

Vers un entrepôt de données pour le trafic routier

C. Bauzer-Medeiros⁽¹⁾, O. Carles⁽²⁾, F. Devuyst⁽³⁾, B. Hugueney⁽⁵⁾, M. Joliveau⁽³⁾,
G. Jomier⁽⁵⁾, M. Manouvrier⁽⁵⁾, Y. Naïja⁽⁵⁾, G. Scemama⁽²⁾, L. Steffan⁽⁵⁾

⁽¹⁾Institute of Computing (IC) - University of Campinas
Caixa Postal 6176 13084-971 Campinas, SP - Brazil
cmbm@ic.unicamp.br

⁽²⁾INRETS - Laboratoire GRECIA
2, Avenue du Général Malleret Joinville 94114 Arcueil Cedex
{carles, scemama}@inrets.fr,

⁽³⁾Ecole Centrale de Paris - Laboratoire MAS
Grande Voie des Vignes F-92 295 Châteney-Malabry Cedex
{florian.de-vuyst, marc.joliveau}@ecp.fr

⁽⁵⁾Université Paris Dauphine - Laboratoire LAMSADE
Place du Maréchal de Lattre de Tassigny 75775 Paris Cedex 16
{hugueney, jomier, manouvrier, naija, steffan}@lamsade.dauphine.fr

Résumé. Cet article présente la démarche multidisciplinaire que nous avons adoptée pour construire un système d'information pour l'aide à la décision dans la gestion du trafic routier. L'architecture du système, le schéma de l'entrepôt de données ainsi que les différentes représentations numériques et symboliques des séquences spatio-temporelles, stockées dans l'entrepôt, y sont détaillés.

1 Introduction

Les entrepôts de données, originellement utilisés dans le domaine du commerce et de la gestion, commencent à l'être dans d'autres domaines, comme par exemple les applications scientifiques. La possibilité de faire des analyses et des corrélations sur des agrégations créées dynamiquement à partir de plusieurs dimensions est un des avantages offerts à ces domaines par cette modalité d'organisation de données.

Cet article aborde les problèmes associés à un de ces nouveaux domaines d'application, celui des données spatio-temporelles, mesurées en temps réel, où les sources des données correspondent à des centaines de capteurs qui enregistrent périodiquement des mesures sur des phénomènes spécifiques. La localisation spatiale de ces capteurs, la périodicité des prises de valeurs et la variation de ces valeurs dans le temps et l'espace représentent quelques unes des variables qui doivent être prises en compte pour extraire l'information souhaitée.

En particulier, cet article concerne les problèmes de gestion du trafic routier urbain, où les capteurs sont placés sur les axes routiers d'une ville et transmettent constamment à des centrales de données de mesures sur la circulation à chaque intervalle de temps. L'analyse de ces données constitue une base pour les prises de décisions sur la circulation à plusieurs niveaux, allant de la macro-échelle (par exemple, la construction d'un viaduc), jusqu'aux décisions ponctuelles (par exemple l'ajustement d'un feu de

circulation). Le tracé des lignes d'autobus, la spécification de voies réservées, la logistique de la sécurité dans un quartier lors d'un événement qui attire la population sont quelques exemples des décisions d'action qui peuvent être prises à partir de l'analyse de ces données.

Il existe déjà beaucoup d'outils développés pour aider cette analyse. Les recherches de l'INRETS (Institut National de REcherche sur les Transports et leur Sécurité) ont conduit à l'élaboration de tels outils [Scemama *et al.*, 2000, Scemama et Carles, 2004]. D'une manière générale, ces outils produisent des rapports sur des périodes (jour, semaine...) et traduisent des états du trafic. Un observatoire permet déjà d'explorer les données suivant plusieurs dimensions : le temps, l'espace des coordonnées géographiques, les entités du réseau de transport (tronçon de route, itinéraire), les indicateurs, les éléments du diagnostic (perturbation, événements). Plus précisément, il permet de visualiser la localisation des principaux points noirs sur le réseau, de consulter et de comparer les courbes des divers indicateurs de trafic. Il s'agit d'un outil de consultation Web accessible à un nombre quelconque d'utilisateurs. Cependant, il repose sur un archivage journalier sous forme de fichiers rendant cet outil peu adapté à des traitements et requêtes complexes tels qu'ils apparaissent dans les outils d'entreposage. Ceci justifie l'intérêt pour les entrepôts de données et plus particulièrement les entrepôts de données spatiales qui couplent les fonctionnalités d'un Système d'Information Géographique (SIG) à celle d'un entrepôt OLAP (*On-Line Analytic Processing*).

Cet article présente la démarche engagée pour mener des travaux qui combinent, d'un côté, les résultats de la recherche sur les entrepôts de données spatiales avec ceux qui concernent les séries temporelles et, de l'autre, les résultats de la recherche en modélisation mathématique et en fouille de données. Cette démarche, de nature multidisciplinaire, est développée dans le cadre du projet CADDY (*Contrôle de l'Acquisition de Données temporelles massives, stockage et modèles DYnamiques*) de l'ACI "Masse de Données". Son originalité provient de cette coopération, où les résultats de plusieurs domaines sont combinés pour produire un système d'information pour l'aide à la décision dans la gestion du trafic routier.

L'article est organisé comme suit. La section 2 présente une description générale de l'architecture du système d'information en cours de construction, basé sur un entrepôt de données spatiales du trafic routier. Cette description permet d'établir un cadre général pour la suite de l'article. La section 3 présente le domaine du trafic routier et les premières démarches suivies pour élaborer l'entrepôt de données. La section 4 décrit l'aspect multi-représentation des données, qui exploite les plusieurs dimensions introduites par l'entrepôt. La section 5 présente l'état de l'implémentation du système. La section 6 présente l'état de l'art sur la gestion des données pour le trafic routier et les entrepôts dans ce domaine. La section 7 présente les conclusions de l'article.

