Université des Antilles

Ecole doctorale 589

Milieu Insulaire Tropical a Risques :
protection, valorisation, santé et développement

Laboratoire de Mathématiques Informatique et Applications (LaMIA, EA 4540)

Thèse de doctorat présentée en vue de l'obtention du grade de docteur en spécialité :
Informatique

Soutenue le 6 Avril 2021,

Campus de Schoelcher, Martinique.

Erol ELISABETH

Fouille de données spatio-temporelle, Résumé de données et Apprentissage automatique

Applications : système de recommandation touristique, données médicales, détection des transactions atypiques dans le domaine financier et prédiction d'activité dans les TPE

Membres du jury
Mr AGOSTINELLI Serge, Professeur, Université des Antilles
Mr ALCANTARA Christophe, MCF HDR, Université Toulouse 1 Capitole
Mme BATAZZI Claudine, Professeur, Université Côte d'Azur
Mr HASLER Maximilian, MCF HDR, Université des Antilles
Mme POPLIMONT Christine, Professeur, Université Aix-Marseille
Mr RICCIO Pierre-Michel, Professeur, Institut Mines Telecom

Résumé

La fouille de données est une des composantes du Customer Relationship Management ou Gestion de la Relation Client (CRM) largement déployée dans les entreprises. C'est un processus d'extraction à partir de données de connaissances intéressantes, non triviales, implicites, inconnues et potentiellement utiles. Ce processus s'appuie sur des algorithmes issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) pour construire des modèles à partir des données stockées dans des entrepôts de données.

Autrement dit, il s'agit de trouver des schémas pertinents (patterns) ou des ensembles similaires, selon des critères fixés au départ et extraire de ces données les connaissances utiles. Cette connaissance permet d'aboutir à la découverte des motifs fréquents ou non fréquents. L'identification des règles les plus fréquentes, mais aussi dans certains cas des règles les moins fréquentes, a tout son sens selon la nature de la recherche. Avec la fouille de données il s'agit d'opter pour une démarche sans a priori et chercher à faire émerger, à partir des données, des inférences dont il faudra évaluer la qualité. Cette approche permet donc au travers de la découverte des connaissances et de modèles de faire des analyses de prédictions. Compte tenu du volume de données traitées, la principale problématique dans le domaine de la conception de systèmes liés à l'apprentissage est de produire des recommandations de qualité tout en minimisant, si possible, l'effort (les temps de calcul) requis.

L'objectif de cette thèse est précisément de déterminer des modèles, établis à partir de clusters au service de l'amélioration de la connaissance du client au sens générique, de la prédiction de ses comportements et de l'optimisation de l'offre proposée. Ces modèles ayant vocation à être utilisés par des utilisateurs spécialistes du domaine de données, chercheurs en économie de la santé et sciences de gestion ou professionnels du secteur étudié, ces travaux de recherche mettent l'accent sur l'utilisabilité des environnements de fouille de données.

Au-delà de cet objectif, ce mémoire de thèse montre comment un ingénieur en informatique, chef d'entreprise a fait le lien entre son terrain professionnel et une recherche académique avec pour objectif la performance des organisations (entreprises

ou collectivités). C'est donc une thèse appliquée qui s'intéresse à la fouille de données spatio-temporelle. Elle met particulièrement en évidence une approche originale pour le traitement des données avec un but d'enrichissement des connaissances pratiques du domaine. Le traitement utilise d'une part des algorithmes traditionnels ; d'autre part, il 'mixe' des algorithmes connus notamment dans le cadre de réseaux de neurones ; et enfin, il expérimente des approches nouvelles dans le résumé de données et l'apprentissage automatique à l'application de système de recommandation.

Cette thèse comporte un volet applicatif en quatre chapitres qui correspond à quatre systèmes que nous avons dû mettre en place dans les projets développés dans notre entreprise :

- Un modèle pour la mise place d'un système de recommandation basé sur la collecte de données de positionnement GPS. Les traitements de l'échantillonnage peuvent être obtenus par des algorithmes polynomiaux. Le résumé de données permet de réduire un ensemble de données volumineux pour être utilisées en amont des techniques supervisées ou non supervisées. Elles sont notamment très complémentaires des techniques non supervisées.

- Un outil de résumé de données optimisé pour la rapidité des réponses aux requêtes au programme de médicalisation des systèmes d'information (PMSI). Ici, il est important de mesurer la qualité des données, mais aussi la taille de ces données, car les règles évoluent en fonction de la taille de ces données.

- Un outil d'apprentissage automatique pour la lutte contre le blanchiment dans le système financier. Ici, la connaissance du client constitue l'élément fondamental du dispositif et nous devons nous doter d'un dispositif d'analyse des faits ou opérations dont le caractère atypique permet une identification automatique de celui-ci. L'identification passe par un outil de classification, mais aussi un outil d'apprentissage automatique grâce à la rétropropagation.

- Un modèle pour la prédiction d'activité dans les TPE qui sont météo-dépendantes (tourisme, transport, loisirs, commerce, etc.). Le problème est ici d'identifier les algorithmes de classification et de réseaux de neurones en vue d'une analyse de données dont le but est d'adapter la stratégie de l'entreprise aux mouvements conjoncturels.

Dans ce travail de thèse, ces quatre applicatifs sont les prétextes à un corpus théorique autour des méthodes liées à la classification. L'importance du résumé de données aborde une problématique nouvelle : l'adaptation de la structure de présentation des données en fonction des données analysées. C'est grâce à cette réorganisation que les modalités d'une dimension sont agrégées selon l'ordre de leur proximité. Les règles d'associations permettent une plus grande efficacité à la classification. Elles informent sur les interactions entre les données et trouvent des régularités dans le comportement des producteurs de données. Si la classification se focalise sur le regroupement des données en ensembles de classes prédéfinies, les motifs séquentiels permettent des catégories particulières selon les motifs d'usage à partir des algorithmes d'apprentissage non supervisés comme le K-means. Il permet d'appréhender des données sans label et va trouver des patterns ou une structuration des données suivant une logique floue et la combinaison des réseaux neurones en rétropropagation de Kohonen.

Enfin pour conclure nous revenons sur notre parcours et tout l'intérêt qu'il y a pour un ingénieur à envisager une recherche plus académique.

Abstract

Data mining is one of the components of Customer Relationship Management (CRM), widely deployed in companies. It is the process of extracting interesting, non-trivial, implicit, unknown and potentially useful knowledge from data. This process relies on algorithms from various scientific disciplines (statistics, artificial intelligence, databases) to build models from data stored in data warehouses.

In other words, it is about finding relevant models (patterns) or similar sets, according to criteria fixed at the beginning and extracting useful knowledge from this data. This knowledge makes it possible to discover frequent or infrequent patterns. Identifying the most frequent periods, but also in some cases the least frequent ones, makes sense depending on the nature of the research. With data mining, it is about opting for an approach without a priori and seeking to extract, from the data, inferences with an assessed quality. This approach therefore allows to perform prediction analyzes, through the discovery of knowledge and models. Given the volume of processed data, the main issue in the field of learning systems' design is to produce quality recommendations while minimizing, if possible, the effort (calculation time) required.

