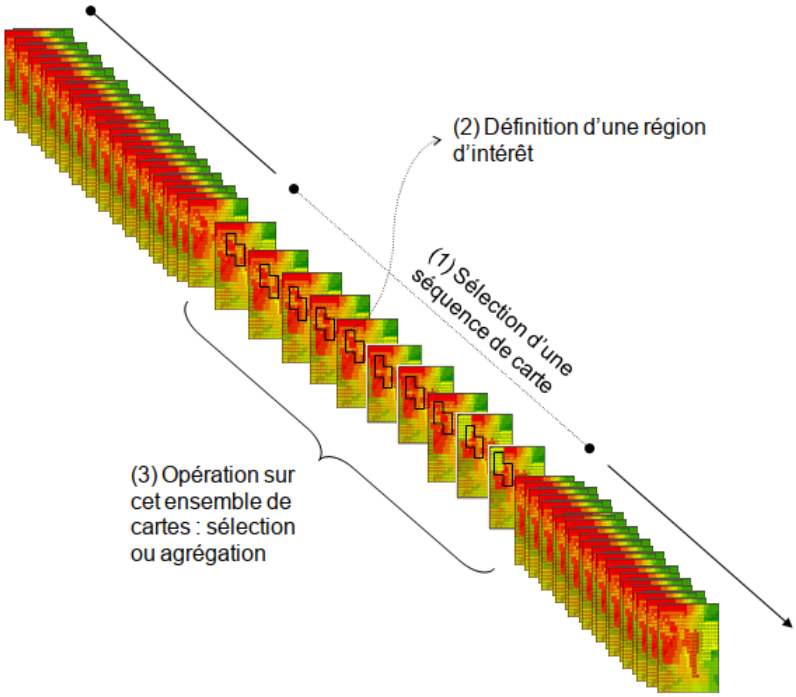

FIG. 1 – *Opérations ciblées (Kang et al., 2015; En-Nejjary et al., 2021; En-Nejjary et al., 2021)*

Quelques exemples d'améliorations des performances de traitement

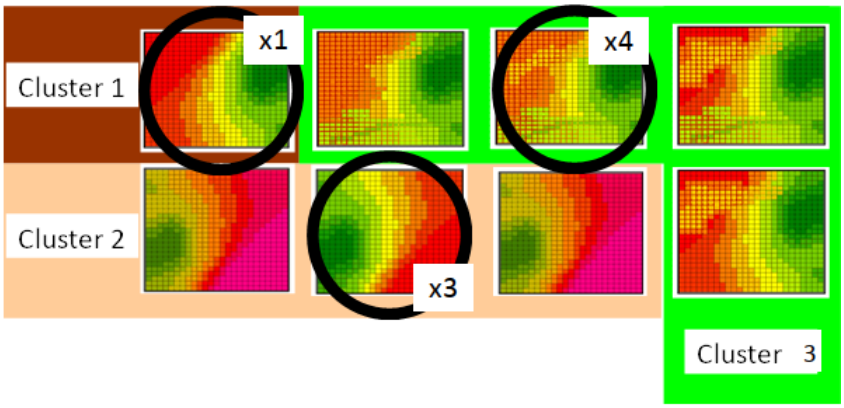

FIG. 2 – *Définition de clusters pour améliorer les agrégations*

Références

Bimonte, S. (2007). *Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation*. Ph. D. thesis. Thèse de doctorat dirigée par Laurini, Robert et Tchounikine, Anne Informatique Lyon, INSA 2007.

Boulil, K., S. Bimonte, et F. Pinet (2012). A UML & spatial OCL based approach for handling quality issues in SOLAP systems. In L. A. Maciaszek, A. Cuzzocrea, et J. Cordeiro (Eds.), *ICEIS 2012 - Proceedings of the 14th International Conference on Enterprise Information Systems, Volume 1, Wroclaw, Poland, 28 June - 1 July, 2012*, pp. 99–104. SciTePress.

Bédard, Y., E. Bernier, T. Badard, N. Chrisman, S. Roche, G. Edwards, M. Mostafavi, J. Pouliot, M. Gervais, F. Hubert, S. Larrivée, M.-J. Proulx, S. Rivest, M. Nadeau, E. Dubé, J. Brodeur, et R. Devillers (2007). Research in geographic information systems, data management and dissemination, and new geospatial technologies. *GEOMATICA 61*(3), 288–314.

Damiani, M. L. et S. Spaccapietra. Spatial data warehouse modelling. In *Processing and Managing Complex Data for Decision Support*.

En-Nejjary, D., F. Pinet, et M. Kang (2021). Spatial data sequence selection based on a user-defined condition using GPGPU. *ISPRS Int. J. Geo Inf. 10*(12), 816.

En-Nejjary, D., F. Pinet, et M.-A. Kang (2021). Spatial data sequence selection based on a user-defined condition using gpgpu. *ISPRS International Journal of Geo-Information 10*(12).

Kang, M., F. Pinet, M. Schneider, J.-P. Chanet, et F. Vigier (2004). How to design geographic database ? Specific UML profile and spatial OCL applied to wireless Ad Hoc networks. In *7th Conference on Geographic Information Science (AGILE'2004), Heraklion, GRC, April 29-May 1 2004*, pp. 289–299.

Kang, M., M. Zaamoune, F. Pinet, S. Bimonte, et P. Beaune (2015). Performance optimization of grid aggregation in spatial data warehouses. *Int. J. Digit. Earth 8*(12), 970–988.

Laurini, R. (2001). *Information Systems for Urban Planning : A Hypermedia Cooperative Approach*.

Laurini, R. et D. Thompson (1992). *Fundamentals of spatial information systems*.

Malinowski, E. et E. ZimÃ¡nyi (2008). *Advanced Data Warehouse Design : From Conventional to Spatial and Temporal Applications*.

Papajorgji, P., F. Pinet, A. Miralles, E. Jallas, et P. M. Pardalos (2010). Modeling : A central activity for flexible information systems development in agriculture and environment. *Int. J. Agric. Environ. Inf. Syst. 1*(1), 1–25.

Summary

This communication focuses on an overview of existing work on the processing of agricultural and environmental data. It shows some examples of performance improvement that can be obtained in the processing of georeferenced data. We focus on operations usually found in data warehouses.