2 Architecture du système d'information

Le projet CADDY, démarré en juillet 2003, fait intervenir des aspects de modélisation mathématique, de statistiques, d'analyse de données et de bases de données. Il a pour axe de recherche principal l'étude de méthodes d'acquisition, de représentation et de recherche des séquences temporelles (*time series*) multivariées interdépendantes. L'ob-

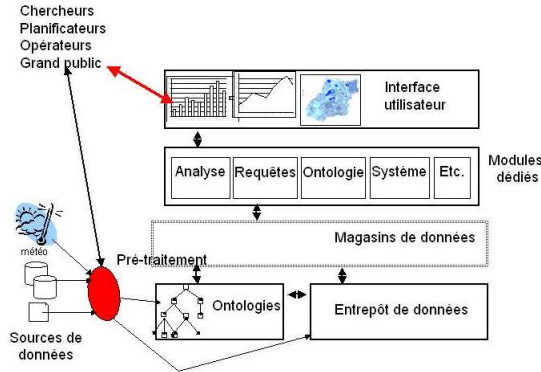


FIG. 1 – Architecture du système de trafic routier.

jectif est de proposer une méthodologie générale permettant d'améliorer le fonctionnement des systèmes complexes manipulant ce type de données. Pour bien vérifier l'applicabilité des modèles développés et tester les prototypes implantés, une collaboration avec l'INRETS a été établie sur le domaine des systèmes du trafic routier. Grâce à cette coopération, il a été possible aux membres du projet CADDY de disposer d'un grand volume de données spatio-temporelles, collectées par des centaines de capteurs routiers dans une grande ville française.

La figure 1 montre, de façon schématique, l'architecture du système d'information qui en cours de développement au sein de CADDY. Sa présentation en couches permet une meilleure compréhension de ses fonctionnalités : *Stockage*, *Modules dédiés*, *Interface utilisateur*. Le but de ce système est de permettre à ses utilisateurs de prendre des décisions, à court et à long terme, sur la gestion du trafic routier et l'aménagement des routes.

Comme le montre la figure 1, plusieurs sources de données sont considérées : les données spatio-temporelles fournies par les capteurs (des séquences journalières de débit et de taux d'occupation, associées au réseau routier - détaillées dans la section 3.1) ; des annotations textuelles et des fichiers de documentation ; des données géographiques (surtout associées aux tracés urbains) ; des informations météorologiques (compte tenu de l'influence des grandes perturbations climatiques sur les écoulements dans les réseaux routiers). Les sources doivent être nettoyées, puis stockées dans un *Entrepôt de Données* qui appartient à la couche *Stockage*.

Au niveau *Stockage*, on trouve trois composants principaux : l'entrepôt, des magasins de données et un ensemble d'ontologies. La création et la maintenance de l'entrepôt forment une base pour le système d'information ; ses détails sont présentés dans la section 3.2. L'ensemble d'ontologies permet d'organiser les définitions et la terminologie utilisées par le domaine de l'application et aide également à la construction et à la maintenance de l'entrepôt. Ontologies est au pluriel car il s'agit de plusieurs dimensions ou axes de connaissance, qui sont fournis par différentes demandes et profils d'utilisateurs. Ces dimensions doivent être intégrées en utilisant des méthodes proposées dans la

littérature [Wache *et al.*, 2001]. De plus, les ontologies sont utilisées pour le traitement des requêtes posées à l'entrepôt et pour la création et la maintenance des *magasins de données* (*DataMarts* en anglais) qui contiennent des "vues matérialisées" sur l'entrepôt. Ces magasins sont créés au fur et à mesure de l'apparition de nouvelles demandes sur des sous-ensembles des données stockées.

La couche *Interface utilisateur* est responsable de tous les types d'interaction d'un utilisateur avec le système. Il est prévu de réaliser différentes formes de visualisations interactives pour un ensemble de données. A présent, le système dispose déjà de visualisations sous forme de carte (d'une ville), de table de données, et de courbes qui montrent l'évolution temporelle du trafic pour un ou plusieurs capteurs (voir section 5).

La couche *Modules Dédiés* contient un ensemble de composants qui interviennent dans le traitement des demandes des utilisateurs. Elle aide à transformer une interaction de l'utilisateur au niveau *Interface* dans un ensemble d'accès à la couche *Stockage*. Elle est aussi responsable du traitement des données récupérées depuis la couche *Stockage* et de leur transformation en vue de leur visualisation par l'*Interface*.

Parmi ces modules, on peut mettre l'accent sur ceux qui sont responsables de l'analyse des données des capteurs, du traitement des requêtes, de l'organisation et de la maintenance des ontologies et des interventions au niveau système. Les modules d'*Analyse* aident, entre autres, à la recherche de motifs dans les séries spatio-temporelles des données de capteurs (voir [Hugueney, 2003] pour plus de détails). Le module *Requêtes* a pour buts : la traduction de termes, l'intégration des résultats partiels des requêtes envoyées à l'entrepôt et aux magasins et la correspondance entre le niveau *Interface* et le niveau *Stockage*. Le module *Système* est responsable des opérations de maintenance du système, qui sont utilisées exclusivement par les opérateurs du système d'information (par exemple, pour le nettoyage des données ou la sauvegarde). Le module *Ontologie* concerne la construction et la mise à jour des ontologies, en utilisant des outils disponibles dans le domaine public. Nous prévoyons à cet effet d'utiliser Protégé[Pro, 2006], compte tenu de l'expérience de membres de l'équipe dans ce domaine.

Les utilisateurs de ce système, comme le montre la figure 1, sont principalement des décideurs et des chercheurs dans le domaine routier. Parmi les décideurs, sont particulièrement pris en compte les experts des compagnies de transport public ainsi que les hommes politiques. Le système doit évidemment être aussi accessible aux opérateurs qui ont besoin des manipulations spécifiques sur les données. Des fonctionnalités sont également prévues pour le public en général (par exemple, pour permettre de visualiser l'évolution du trafic sur un tronçon de route).