The objective of this thesis is precisely to determine models, established from clusters at the service of improving the knowledge of customers in the generic sense, the prediction of its behavior and the optimization of the proposed offer. As these models are intended for use by users who are specialists in the field of data, researchers in health economics and management sciences or professionals of the studied sector, this research work emphasizes on the user-friendliness of data mining environments. .

Beyond this objective, this thesis dissertation shows how a computer engineer, company manager created a link between his professional field and academic research with aim : the performance of organizations (companies or communities). It is therefore an applied thesis which is interested in spatio-temporal data mining. It particularly highlights an original approach to data processing with the aim of enriching practical knowledge in the field. The processing uses traditional algorithms on the one hand; on the other hand, it "mixes" known algorithms, particularly in the context of neural networks; and finally, it experiments with new approaches to data synthesis and machine learning for the application of recommendation systems.

This thesis includes an application component in four chapters which corresponds to four systems that we had to set up in the projects developed by our company:

- A model for setting up a recommendation system based on the collection of GPS positioning data. The sampling treatments can be obtained by polynomial algorithms. Data summary is used to reduce a large set of data to be used upstream of supervised or unsupervised techniques. In particular, they are very complementary to unsupervised techniques.

- A data summary tool optimized for quick responses to requests from the Medical Information Systems Program (PMSI). Here, it is important to measure the quality of data, but also the size of the data, because the rules evolve according to the size of this data.

- A machine learning tool for the fight against money laundering in the financial system. Here, knowing the customer is the fundamental element of the system, we must equip ourselves with a mechanism for analyzing the facts or operations, the atypical nature of which allows them to be automatically identified. Identification involves a classification tool, but also a backpropagation machine learning tool.

- A model for predicting activity in VSEs which depends on weather (tourism, transport, leisure, commerce, etc.). The problem here is to identify classification algorithms and neural networks for data analysis aimed at adapting the company's strategy to economic changes.

In this thesis work, these four applications are the pretexts for a theoretical corpus around the methods linked to classification. The importance of data summary addresses a new issue: the adaptation of the data presentation structure according to the analyzed data. It is thanks to this reorganization that the modalities of a dimension are aggregated in the order of their proximity. Association rules allow greater efficiency in classification. They provide information on interactions between data and find regularities in the behavior of data producers. Meanwhile classification focuses on grouping data into sets of predefined classes, sequential patterns allow special categories based on patterns of use from unsupervised learning algorithms such as K-means. It enables us to understand unlabeled data and finds patterns or data structuring following fuzzy logic and the combination of Kohonen backpropagation neural networks.

Finally, to conclude, let's have a look back at our journey and all the interest there is for an engineer to consider more academic research.

Remerciements

Je tiens à adresser mes plus sincères remerciements au Professeur Serge AGOSTINELLI, qui m'a aidé pour cette thèse, mais surtout qui a compris mon profil d'ingénieur et m'a amené à découvrir une nouvelle façon d'appréhender la recherche informatique. Il m'a conseillé avec toute sa bienveillance naturelle ; nos échanges ont toujours été constructifs et j'espère qu'ils continueront après cette thèse.

Je souhaite également remercier Professeur Richard NOCK pour son accompagnement, ses critiques toujours constructives, son soutien aux projets industriels.

Je remercie aussi Professeur Serge AGOSTINELLI, Docteur Christophe ALCANTARA, Professeur Claudine BATAZZI, Docteur Maximilian HASLER, Professeur Christine POPLIMONT, Professeur Pierre-Michel RICCIO de me faire l'honneur d'être membre du jury de ma thèse.

Je voudrais aussi montrer toute ma gratitude envers tous les membres de ma famille et mes proches qui ont toujours été là et m'ont toujours motivé sans faillir dans la poursuite de mes recherches.

Bien sûr, je tiens à montrer tout particulièrement ma reconnaissance envers mes deux plus proches collaborateurs, Didier Priam et Sévrine Charles-Donatien qui comprennent que pour que la dimension recherche puisse intégrer l'entreprise, il faut pouvoir se surpasser, travailler plus et garder un esprit ouvert à la recherche et à l'innovation.

Merci aussi à Didier Largange, avec qui les échanges techniques et les confrontations de points de vue ont toujours été sources de progrès communs.

Merci au Docteur Charlotte DEVRAUX pour ses apports et éclaircissements dans le cadre des travaux du PMSI.

Merci à Nathalie SEBASTIEN -qui a une place toute particulière- pour la co-écriture des papiers sur le PMSI, la co-animation des conférences en Martinique et Haïti et pour tout le reste.

A toutes ces personnes dont j'ai peut-être omis de faire mention, je vous prie de recevoir toute ma gratitude.

Table des matières

1. Introduction

> «… le traitement des données en masse relève d'une multitude de problèmes à la fois scientifique et de terrain »
>
> *Agostinelli, 2018 [1].*

Mon expérience en tant que responsable d'entreprise, ou bénévole au sein de certaines organisations m'a permis de manipuler plusieurs types de données. Ce mémoire s'appuie sur les travaux autour des données sur lesquelles j'ai travaillé de 2008 jusqu'à nos jours.

De 2008 à 2011, lors de ma première inscription en thèse, j'ai cherché la valorisation des données de suivi GPS installées sur des véhicules de location pour les touristes. Ce projet comportait un volet applicatif à terme important qui vise à fournir une méthode pour la mise place d'un système de recommandation basé sur la collecte de données de positionnement GPS enregistré en temps réel. Il s'agit de créer, dans le cadre d'un apprentissage des déplacements, des classes représentant les caractères de ces déplacements (positions géographiques, durée, ordonnancement).

Le Pr Serge Agostinelli m'a ensuite accompagné pour cette thèse. Il m'a montré le cheminement logique de ma réflexion et aidé à trouver un compromis acceptable entre la mise à distance des contraintes de terrain et la rédaction académique nécessaire à cette distanciation. Il m'a permis d'articuler les choix du professionnel et les nécessités de la recherche appliquée.

J'ai obtenu ensuite de l'agence technique de l'information sur l'hospitalisation (ATIH), l'accès complet au Programme de Médicalisation des Systèmes d'information anonymisé pour le volet Médecine - Chirurgie - Obstétrique (M.C.O). Les établissements de santé publics et privés disposent de volumes importants d'informations quantifiées et standardisées sur leur activité médicale au travers du Programme de Médicalisation des Systèmes d'information (PMSI) qui s'inscrit dans la réforme hospitalière avec comme ambition l'optimisation de l'organisation de l'offre de soins et la réduction des inégalités de ressources entre les établissements de santé (Ordonnance n°96-346 du 24 avril 1996). Suite à une nouvelle demande à l'ATIH, cet accès a été élargi en 2017 aux données sur la Psychiatrie (P.S.Y), les soins de suite et de réadaptation (S.S.R) et l'hospitalisation à domicile (H.A.D) pour la période 2007-2019.