Pour terminer avec cette description générale du système, il est à noter qu'il existe des systèmes de même nature – voir section 3.1. L'originalité de notre approche est basée sur deux axes principaux : a) la recherche multidisciplinaire et complémentaire de plusieurs branches de l'informatique et des mathématiques ; et b) la construction d'un entrepôt de données spatio-temporelles, qui, associé aux ontologies et à la création des magasins de données, permettra de suivre l'évolution des demandes des utilisateurs pour de multiples types de visualisation exploratoire. Les sections qui suivent présentent la construction de cet entrepôt et l'état actuel du projet.

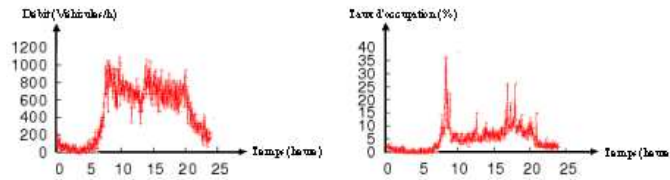


FIG. 2 – Exemples de séquences journalières de débit et de taux d'occupation.

3 Stockage des données du trafic routier

3.1 Variables macroscopiques et réseau routier

Dans le domaine du transport, le trafic est décrit à l'aide de variables macroscopiques temporelles mesurées par des capteurs implantés le long des axes routiers. Les variables macroscopiques sont le débit (exprimé en nombre de véhicules par unité de temps - généralement minute ou heure) et le taux d'occupation (exprimé en pourcentage). Les capteurs de mesure des variables du trafic routier sont reliés à des détecteurs, eux-mêmes faisant partie d'une station de mesure. La station de mesure est équipée d'un moyen d'acquisition en lien avec les détecteurs, d'un moyen de stockage et d'un moyen de transmission permettant de restituer les mesures effectuées. Nos données proviennent de capteurs fixes enfouis dans les chaussées et qui permettent de mesurer directement le taux d'occupation, c'est-à-dire le rapport entre le temps de passage d'un véhicule sur le capteur et l'intervalle de temps de la mesure. Par déduction, on obtient le débit, qui traduit le nombre de véhicules qui passent en un point x donné sur le réseau pendant un intervalle de temps delta t .

Le système de supervision CLAIRE [Scemama *et al.*, 2000, Scemama et Carles, 2004] de l'INRETS modélise le réseau routier par un graphe orienté où chaque arc représente une voie de la chaussée sur un tronçon de route. Un axe routier (rue, route, autoroute) est constitué d'une succession d'arcs, c'est-à-dire d'un chemin dans le graphe. Les mesures des variables macroscopiques de chaque capteur sont associées aux arcs modélisant le réseau routier.

Nous travaillons actuellement sur un jeu de données constitué par le système de supervision CLAIRE. Il s'agit de mesures de débit et de taux d'occupation localisées sur le graphe modélisant le réseau routier d'une grande ville française. Ce jeu de données sera étendu par la suite.

Les mesures de débit et de taux d'occupation constituent des séquences temporelles (*time series*). La figure 2 en présente un exemple pour chacune des deux variables. Dans le cas du projet CADDY, les valeurs de ces séquences sont enregistrées toutes les 3 minutes par plus de 400 capteurs et forment un ensemble d'environ 400 000 valeurs élémentaires par jour. L'interdépendance entre ces variables est exprimée par un diagramme qui exprime le débit (q) en fonction du taux d'occupation (τ). Il est calculé à partir de méthodes statistiques de régression permettant de trouver une fonction

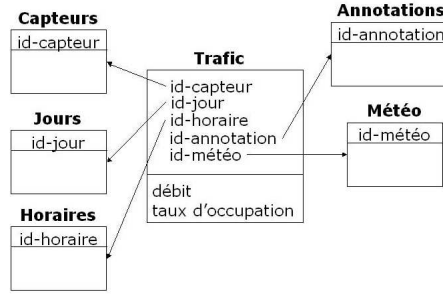


FIG. 3 – Schéma de l'entrepôt de données

polynomiale, logarithmique ou exponentielle passant au plus près du nuage des points mesurés. Ce diagramme permet de détecter des congestions de trafic, des points critiques, etc. Il est à noter que pour un débit donné, la circulation peut être fluide ou saturée. Un débit nul peut notamment signifier que le tronçon est vide ou que le taux d'occupation a atteint sa valeur limite. Plus de détails sur les variables macroscopiques du trafic routier et sur leurs relations figurent dans [Cohen, 1993].

Une main courante d'annotations est disponible en complément des mesures précédentes. Elle reflète des événements, positionnés dans l'espace et dans le temps, ayant pu affecter le trafic : incident sur la voie publique, panne de feux de signalisation, chute de neige, etc. Des données météorologiques sont aussi disponibles. Ils indiquent la température et les précipitations cumulées durant la journée.

3.2 Construction de l'entrepôt

L'entrepôt de données pour le stockage et la gestion des séries temporelles du trafic routier est actuellement en cours de construction. Dans un premier temps, les insertions en temps réel à partir des capteurs ne sont pas prévues. Nous souhaitons d'abord développer les principales fonctionnalités et les valider auprès des utilisateurs qui forment la catégorie des décideurs. Les données de base peuvent être organisées selon différents axes (capteurs, temps, valeurs mesurées). Chacun de ces axes encapsule plusieurs informations – par exemple, l'axe du *temps* contient des attributs sur les horaires, les dates, etc.