Depuis 2014 dans le cadre professionnel, j'ai eu à diriger le projet de développement d'un système de paiement. Hors volet technique et ingénierie, j'ai dû faire un choix concernant le système de détection de transactions frauduleuses : trouver un prestataire ou élaborer une solution interne. J'ai fait le second choix en cherchant une solution informatique afin de répondre aux exigences du Groupe d'action financière (GAFI). Le GAFI est un organisme intergouvernemental créé en 1989 par les Ministres de ses états membres. Les objectifs du GAFI sont l'élaboration des normes et la promotion de l'application efficace de mesures législatives, réglementaires et opérationnelles en matière de lutte contre le blanchiment de capitaux, le financement du terrorisme et les autres menaces liées à l'intégrité du système financier international. Ce choix a abouti à une approche qui vise à identifier les transactions atypiques.

En 2019, suite à un échange avec un responsable de TPE j'ai tenté de formaliser une problématique récurrente pour les entrepreneurs qui est de pouvoir quantifier le potentiel de leurs activités en fonction des paramètres qu'ils utilisent de façon empirique (météo, situation des transports, grèves, événements géographiques proches, saisonnalité ...). Ce travail est présenté dans le quatrième volet de ce mémoire à travers une approche algorithmique.

Ces quatre expériences « de terrain » ont un point commun : il est possible de proposer des solutions scientifiques en utilisant des outils informatiques et notamment solutions issues de la fouille de données. Ce mémoire s'attelle donc à expliquer les travaux et l'orientation scientifique qui ont permis d'arriver à des solutions fonctionnelles.

1.1.Le terrain et ses problèmes

La petite entreprise se doit d'être agile et certaines problématiques demandent un peu de recul que l'approche scientifique permet d'appréhender. Une expérience du terrain et un retour à la théorie systématique permettent de faire gagner du temps à ces deux pans qui parfois s'opposent.

La théorie, la recherche permet bien souvent de porter une orientation et parfois des solutions que l'entreprise sur le terrain pourra mettre en œuvre avec une grande efficacité notamment sur de petits territoires, ou de petites entreprises.

1.1.1. Définition générale de la fouille de données

Le data mining est un processus d'extraction de connaissances valides et exploitables à partir de grands volumes de données. Il a pour vocation d'être utilisé dans un environnement professionnel. Il se distingue de l'analyse de données et de la statistique par les points suivants :

- Le data mining se situe à la croisée des statistiques, de l'intelligence artificielle et des bases de données. Contrairement aux approches statistiques, le data mining ne nécessite pas que l'on établisse une hypothèse de départ qu'on doit de vérifier. les corrélations intéressantes sont déduites des données elles-mêmes et le logiciel n'est là que pour aider l'utilisateur à les mettre en évidence.
- Les connaissances extraites par le data mining ont vocation à être intégrées dans un schéma organisationnel. Le data mining impose donc d'être capable d'utiliser de manière opérationnelle les résultats des analyses effectuées, souvent dans des délais très courts. Le processus d'analyse doit permettre à l'organisation une réactivité (très) importante.
- Les données traitées sont issues des systèmes de stockage en place dans l'organisation et sont ainsi hétérogènes, multiples, plus ou moins structurées. A priori, elles ne sont pas destinées à l'analyse, sauf dans le cas d'un entrepôt de données, et cela impose de disposer de systèmes performants de préparation ou de manipulation de données.
- Le data mining se propose de transformer en information, ou en connaissance, de grands volumes de données qui peuvent être stockés de manières diverses, dans des bases de données relationnelles, dans un (ou plusieurs) entrepôt(s) de

données (datawarehouse), mais qui peuvent aussi être récupérées de sources riches « bien renseignées » plus ou moins structurées comme internet, ou encore en temps réel (sollicitation d'un centre d'appel, retrait d'argent dans un distributeur à billets...). Lorsque la source n'est pas directement un entrepôt de données, il s'agit très souvent de construire une base de données ou un datamart dédié à l'analyse et aux analystes. Cela suppose d'avoir à sa disposition une palette d'outils de gestion de données (data management). On peut également structurer les données de l'entrepôt sous forme d'un hypercube qui signifie OLAP en anglais (online analytical processing) traitement analytique en ligne, même si cela est assez rare en matière de data mining. Parmi les utilisations du data mining, on peut citer certains exemples (analyser les comportements des consommateurs dans la grande distribution, prédire le taux de réponse à un mailing, dans le secteur bancaire prédire la perte d'un client, détecter des comportements anormaux ou atypiques, etc.).

1.1.2. Principes et méthodes

Le data mining est différent de la statistique qui fixe une hypothèse à confirmer par le traitement des données. Avec le data mining, on adopte une démarche sans a priori et on cherche à faire émerger, à partir des données, des inférences dont il faudra évaluer la qualité.

Le data mining se sert d'algorithmes issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) pour construire des modèles à partir des données, c'est-à-dire trouver des schémas « intéressants » (des patterns ou motifs en français) selon des critères fixés au départ, et extraire de ces données la connaissance utile. Cette connaissance peut être les motifs fréquents ou non fréquents. Il peut être intéressant d'obtenir les règles les plus fréquentes, mais aussi dans certains cas les règles les moins fréquentes en fonction de ce que l'on cherche.

1.1.2.a. Le mode non-supervisé

Les algorithmes non-supervisés permettent de travailler sur un ensemble de données dans lequel aucune des données ou des variables à disposition n'a d'importance

particulière par rapport aux autres, c'est-à-dire un ensemble de données dans lequel aucune variable n'est considérée individuellement comme la cible, l'objectif de l'analyse.

On les utilise par exemple : pour dégager d'un ensemble d'individus des groupes homogènes (typologie) ; pour construire des normes de comportements et donc des déviations par rapport à ces normes (détection de fraudes nouvelles ou inconnues à la carte bancaire, à l'assurance maladie...) ; pour réaliser de la compression d'informations (compression d'image).

Sans être exhaustif, les techniques disponibles sont :

- Techniques à base de Réseau de neurones : carte de Kohonen (1997) (SOM/TOM) [2]. Les cartes auto adaptatives sont constituées d'un ensemble de neurones lesquels sont reliés par une forme récurrente de voisinage linéaire, rectangulaire ou triangulaire.

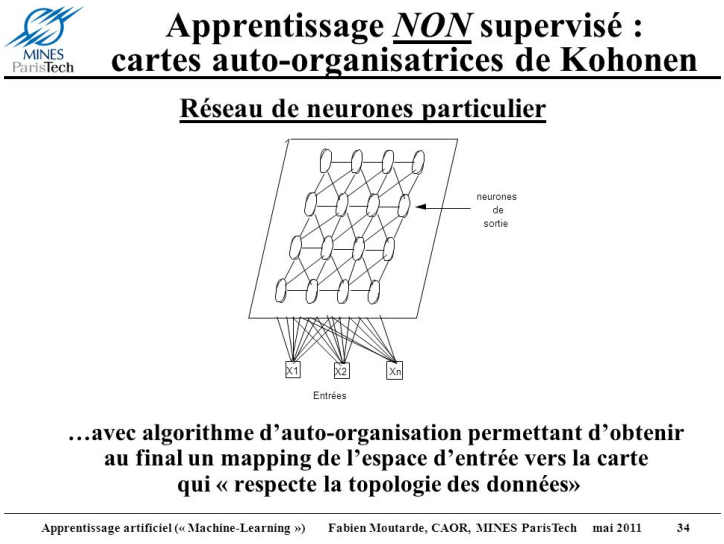


Fig1. Cartes auto-adaptatives (sources Fabien Moutarde (cours) - CAOR- Mines Paris-Tech)

- Techniques utilisées classiquement dans le monde des statistiques : classification ascendante hiérarchique, k-means et les nuées dynamiques (Recherche des plus proches voisins), les classifications mixtes (Birch...), les classifications relationnelles [3] [4].
- Techniques de classification non supervisée et non hiérarchique (clustering) des plus utilisées à partir du k-means :

Etant donné un entier K, K-means partitionne les données en K groupes, ou "clusters", ou "classes" ne se chevauchant pas (donne une Référence). Ce résultat est obtenu en positionnant K "prototypes", ou "centroïdes" dans les régions de l'espace les plus peuplées. Chaque observation est alors affectée au prototype le plus proche (règle dite "de la Distance Minimale"). Chaque classe contient donc les observations qui sont plus proches d'un certain prototype que de tout autre prototype (image inférieure de l'illustration ci-dessous).

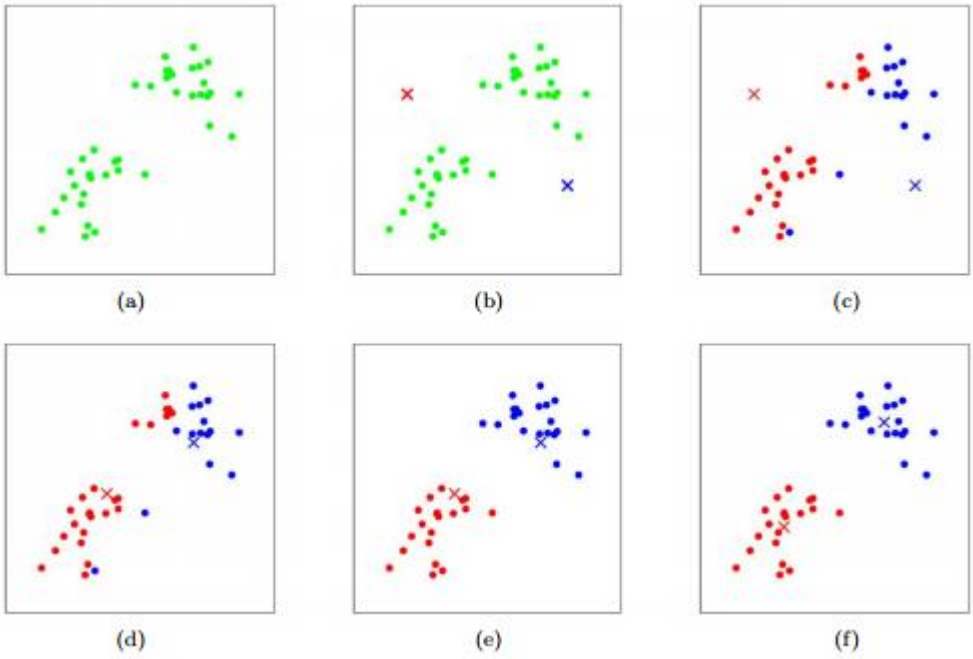


Figure 2 : Chris Piech Standford- K-means algorithm

- Méthodes dites de recherche d'associations [5]. Elles sont à l'origine utilisées pour faire de l'analyse dite de panier d'achats ou de séquences. C'est-à-dire, pour essayer de savoir parmi un ensemble d'achats effectués par un très grand nombre de clients et de produits possibles, quels sont les produits qui sont achetés simultanément pour un supermarché (donne une Référence) [6]. Elles sont également appliquées à des problèmes d'analyse de parcours de navigation de site web (donne une Référence). Ces techniques peuvent donc être utilisées de manière supervisées : algorithmes a priori qui sont les références [5], GRI et Carma utilisés notamment en pharmacie [5] , méthode ARD ou le fameux PageRank de Google utilisent ces méthodes.

Ces techniques nous orientent vers l'identification d'outils algorithmiques qui nous permettons de tester et de tenter de proposer des solutions à nos problématiques liées aux quatre thématiques. Elles ont pour socle commun l'apprentissage automatique et la classification.

Les règles d'associations, motifs séquentiels seront principalement utilisés et retenus dans le cadre de « géotourisme » car les résultats permettent clairement une excellente filtration des parcours des touristes et les relations fortes entre deux sites visités. Les kmeans permettent aussi de classifier les touristes avec un résultat intéressant qui met en évidence les groupes de touristes visitant des lieux similaires.

1.2. Le Géotourisme

Le poids économique du tourisme mondial place le géotourisme à la fois comme première industrie avec des parts de marché qui ne cessent de croître, mais également comme une économie de services en profonde et perpétuelle mutation. C'est l'offre qui crée la demande caractérisée par de fortes disparités de comportements suivant la diversité des attentes des clientèles et des revenus. Or la compétitivité des destinations touristiques passe par une compréhension desdits comportements.

Nous nous sommes particulièrement intéressés à l'optimisation du système de modélisation des comportements (résumé de données). Ce projet comporte en effet un volet applicatif important qui vise à fournir une méthode pour la mise place d'un système de recommandations basé sur la collecte de données de positionnement GPS enregistré en temps réel. Il s'agit de créer, dans le cadre d'un apprentissage des déplacements, des classes représentant les caractères de ces déplacements (position géographique, durée et ordonnancement).

Le principal problème dans le domaine de la conception de systèmes de recommandations est de produire des recommandations de qualité tout en minimisant, si possible, l'effort du traitement informatique requis de la part des producteurs et des consommateurs. Ce problème est particulièrement prégnant lorsque le touriste utilise des dispositifs mobiles embarqués.

Beaucoup de sites Web d'e-commerce rendent des services. Dans ce contexte de foisonnement d'informations, une recherche de produits pourrait renvoyer un très grand ensemble d'accès. Sans l'appui d'un système, filtrant les produits non pertinents, comparant des solutions de rechange, choisir la meilleure option peut être difficile, voire impossible pour l'utilisateur connecté par un dispositif mobile. Ces systèmes sont souvent des systèmes de recommandations. L'objectif d'un système de recommandation est d'aider les utilisateurs à faire leurs choix dans un domaine où ils disposent de peu d'informations pour trier et évaluer les alternatives possibles.

Des systèmes de recommandations manuels élaborés aident aussi à la découverte de nouveaux produits ou services, dans un cadre strictement supervisé. C'est la caractéristique principale des approches traditionnelles des systèmes de recommandation : les systèmes rassemblent des préférences d'utilisateur en interrogeant explicitement l'utilisateur. Le système exploite les préférences acquises pour activer

l'algorithme spécifique de recommandation. Bien que ces préférences tendent à être fiables, l'approche a plusieurs inconvénients. D'abord, les utilisateurs doivent avoir assez de connaissances sur le domaine pour rendre la préférence explicite selon le modèle de produit (par exemple, les attributs du produit). En second lieu, les préférences incertaines ou inachevées peuvent devenir claires pendant que les utilisateurs agissent avec le système et comprennent mieux ce qu'ils veulent et quels produits sont disponibles, ce qui signifie qu'ils ne peuvent pas être demandés avant que le système ne fournisse quelques recommandations. Troisièmement, peu d'utilisateurs sont disposés à indiquer leurs préférences jusqu'à ce qu'ils reçoivent un certain "bénéfice" du système.