La conception de cet entrepôt met en valeur non seulement ces axes mais également le fait que l'évolution des valeurs de séquences temporelles issues du trafic routier est fortement corrélée avec l'activité humaine aux alentours de l'emplacement des capteurs. Cette observation, déjà signalée par les experts, a aussi été constatée lors des expériences préliminaires que nous avons menées sur les données de base.

Par exemple, la figure 2, concernant le débit et le taux d'occupation d'un jour de semaine ordinaire (de lundi à vendredi non férié), illustre que le débit augmente pendant les heures de pointe (entre 8h et 10h et entre 17h et 20h) et que, de manière corrélée, les pics du taux d'occupation sont visibles pendant ces mêmes tranches horaires. Les pics

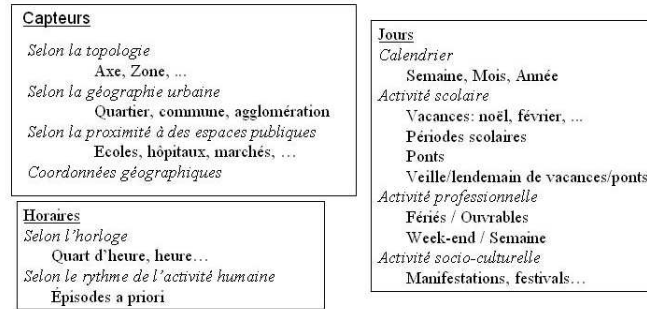


FIG. 4 – Dimension de l'entrepôt de données.

peuvent être décalés ou disparaître lorsqu'il s'agit d'un jour de week-end ou d'un jour férié. De plus, la position du capteur à la proximité d'une école ou d'un marché influe sur l'état de trafic à certaines heures (heures de sortie d'écoles, heures d'ouverture du marché). Les conditions météorologiques et des événements spécifiques (travaux, accidents) doivent également être pris en considération.

La figure 3 montre le noyau du schéma de l'entrepôt. Elle représente un schéma en étoile avec une table de faits "Trafic" qui est liée à cinq tables de dimensions: (1) "Jours" et (2) "Horaires" (pour faire l'analyse selon l'activité humaine à travers le temps), (3) "Capteurs" (permettant d'évaluer l'effet de l'emplacement des capteurs), (4) "Météo" et (5) "Annotations" (pour analyser le trafic selon les conditions météorologiques et/ou la présence d'un événement). A un jour donné, à un horaire fixé, pour un capteur donné, pour certaines conditions météorologiques et pour un événement donné, deux mesures sont enregistrées dans la table des faits: le débit et le taux d'occupation.

Une partie du contenu des dimensions "Jours", "Horaires" et "Capteurs" est détaillé sur la figure 4. Plutôt que de décrire le schéma complet, nous avons choisi d'indiquer quelques indicateurs et attributs choisis pour les agrégations souhaitées. Le choix des attributs pour chacune des dimensions a été réalisée à partir des partitions emboîtées sur les valeurs des attributs de manière à avoir plusieurs niveaux d'agrégation.

On peut remarquer, par exemple, que les données issues de la dimension "Capteurs" peuvent être vues à plusieurs niveaux d'agrégation selon, entre autres, la topologie du graphe routier, la géographie de la ville, ou l'influence que peuvent avoir des éléments urbains spécifiques situés aux alentours d'un capteur sur les données mesurées (école, aéroport, gare routière, etc). On remarque que la présence des coordonnées géographiques associées au capteur permet d'intégrer les informations de l'entrepôt à celles d'autres systèmes d'information urbains dont les données sont aussi liées à des coordonnées géographiques, par exemple, le cadastre, ou la gestion des égouts. Cela permet aussi, au niveau interface, la visualisation de l'évolution des conditions du trafic sur une carte de la ville et l'interaction de l'utilisateur avec cette carte pour poser ses requêtes.

Dans le même esprit, la dimension "Jours" permet de dériver des relations entre

les valeurs mesurées et des différents types d'événement temporel associés à des activités humaines. Les attributs sont répartis en quatre catégories, selon : (1) *les partitions classiques du calendrier* (jour, mois, trimestre, année, etc), (2) *l'activité scolaire*, (3) *l'activité professionnelle* et (4) *l'activité socio-culturelle* (grève, défilé, festivals, manifestations sportives, etc.).

Les attributs de la dimension "*Horaires*" sont répartis selon : (1) *les heures de l'horloge* et (2) *les rythmes de l'activité humaine*, à plusieurs niveaux de granularité (nuit, petit matin, matin, etc.). La même politique a été suivie pour définir les attributs des autres dimensions.

Quelques expérimentations déjà réalisées sur les données montrent que certaines de ces agrégations sémantiques apparaissent dans les valeurs mesurées. Par exemple, nous avons exécuté l'algorithme de classification de données, K-means [McQueen, 1967] sur les valeurs mesurées par plusieurs capteurs, sur plusieurs mois, pour déterminer des motifs sur la dimension temporelle (Jour). Après des essais, trois classes ont été identifiées: celle regroupant les jours ouvrables d'une semaine, celle des samedis des ponts et celle des dimanches et des jours fériés. Cela indique que les données vérifient l'hypothèse indiquant que le comportement du trafic pendant un jour de fête est similaire à celui des dimanches dans une même région.

Le volume des séries temporelles collectées met celles-ci hors de portée non seulement de l'étude directe (par exemple par visualisation), mais il interdit aussi l'utilisation de n'importe quel algorithme non trivial dans le cadre d'analyse en-ligne. Pour cette raison, on doit utiliser des représentations plus compactes. La réduction du nombre des dimensions des séries temporelles est effectuée en agrégeant les instants de mesure en épisodes temporels. Une séquence peut être résumée numériquement ou représentée de manière symbolique (voir section 4). Le choix du type de résumé est fortement lié au type de requête auquel le système doit pouvoir répondre.