Le système est essentiellement algorithmique et nécessite une puissance de calcul importante pour définir un profil de goûts uniques. Ces algorithmes peuvent être utilisés dans un cadre des objets connectés tels qu'un module de suivi GPS installé dans des véhicules de location destinés aux touristes. Ce mode de transport destiné aux particuliers notamment au départ des aéroports, est largement utilisé par les touristes, mais ils sont aujourd'hui orientés vers le e-marketing et l'amélioration de la relation client (e-CRM). Il devient donc nécessaire d'imaginer les extensions des modèles et un élargissement des utilisations dans la recherche d'information, l'analyse des usages et la personnalisation des solutions ou réponses.

La plupart des approches fondées sur la fouille de données sont principalement des approches statistiques où l'ordre d'occurrence d'événements dans l'historique n'est pas pris en compte lors du calcul de recommandation. Or, dans la modélisation des parcours et des durées, cet ordre est une composante très importante et, une autre limitation dans notre cadre d'étude.

Les systèmes existants sont principalement orientés dans le domaine d'aide à "l'achat" sur le web. L'acheteur potentiel devient une structure de données qui décrit les centres d'intérêt dans l'espace des objets commerciaux à recommander. Notre approche se caractérise par la prise en compte dans la recommandation de la notion de positionnement géographique, mais également de donnée temporelle.

Les problèmes sur lesquels il faudra se pencher pour la mise en œuvre de notre approche concernent d'une part, la définition des méthodes et des techniques de mesure de similarités entre comportements et d'autre, la définition des techniques de recommandations personnalisées.

Ce projet comporte un volet applicatif important : fournir une méthode pour la mise en place d'un système de recommandation basée sur la collecte de données de positionnement GPS de véhicules loués pas les touristes. Il s'agit dans le cadre d'un apprentissage des déplacements de créer des classes représentants les caractères de ces déplacements (position géographique, durée, ordonnancement).

1.2.1.Résumé de données de positions GPS

La croissance du volume des données à traiter ouvre la voie à de nouveaux champs d'étude. Dans certains cas, il est possible d'atténuer l'influence de la quantité de données sur les temps de réponse des traitements et algorithmes. Les traitements concernés peuvent se satisfaire d'approximations, généralement de moindre qualité que des résultats non approximatifs, mais obtenus en des temps plus acceptables. C'est le cas de l'échantillonnage au sein d'une population ou des problèmes NP complets dont une solution approchée peut être obtenue par des algorithmes polynomiaux. Le résumé de données permet de réduire un ensemble de données volumineux à un ensemble de taille plus réduite, épuré de ce que l'on considérera comme de l'information non pertinente ou non signifiante, comme du bruit. Elles sont ainsi très souvent, mais pas systématiquement, utilisées en amont des techniques supervisées ou non supervisées. Elles sont notamment très complémentaires des techniques non supervisées.

Le système de géolocalisation GPS installé dans les véhicules envoi des données en continu. Le Global Positioning System (GPS) est un système de positionnement mondial. Ce système imaginé par le physicien D. Fanelli et mis en place à l'origine par le Département de la Défense des États-Unis. Les données transmises par les satellites peuvent être librement reçues et exploitées. Un récepteur peut connaître sa position sur la surface de la Terre, avec une précision aujourd'hui de l'ordre de 10 mètres en moyenne. Dans notre étude 12 véhicules ont été équipés. Ces véhicules ont envoyé des données pendant plusieurs mois.

Le système GPS comprend au moins 24 satellites activés en orbite à 20 200 km d'altitude. Ces satellites émettent en permanence sur deux fréquences L1 (1 575,42MHz) et L2 (1 227,60 MHz) un signal complexe, constitué de données numériques et d'un ensemble de codes pseudo-aléatoires, daté précisément grâce à leur horloge atomique. Les données numériques, transmises à 50 bit/s, incluent en particulier des éphémérides permettant le

calcul de la position des satellites, ainsi que des informations sur leurs horloges internes. Les codes sont un code C/A (acronyme de *coarse acquisition*, acquisition grossière) à 1,023 Mbit/s et de période 1 ms, et un code P (pour précision) à 10,23 Mbit/s avec une période de 280 jours. Le premier est librement accessible, le second est réservé aux utilisateurs autorisés ; il est le plus souvent chiffré. Les récepteurs commercialisés dans le domaine civil utilisent le code C/A. Quelques rares utilisateurs civils spécialisés, comme les organismes de géodésie, ont accès au code P.

Un récepteur GPS qui capte les signaux d'au moins quatre satellites équipés de plusieurs horloges atomiques peut, en calculant les temps de propagation de ces signaux entre les satellites et lui, connaître sa distance par rapport à ceux-ci et par trilatération, situer précisément en trois dimensions n'importe quel point placé en visibilité des satellites GPS2, avec une précision de 15 à 100 mètres pour le système standard. Le GPS est ainsi utilisé pour localiser des véhicules roulants, des navires, des avions, des missiles et même des satellites évoluant en orbite basse. La précision de la position horizontale est de l'ordre de 10 mètres, la position verticale est fausse et varie énormément. Le GPS étant un système développé pour les militaires américains, une disponibilité sélective a été prévue : certaines informations, en particulier celles concernant l'horloge des satellites, peuvent être volontairement dégradées et priver les récepteurs qui ne disposent pas des codes correspondants de la précision maximale.

Pendant de nombreuses années, les civils n'avaient ainsi accès qu'à une faible précision (environ 100 m). Le 1er mai 2000, le président des Etats-Unis a mis fin à cette dégradation volontaire. Certains systèmes GPS conçus pour des usages très particuliers peuvent fournir une localisation à quelques millimètres près. Il utilise alors un système différentiel. Le GPS différentiel (DGPS), corrige ainsi la position obtenue par GPS conventionnel par les données envoyées par une station terrestre de référence localisée très précisément. La station terrestre connaissant son "imprécision" elle l'applique à la position reçue.

Dans certains cas, seuls trois satellites peuvent suffire. La localisation en altitude (axe des Z) n'est pas d'emblée correcte alors que la longitude et la latitude (axe des X et des Y) sont encore bonnes. On peut donc se contenter de trois satellites lorsque l'on évolue au-dessus d'une surface « plane » (océan, mer). Ce type d'exception est surtout utile au positionnement d'engins volants (tels les avions) qui ne peuvent pas se reposer sur le seul

GPS, trop imprécis pour leur donner leur altitude. Il existe un modèle de géoïde mondial nommé « Earth Gravity Model 1996 » ou EGM964 associé au WGS 84 qui permet, à partir des coordonnées WGS 84, de déterminer des altitudes rapportées au niveau moyen des mers avec une précision d'environ 1 mètre. Des récepteurs GPS évolués incluent ce modèle pour fournir des altitudes plus conformes à la réalité.

Les règles d'association sont devenues un concept majeur en fouille de données pour représenter les relations quasi-implicatives entre des variables booléennes (dénommées items). Depuis les premiers travaux d'Agrawal (1993) [5] , de nombreux algorithmes ont été proposés pour découvrir efficacement de telles connaissances dans de grandes bases de données. Tous engendrent d'énormes quantités de règles, dues à l'explosion combinatoire du nombre de conjonctions d'items traitées. Les présents travaux visent à trouver des solutions au problème du volume de données, de l'explosion combinatoire notamment pour les règles d'associations et les motifs séquentiels. Il est nécessaire de faire un premier résumé de données. Ce résumé de données consiste à définir ce qu'est un arrêt et à le coder en terme informatique.

1.2.2.La collecte de données de comportements

Le boîtier GPS de géolocalisation installé dans le véhicule a une mémoire qui lui permet de ne pas perdre les données collectées en cas de réseau GSM / GPRS indisponible. En fonction du matériel embarqué la forme des données reçues est sensiblement la même : La trame d'une position reçue via https (du boitier vers le serveur est décomposable)

- numéro automatique d'insertion de ligne (numauto = ex 5)
- numéro de boîtier gps embarqué (numbeepway = ex 344691000067372)
- l'ip publique du boîtier gps (ip = ex 193.251.163.1)
- la date et l'heure de la position collectée (date heure = ex 2009-12-22 00:00:37)
- la latitude de la position collectée (latitude = ex 1436.0512)
- l'orientation NS de la position collectée (ns = ex N)
- la longitude de la position collectée (longitude = ex 06056.1889)
- l'orientation EW de la position collectée (ew = ex W)
- la hauteur de la position collectée (hauteur = ex 138)

- la vitesse de la position collectée (vitesse = ex 16.01)

- des informations complémentaires de la position collectée (info =; ;20.00) (cumul en km depuis le dernier démarrage)

Une trame peut donc avoir la forme suivante :

5 344691000067372 193.251.163.1 2009-12-22 00:00:37 1436.0512 N 06056.1889 W 138 16.01 ; ;20.00

Un trajet étant la succession des arrêts entre le premier arrêt à partir de la première location jusqu'au retour au parking du loueur. Un premier traitement a été réalisé afin de générer pour les arrêts collectés (position longitude, latitude) une table de succession des arrêts pour chaque trajet.

1.2.3.Les jeux de données

Les jeux de données TRAJET sont constitués de succession d'arrêts caractérisés par un début : la prise du véhicule dans le parking du loueur et une fin par le retour de ce véhicule dans le parc.

La création du jeu TRAJET prend la forme suivante en langage SQL.

```
CREATE TABLE ' jeux_arret '    (
'num' int (11) NOT NULL AUTO_INCREMENT,
'nom_jeux ' varchar (100) NOT NULL,
' ligne_jeux ' int (20) NOT NULL,
'num_beepway ' var char (250) NOT NULL,
'date ' datetime NOT NULL,
' duree_arret ' var char ( 20) NOT NULL,
' num_point ' int ( 11 ) NOT NULL,
' data ' text NOT NULL,
PRIMARY KEY ( 'num' ) )
```

Il est possible après de manipuler ces données simplement.

Dans cette étude le jeu de données est constitué d'un ensemble d'arrêts. Une structure simple et ordonnée du stockage permet d'avoir les bases pour des traitements complexes et multiples (règles d'association, motifs séquentiels, q-motifs…kmeans...). Ce mémoire

explique comment analyser ces données avec les règles d'association et les motifs séquentiels.

1.3. Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie

Depuis la loi du 31 juillet 1991 portant réforme hospitalière puis l'arrêté du 20 septembre 1994 et la circulaire du 10 mai 1995, les établissements de santé publics et privés ont l'obligation de procéder à l'évaluation et à l'analyse de leur activité médicale, mais également à la transmission de celle-ci aux services de l'Etat et de l'Assurance maladie. La mutation de leur environnement les a conduits à mettre en place le Programme de médicalisation des systèmes d'information (PMSI). Il permet de disposer d'informations quantifiées et standardisées, tenant compte notamment des pathologies et des modes de prise en charge, utilisées tant pour la gestion des établissements que pour le financement des établissements de santé et l'organisation de l'offre de soins.

Compte tenu du volume important de données produites par les établissements de santé et sa croissance exponentielle, la fouille de données permet, grâce aux modèles exploratoires développés, une utilisation optimale de celles-ci et une compréhension de l'activité hospitalière. Le principal problème dans le domaine de la conception de systèmes liés à l'apprentissage est de produire des recommandations de qualité tout en minimisant, si possible, l'effort (les temps de calculs) requis. Pour faire face à l'explosion du volume des données, un nouveau domaine technologique a vu le jour : le Big Data. Inventées par les géants du web, ces solutions sont dessinées pour offrir un accès en temps réel à des bases de données géantes. L'objectif est précisément de déterminer des modèles, établis à partir de « clusters » ou k-means (regroupement d'items similaires dans des classes de données représentatives), au service de l'amélioration de la connaissance du client au sens générique, de la prédiction de ses comportements et de l'optimisation de l'offre proposée.

Ces modèles ont vocation à être utilisés par des utilisateurs spécialistes du domaine de données, chercheurs en économie de la santé et sciences de gestion ou professionnels du secteur étudié. Les outils développés comme la cartographie de la répartition du nombre de diagnostique principaux par commune mettent l'accent sur « l'utilisabilité » des environnements de fouille de données comme outil d'aide à la décision.

Pour ce faire, cette approche conduite par une équipe pluridisciplinaire (informatique, santé publique et sciences de gestion) s'appuie sur deux prérequis : (1) l'utilisation des

connaissances du domaine dans l'ensemble du processus de fouille ; (2) l'amélioration de la compréhensibilité et de la confiance de l'utilisateur dans le modèle grâce une association de celui-ci dans sa création.

Nous retiendrons comme axe pour notre réflexion un environnement « symptomatique » de ces questions : l'hôpital. Deux raisons essentielles justifient ce choix. En premier lieu, les établissements de santé publics et privés disposent de volumes importants d'informations quantifiées et standardisées sur leur activité médicale, au travers du Programme de Médicalisation des Systèmes d'information (PMSI) qui s'inscrit dans la réforme hospitalière avec comme ambition l'optimisation de l'organisation de l'offre de soins et la réduction des inégalités de ressources entre les établissements de santé (Ordonnance n°96-346 du 24 avril 1996).