4 Multi-représentation de longues séries temporelles

Afin de ne pas limiter a priori le champ des analyses, on cherche à construire des représentations préservant au maximum l'information présente dans les données, mais sans avoir de connaissances sur ce qui constitue justement cette information. On utilise donc des représentations sans perte d'information et on cherche les paramètres de ces représentations de façon à minimiser l'erreur de modélisation, par exemple au sens des moindres carrés. Dans une démarche d'extraction d'information, on décrit les séries temporelles à l'aide d'un alphabet symbolique constituant un univers du discours. Ceci permet aux utilisateurs une manipulation explicite des informations selon différents besoins.

Résumés numériques : Contrairement à la compression de données, les résumés numériques permettent des analyses sur des extraits de séries temporelles. Il faut donc permettre une localisation temporelle des informations. Ceci nous a amené à une partition du domaine de définition temporel en épisodes. Les modèles les plus simples comme les modèles linéaires d'ordre 0 (constant par morceaux) et d'ordre 1 permettent une approximation aussi bonne que l'on veut pour autant que la partition en épisodes soit

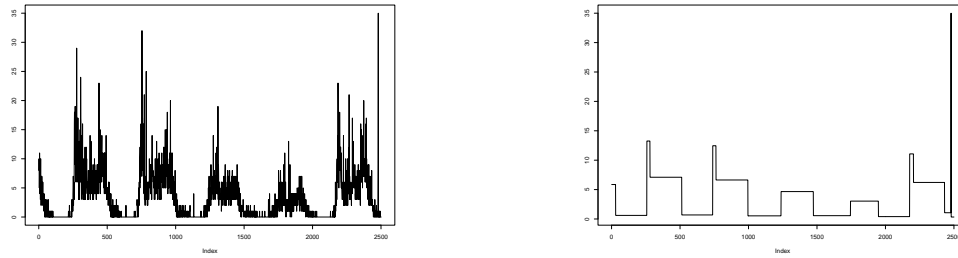


FIG. 5 – Extrait de série temporelle de taux d'occupation et sa modélisation par APCA.

suffisamment fine. Les choix d'un modèle et d'un nombre d'épisodes pertinents doivent généralement être liés aux données à représenter. Le découpage en épisodes peut être régulier et fixé a priori ou bien adapté localement aux données. PAA (*Piecewise Aggregate Approximation*) [Keogh et al., 2000] et APCA (*Adaptive Piecewise Constant Approximation*) [Chakrabarti et al., 2002], illustrées par la figure 5, sont deux types de représentations numériques de séries temporelles par des modèles linéaires d'ordre 0 associés à des découpages qui sont respectivement réguliers et adaptatifs.

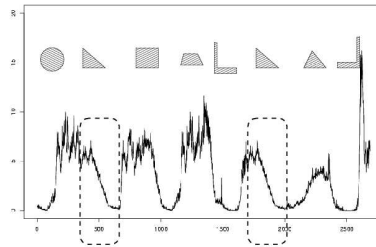


FIG. 6 – Extrait de série temporelle de taux d'occupation et sa modélisation par CBSR.

Représentations symboliques: Les outils d'analyse de données doivent permettre à l'utilisateur de poser des requêtes sur les concepts qu'il connaît, donc des symboles. Nos représentations symboliques associent un symbole à une classe d'extraits de séries temporelles considérés comme équivalents. Des symboles peuvent par exemple être associés à différents profils typiques de circulation journalière ou des niveaux typiques d'encombrement des axes routiers. Ces alphabets de symboles sont obtenus par classification. Dans le cas d'extraits définis sur des épisodes de même taille (par exemple une journée), on peut effectuer une classification numérique non supervisée pour calculer des profils prototypes. C'est le principe de la CBSR (*Clustering Based Symbolic Representation*), illustrée à la figure 6. Pour une représentation symbolique

basée sur un découpage épisodes adapté localement aux données, on peut effectuer une classification automatique des paramètres des modèles. Un cadre général de définition de plusieurs représentations symboliques est présenté dans [Hugueney, 2003].

5 État de l'implémentation du système

Jusqu'à présent, nous avons expérimenté plusieurs approches de représentation numériques et symboliques sur les séries temporelles de trafic routier mises à notre disposition. Nous avons également développé un outil permettant à l'utilisateur de suivre l'évolution des données spatio-temporelles dans une section choisie d'une ville, pour un intervalle de temps donné. Cet outil intégrera l'entrepôt de données en cours de construction. L'interaction de l'utilisateur est faite à travers une carte du réseau routier de la ville, afin d'afficher les renseignements souhaités.

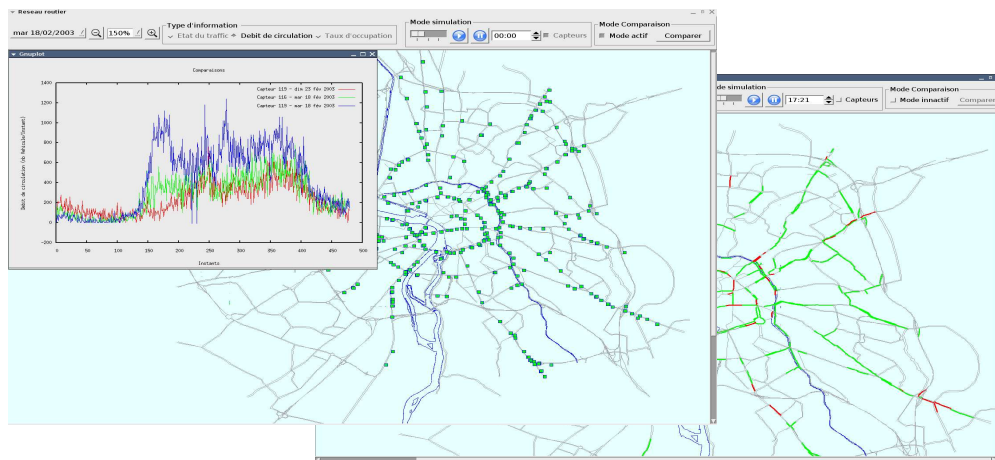


FIG. 7 – Différentes vues disponibles du même ensemble de données.