Compte tenu du volume de données enregistrées, l'usage interne (établissements de santé) et externe (assurances maladie, services de l'Etat) de ces informations soulève des problématiques d'exploitation optimale. En second lieu, l'hôpital est confronté depuis 2002 à des mutations importantes avec comme objectif une amélioration de la performance de la gestion hospitalière tout en maîtrisant les dépenses d'hospitalisation, qui sont en augmentation continue (Ordonnance n°2003-850 du 04 septembre 2003, ordonnance n°2005-406 du 2 mai 2005, Loi du 21 juillet 2009 dite Loi Hôpital Patients Santé Territoires (HPST)). Ces réformes se sont accompagnées d'un nouveau mode d'allocation des ressources avec l'application d'une tarification à l'activité (T2A) pour les établissements dits MCO, qui affecte plus spécifiquement les établissements auparavant sous dotation globale, dont les activités seront désormais fonction de l'activité réalisée. Elle remet en cause les positions historiques entre le secteur privé et le secteur public « subventionné » (établissements publics de santé et établissements privés participant aux missions de service public) et obligent à un changement de paradigme, où la performance des établissements est centrale et soulève des questions de connaissance de son « client », le patient et de l'offre de soins proposée. Dans ce contexte, le PMSI prend toute sa place en tant qu'outil de gestion.

Dans le cadre de mon activité professionnelle liée, j'ai eu l'opportunité de travailler sur des procédures Lamda. Pendant une période de 2 années, via une procédure réglementaire stricte, des informations médicales relatives à l'activité de l'année n :

l'établissement de santé peut ainsi envoyer des séjours non codés de l'année n, mais aussi renvoyer des séjours recodés après un contrôle qualité codage ou facture interne.

Cette procédure s'appelle LAMDA ; elle doit faire l'objet d'une demande de la part de la direction de l'établissement à l'Agence Régionale d'Hospitalisation (ARS … de santé à compter d'avril 2010). Cette procédure a été utilisée afin de vérifier de façon algorithmique que certains actes ou codages n'étaient pas oubliés par l'équipe de codage. Par exemple il est facile d'oublier qu'un acte a été fait le week-end. Un passage dans un algorithme déterminant les dimanches permet de supprimer tous les oublis.

J'ai aussi mis en oeuvre des règles d'associations dans les bases de données PMSI MCO (Médecine, chirurgie, obstétrique et odontologie) gérées par l'ATIH, qui regroupent l'ensemble des RSS (résumés de sorties standardisées) des établissements de santé publics et privés de France pour les années 2007 à 2012 relatifs à des prises en charge de patients (activités médicales avec ou sans hébergement, cancérologie) et comportent plus de 132 millions d'enregistrements. Cette utilisation était conditionnée par un avis favorable de la CNIL et de l'ATIH. Ces résultats ont été présentés lors d'un congrès de la FHM Fédération Hospitalière de Martinique (2014) [8] et sur le volet méthodologique à Haiti – E2Tech (2015).

ATIH Agence Technique de l'Information sur l'Hospitalisation, instituée par le décret n°2000-1282 du 26 décembre 2000, repris dans le code de la santé publique aux articles R. 6113-33 et suivants, dont les missions portent essentiellement sur la collecte des données et de la gestion des référentiels, le calcul des tarifs et des coûts de prestation et la contribution au suivi et à l'analyse financière et médico-économique de l'activité des établissements de santé et l'évolution des classifications.

1.3.1. Les enjeux de l'hôpital

L'hospitalisation française a connu depuis plusieurs années des changements majeurs, avec comme objectifs centraux la description de l'activité médicale et plus récemment la performance du système de soins, qui repose très largement sur l'hôpital. Les établissements de santé sont inscrits, par ces réformes, dans une dynamique d'amélioration de la performance de la gestion hospitalière qui doit concilier la maîtrise des dépenses, qui depuis 2002 augmentent de façon continue à plus de 2,5% par an en

volume, la qualité des soins et la garantie de l'accès des soins à tous. Les principaux leviers de ces réformes sont l'évaluation de la qualité des soins et des pratiques, notamment grâce à la certification des établissements, la rénovation des règles de planification hospitalière avec l'instauration de la contractualisation (Ordonnance n°2003-850 du 4 septembre 2003), le déploiement d'une nouvelle organisation au sein de l'hôpital public axé sur le développement du pilotage économique avec la création des pôles et la réforme des instances institutionnelles (Ordonnance n°2005-406 du 2 mai 2005). Ces principes sont réaffirmés à chaque campagne tarifaire annuelle.

Ces réformes se sont accompagnées de l'instauration de la T2A, qui vise à une répartition plus équitable des moyens financiers entre établissements assurant des missions de service public (établissements publics et établissements privés assurant des missions de service public) et à une responsabilisation des acteurs, les recettes dépendant majoritairement de l'activité produite par les établissements. Il est à noter que celle-ci n'est pas neutre sur le secteur privé, dans la mesure où elle marque également un rapprochement des modes d'allocation des ressources entre secteur public et privé.

L'instauration de la T2A, qui s'appuie sur le PMSI, contraint les établissements de santé à changer de paradigme. En effet, la T2A vise à améliorer la transparence : elle assure en effet une plus grande transparence dans le financement des soins hospitaliers en liant le financement à la production des soins. Par ailleurs, elle est voulue comme un mécanisme « équitable » dans la mesure où on paie le même prix pour un même service pour tous les fournisseurs de soins. Cette équité dépend toutefois de la fiabilité de la classification de l'activité en groupes tarifaires : il est impératif que cette classification soit suffisamment fine, et les groupes suffisamment homogène, pour que les établissements qui attirent systématiquement les patients les plus lourds ne soient pas pénalisés. Il faut également bien prendre en compte les facteurs exogènes liés au contexte local et que les établissements ne contrôlent pas, car ils peuvent influencer fortement les coûts. La T2A vise également à améliorer l'efficience, à la fois de chaque établissement individuellement et de l'ensemble du marché : elle introduit en effet une forme de compétition stimulant l'efficience dans un contexte où ces pressions compétitives étaient inexistantes jusqu'alors. Ceci suppose toutefois que les prix reflètent correctement les coûts des producteurs les plus efficients.

Il est donc désormais essentiel pour les établissements d'être en mesure de décrire leur activité médicale, mais également d'en comprendre les leviers et les incidences financières. Ceci passe par une connaissance du type de patients pris en charge (pathologies, durées du séjour, âge, lieu de résidence) avec une mise en perspective sur temps long.

En effet, à l'instar d'autres entreprises de services elles sont en concurrence avec d'autres acteurs, offreurs de soins publics ou privés, à un patient qui est un acteur à part entière du processus de soins. Le patient est le consommateur du service proposé (le soin), mais il va influer tout au long de la prise en charge sur le service produit. De ce fait, il existe un facteur d'incertitude du fait du caractère « intrinsèquement hétérogène » de la cible (Bancel-Charensol et Jougleux, 1997).

Il apparaît donc difficile pour cette activité de service « *obtenir de clients (…) des normes comportementales prévisibles et pratiquement impossibles. Ces participants à la production faiblement encadrés introduisent des incertitudes fortes sur les processus ; leurs résultats et sur leur qualité. Chacun revendique le droit de se comporter comme une exception* » (Gadrey, 1996). Cette difficulté est accrue lorsque les producteurs de soins appréhendent les patients comme des « exceptions ». L'environnement contraint des établissements ne leur permet plus d'accepter comme une fatalité ces situations. Ils ont donc l'obligation de réduire cette incertitude par une meilleure connaissance de la « cible », le patient.