La figure 7 montre les principales options d'interaction. A partir du menu déroulant en haut à gauche, l'utilisateur peut se projeter à travers le temps et sélectionner les dates qui l'intéressent. Par l'intermédiaire des fonctions de zoom, il peut aussi faire une analyse multi-échelle.

La boîte "Type d'information" permet à l'utilisateur de sélectionner le type d'information (débit, taux d'occupation ou état de la route). En déplaçant le curseur sur la carte, l'utilisateur est en mesure d'identifier chaque capteur (les petits carrés visibles sur la figure 7) et afficher les séries temporelles associées à un ou plusieurs capteurs.

L'exploitation peut être spatiale, temporelle ou spatio-temporelle. En affichant plusieurs séries temporelles issues des capteurs, l'utilisateur va pouvoir procéder à diverses comparaisons :

Comparaison spatio-temporelle : En sélectionnant le même capteur lors de deux journées distinctes, on peut procéder à une analyse temporelle du trafic – voir figure

8.(a) montrant les différences de comportement entre un jour de semaine et un dimanche. En sélectionnant deux capteurs géographiquement distants lors d'une même journée, on peut procéder à une analyse spatiale du trafic – voir figure 8.(b) montrant les différences de comportement entre le centre ville (courbe verte/claire) et la périphérie (courbe rouge/foncée).

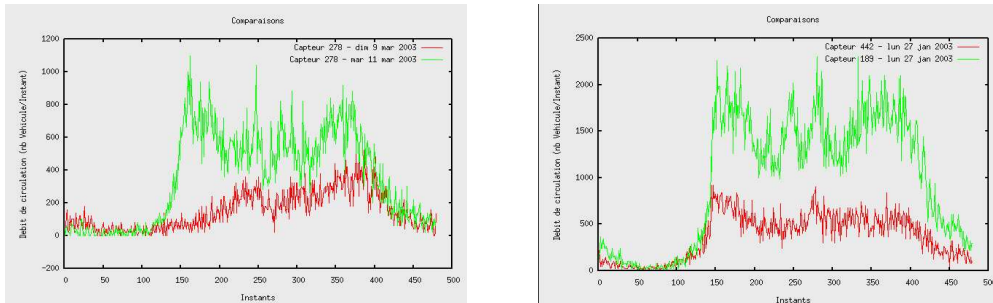


FIG. 8 – (a) : débit de circulation à un capteur pour un dimanche (rouge/foncé) et un mardi(vert/clair); (b) : débit de circulation pour deux capteurs distincts le même jour.

Analyse de corrélations, espace des phases : Les exemples précédents ne concernent que le débit, mais on peut tout aussi bien traiter d'autres variables comme le taux d'occupation (cf.figure 9.(a)) ou des états de trafic. Nous nous sommes aussi intéressés aux corrélations qui peuvent exister entre le débit et le taux d'occupation pour un capteur donné – figures 9.(b) et 9.(c).

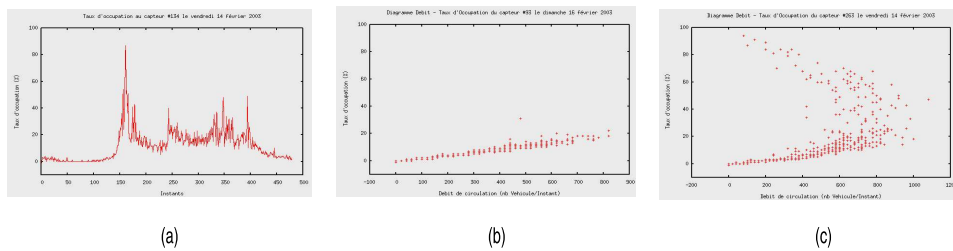


FIG. 9 – (a) : taux d'occupation d'un capteur à une date donnée; (b) : diagramme montrant un régime non congestionné durant la journée; (c) : diagramme pour un état journalier composé de sous-états successifs non congestionnés et congestionnés.

L'outil possède également un module de visualisation permettant d'afficher dynamiquement l'état de la circulation à des différents moments de la journée (cf. partie droite de la figure 7). Ce module peut, entre autres, permettre à l'utilisateur d'identifier de manière intuitive quelques motifs de corrélation spatio-temporelle.

6 Etat de l'art

6.1 Modélisation du trafic routier et fonctionnalités associées

Le trafic routier est modélisé selon différentes approches classées selon un niveau de granularité ou niveau de détail lié au nombre de véhicules. Les modélisations sub-microscopiques prennent en compte toutes les composantes des véhicules et leurs interactions avec l'environnement. Les modèles microscopiques décrivent le comportement de chaque véhicule et conducteur dans un environnement modélisé avec lequel ils interagissent [Carles, 2001]; les techniques multi-agents sont fréquemment utilisées dans ce cadre [Balbo et Pinson, 2005]. L'approche mésoscopique décompose le trafic en pelotons de véhicules. Enfin, l'approche macroscopique consiste à modéliser l'écoulement de flux de véhicules dans les sections du réseau de déplacement. La modélisation est difficilement indépendante des applications et le graphe routier peut apparaître avec différents niveaux de granularité [Carles *et al.*, 2003], avec des étiquettes associées pour ajouter de la sémantique. Cette modélisation se trouve au coeur des applications du domaine des transports et de leur intégration [Carles *et al.*, 2003].

On distingue trois fonctionnalités dans les applications du domaine du trafic routier : le recueil des données, l'analyse et l'interprétation de ces données, et la décision et le contrôle qui en résultent. Il y a une grande gamme de problèmes de recherche associées, tels que l'identification et le suivi des états du trafic, ou la définition des actions de contrôle qui peuvent dans certains cas lancer l'exécution de la commande associée. De plus, plusieurs questions peuvent être associées à des problèmes de recherche en bases de données et fouille de données, telles que l'identification de capteurs fautifs, des données aberrantes et la reconstitution et nettoyage des données recoltées.

6.2 SIG et entrepôts de données pour le trafic routier

Les travaux sur les entrepôts de données spatiales traitent essentiellement d'aspects généraux de ces systèmes indépendamment des applications. Ils abordent en particulier l'indexation, l'agrégation spatiale et les interfaces utilisateurs [Ahmed *et al.*, 2004, Rao *et al.*, 2003]. [Fernando *et al.*, 2004] présentent les agrégations des données spatio-temporelles comme le résultat de la combinaison de méthodes d'agrégations spatiales et d'agrégations temporelles. Même si les auteurs présentent plusieurs exemples et examinent les questions d'évaluation des requêtes associées, ils ne s'occupent pas du problème de la croissance continue des données associée aux séries spatio-temporelles, et donc la question d'un entrepôt de données n'est pas considérée.

L'extension des Systèmes d'Information Géographique (SIG) au domaine des transports a donné naissance à l'acronyme *GIS-T* pour *Geographic Information System for Transportation* [Goodchild, 2000, Thill, 2000]. La plupart des SIG fournit des fonctionnalités de connexion avec des SGBD, ce qui permet une gestion souple des données spatiales. Pour la gestion des applications de transport, on les trouve généralement couplés à des systèmes d'information de transport (TIS - *Transportation Information System*) [Thill, 2000], les SIG n'offrant pas toujours toutes les fonctionnalités nécessaires à ce domaine. Dans [Miller et Shaw, 2001], par exemple, les auteurs couplent un SIG avec leur outil de gestion de trafic, afin de pouvoir visualiser les congestions et aider les

utilisateurs à déterminer leur itinéraire. Le système CLAIRE [Scemama *et al.*, 2000, Scemama et Carles, 2004] utilise également un SIG dans ses implémentations actuelles.

A notre connaissance, l'utilisation d'entrepôts de données spatiales pour le trafic routier n'a été faite que dans [Bertini et Rodriguez, 2005, Lu *et al.*, 2005] où un entrepôt est établi à des fins de stockage et de visualisation. La visualisation se base sur l'utilisation de couleurs permettant de différencier la congestion de trafic de la fluidité. La visualisation relève de la fonction de suivi du trafic routier et ne concerne pas les fonctionnalités de décision et de contrôle.

6.3 Approximation de séquences temporelles

La recherche en séries temporelles a produit une vaste gamme de résultats concernant l'approximation de ces séquences. Les méthodes d'approximation peuvent être divisées en deux catégories selon le type de traitement effectué: (1) les fonctions qui représentent les séquences par des segments, et (2) les méthodes de transformation qui modifient l'espace initial de définition des séquences. Dans la première catégorie on trouve les méthodes *PAA* [Keogh *et al.*, 2000], *APCA* [Chakrabarti *et al.*, 2002] ou *PLA* (*Piecewise Linear Approximation*) [Keogh *et al.*, 2001], qui utilisent des fonctions d'ordre n , constantes par morceaux. Les méthodes de transformation permettent de décomposer les séquences temporelles dans un espace de fonctions connues a priori, comme pour la Transformée de Fourier Discrète ou la Transformée en Ondelettes Discrètes, utilisée dans [Faloutsos *et al.*, 1994].

La représentation symbolique permet de décrire une séquence temporelle par un ensemble de symboles discrets. La plupart des approches, dont [Park *et al.*, 2000], utilise la méthode de discrétisation *Equal-length categorisation* pour représenter symboliquement les séquences. Cette méthode divise les domaines de valeurs de la séquence en intervalles de même longueur et associe un symbole à chaque intervalle. Ainsi, chaque valeur de la séquence est associée au symbole de l'intervalle auquel elle appartient. D'autres approches associent des symboles aux segments d'approximation de la séquence obtenus par la méthode *PAA* [Lin *et al.*, 2003] ou par la méthode *APCA* [Hugueney, 2003]. A notre connaissance, la combinaison de représentations multiples de données de trafic routier (représentation symbolique ou sous forme de résumés numériques) n'a jamais été faite.

7 Conclusion

Cet article a présenté une démarche multidisciplinaire pour le traitement de masses de données spatio-temporelles dans le domaine du trafic routier. Cette démarche intègre un entrepôt de données à des fonctionnalités d'analyse et représentation multi-échelle, numérique et symbolique. La plupart des systèmes d'information pour le trafic routier utilise des bases de données classiques pour la gestion des données, ce qui limite les fonctionnalités offertes aux utilisateurs.

Les mécanismes d'agrégation des entrepôts de données permettent, grâce au stockage de données agrégées, de répondre rapidement à des requêtes plus complexes – par exemple, "Le débit moyen par capteur pour les lundis de 2004 entre 12h et 14h".

Lorsque des états symboliques sont calculés à partir des données brutes et de seuils fixés par les experts du domaine du trafic routier, l'entrepôt peut stocker ces états symboliques et permettre ainsi de répondre à des requêtes telles que "Quels sont les états symboliques (ex. *fluide* ou *saturé*) du trafic sur un ensemble de capteurs pour les jeudis du premier semestre 2004 entre les heures 17h et 20h?".

Nous proposons d'étendre les mécanismes des entrepôts de données en utilisant des techniques de résumé plus fines à partir desquelles il est possible d'extraire de l'information et de l'interroger à différents niveaux hiérarchiques. Cela doit permettre de répondre à des requêtes encore plus complexes – par exemple, "Etant donné un événement (ex. match de rugby) devant avoir lieu à un endroit e (ex. le stade Charletty à Paris), à une date d et une heure h , quel est l'état prévu du trafic dans la zone z aux heures $h - i$ ou $h + j$?", "La rue r est-elle bouchée le dimanche entre 14h et 16h?".

Références

- [Ahmed *et al.*, 2004] T.O. Ahmed, M. Miquel, et R. Laurini. Continuous data warehouse: concepts, challenges and potentials. In *12th Int. Conf. on Geoinformatics*, pages 157–164, 2004.
- [Balbo et Pinson, 2005] F. Balbo et S. Pinson. Dynamic modeling of a disturbance in a multi-agent system for traffic regulation. *Decision Sup. Sys.*, 41(1):131–146, 2005.
- [Bertini et Rodriguez, 2005] Matthews S. Hansen S. Delcambre A. Bertini, R.L. et A. Rodriguez. Its archived data user service in portland, oregon: Now and into the future. In *8th Int. IEEE Conf. on Intel. Transport. Sys.*, pages 13–16, Vienna (Austria), Sept. 2005.
- [Carles *et al.*, 2003] O. Carles, G. Scemama, et M. Tendjaoui. Concepts génériques pour la supervision des réseaux multimodaux. *Génie Logiciel*, 65, June 2003.
- [Carles, 2001] O. Carles. *Système de Génération Automatique de Bases de Données pour la Simulation de Situations de Conduite Fondée sur l'Interaction de ses Différents Acteurs*. Thèse de doctorat, Univ. Paul Sabatier, Toulouse, Juillet 2001.
- [Chakrabarti *et al.*, 2002] K. Chakrabarti, E. Keogh, S. Mehrotra, et M. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Trans. on Database Systems (TODS)*, 27(2):188–228, 2002.
- [Cohen, 1993] S. Cohen. *Ingénierie du trafic routier. Eléments de théorie du trafic et applications*. Presses de l'Ecole Nat. des Ponts et Chaussées (ENPC) Paris, 1993.
- [Faloutsos *et al.*, 1994] C. Faloutsos, M. Ranganathan, et Y. Manolopoulos. Fast Subsequence Matching in Time-Series Databases. In *ACM SIGMOD Int. Conf. on Management of Data*, pages 419–429, Mineapolis, USA, 1994.
- [Fernando *et al.*, 2004] I. Fernando, V. Lopez, R. T. Snodgrass, et B. Moon. Spatio-temporal Aggregate Computation: A Survey. Technical Report TR-77, Time Center, 2004.
- [Goodchild, 2000] M. F. Goodchild. Gis and transportation: Status and challenges. *GeoInformatica*, 4(2):127–139, 2000.
- [Hugueney, 2003] B. Hugueney. *Représentations symboliques de longues séries temporelles*. Thèse de doctorat, Univ. Paris 6, 2003.

- [Keogh *et al.*, 2000] E. Keogh, K. Chakrabarti, M. Pazzani, et S. Mehrotra. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, 3(3):263–286, 2000.
- [Keogh *et al.*, 2001] E Keogh, S. Chu, D. Hart, et M.J. Pazzani. An Online Algorithm for Segmenting Time Series. In *IEEE Int. Conf. on Data Mining (ICDM)*, pages 289–296, San Jose, USA, 2001.
- [Lin *et al.*, 2003] J. Lin, E. Keogh, S. Lonardi, et B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, San Diego, California, 2003.
- [Lu *et al.*, 2005] C.T. Lu, L. Sripada, S. Shekhar, et R. Liu. Transportation data visualization and mining for emergency management. *Int. Journal of Critical Infrastructures (Inderscience)*, 1(2/3):170–194, 2005.
- [McQueen, 1967] J. McQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symp. on Math. Statistics and Probability*, pages 281–298, Berkeley, CA : Univ. of California Press, 1967.
- [Miller et Shaw, 2001] H.J. Miller et S.L. Shaw. *Geographic Information Systems for Transportation: Principles and Applications*. New York: Oxford Univ. Press, 2001.
- [Park *et al.*, 2000] S. Park, W-W. Chu, J. Yoon, et C. Hsu. Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases. In *16th IEEE Int. Conf. on Data Eng. (ICDE'00)*, pages 23–32, San Diego (USA), Feb. 2000.
- [Pro, 2006] Protégé, 2006. (date de dernier accès). <http://protege.stanford.edu/>.
- [Rao *et al.*, 2003] F. Rao, L. Zhang, X.L. Yu, Y. Li, et Y. Chen. Spatial hierarchy and olap-favored search in spatial data warehouse. In *6th ACM Int. Workshop on Data Warehousing and OLAP*, pages 48–55, New Orleans (USA), 2003.
- [Scemama *et al.*, 2000] G. Scemama, A. Blaquièrre, et P. Olivero. Claire++ observatory: a new tool to know and assess congestion on road networks. In *7th World Congress on Intel. Transp. Sys.*, Turin (Italy), 6–9 Nov. 2000.
- [Scemama et Carles, 2004] G. Scemama et O. Carles. Claire-siti, public road transport network management control: a unified approach. In *12th IEE Int. Conf. on Road Transport Information & Control (RTIC'04)*, London (UK), April 2004.
- [Thill, 2000] J.C. Thill. Geographic information systems for transportation in perspective. *Transportation Research Part C: Emerging Technologies*, 8(1-6):3–12, 2000.
- [Wache *et al.*, 2001] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, et S. Hübner. Ontology-based integration of information — a survey of existing approaches. In H. Stuckenschmidt, editor, *IJCAI-01 Workshop: Ontologies and Information Sharing*, pages 108–117, 2001.

Summary

This paper presents a multidisciplinary approach to construct an decision support system for traffic. It discusses the system’s architecture, the scheme of its data warehouse, as well as the numerical and symbolical representations of the spatial time series stored in this warehouse.