1.3.1.Les outils

Les modèles présentés répondent à deux des thématiques majeures en fouille de données, à savoir : D'une part l'extraction d'informations et de connaissances de données informatisées et d'autre part, la visualisation d'informations, les approches visuelles et interactives pour la représentation, l'extraction d'informations et de connaissances.

1.3.1.a. Structure de la base de données PMSI MCO

Tous les établissements de santé publics et privés (établissements dits MCO qui exercent des activités de médecine, chirurgie, obstétrique et odontologie mais aussi les activités ambulatoires et la cancérologie, établissements de soins de suite et de réadaptation (SSR), établissements exerçant des activités d'hospitalisation à domicile (HAD), établissements exerçant des activités de santé mentale). Cependant les modalités de recueil de

l'information, de traitement de l'information médicalisée et les règles de tarification diffèrent selon le type d'établissement.

La base de données PMSI MCO est constituée de l'ensemble des données des établissements de santé publics et privés, remontées mensuellement selon un format normé et standardisé, le Résumé de sortie standardisé (RSS). Les RSS sont établis à l'occasion de tout séjour hospitalier dans un établissement public ou privé et retrace le séjour du patient (indication de (ou des) l'unité(s) médicale(s) de prise en charge, description de la prise en charge au sein de

l'unité médicale au travers du Résumé d'unité médicale (RUM). Le RUM contient un nombre limité de rubriques, d'ordre administratif et médical, défini par un arrêté du 22 février 2008 modifié 4. Ces informations sont codées selon des nomenclatures imposées *(Figure 3).*

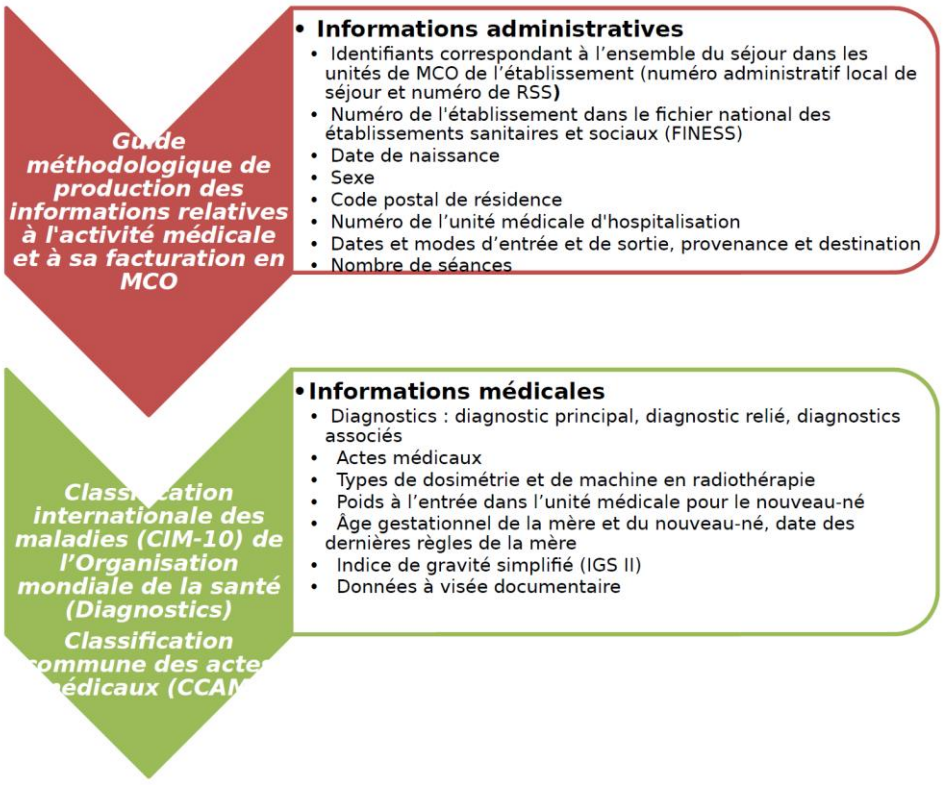


Figure 3 – Rubriques du RUM et nomenclatures de codification

Les informations ainsi recueillies font l'objet d'un traitement automatique aboutissant au regroupement des RSS en classes représentatives et cohérentes. L'Arrêté 4 du 28 février 2008 modifié relatif au recueil, au traitement des données d'activité médicale et des

données de facturation correspondante, produites par les établissements de santé publics ou privés ayant une activité en médecine, chirurgie, obstétrique et odontologie, et à la transmission d'informations issues de ce traitement dans les conditions définies à l'article L. 6113-8 du code de la santé publique (homogène) d'un point de vue médical et économique, appelé groupes homogènes de malades (GHM).

Ce traitement algorithmique de classification permet de regrouper les RSS (Groupage) en fonction du critère médical de prise en charge (i.e. appareil fonctionnel ou motif notoire d'hospitalisation), appelé Diagnostic principal. Ce premier niveau de classement est appelé Catégorie majeure de diagnostic (CMD). Cette classification en GHM peut toutefois être altérée si un évènement médical majeur (acte médical), appelé Acte classant est intervenu au cours de la prise en charge du patient.

Un deuxième niveau de classification en GHM est lié aux Complications ou morbidités associées (CMA), qui permettent de mettre en exergue la sévérité des cas pris en charge, et les incidences sur la durée de la prise en charge (durée de séjour). Ces CMA sont également appelés Diagnostics associés.

L'âge peut également être un facteur déterminant dans la classification en GHM dans la mesure où, l'âge peut être soit de nature à accroître le niveau de sévérité, soit de nature à influer sur la survenance de la pathologie (exemple : limite d'âge de 18 ans pour les affections touchant fréquemment les enfants…). Un focus particulier est fait sur les diagnostics principaux, les diagnostics associés et sur les durées moyennes de séjour.

1.4.Hemdal : Lutte contre le Blanchiment

Blanchiment désigne le « blanchiment » tel que défini par l'article 1 (1) de la Loi AML. Le blanchiment est un des éléments des techniques de criminalité financière. Il consiste dans le fait de dissimuler la provenance d'argent acquis de manière illégale afin de le réinvestir dans des activités légales. Le Blanchiment est une technique érigée en infraction et concerne la vente de substances médicamenteuses et la lutte contre la toxicomanie, telle que modifiée. Le Blanchiment est une infraction de conséquence. Il existe donc toujours en amont une infraction primaire spécifique dont la liste évolue au gré des réformes, des règles applicables en matière de LBC/FT tant au niveau du droit interne qu'au niveau du droit de l'Union Européenne.

Mon activité professionnelle m'a conduit à développer une solution de paiement mobile multidevise. Cette application dans un secteur d'activité réglementé, comporte un volet important qui vise à proposer une approche répondant à la réglementation sur la mise en œuvre d'une approche visant à limiter voir rendre impossible le blanchiment de l'argent circulant entre les clients. Le GAFI régit cette réglementation. La littérature sur le sujet est très peu fournie. *« Paradoxalement au (trop) petit nombre de livres publiés sur le sujet, à l'heure actuelle, tous les domaines sont concernés par la gestion des observations atypiques, seul leur but final diffère. Dans certains domaines, il est seulement nécessaire d'identifier et de supprimer ces individus. Dans d'autres, il faut que les analyses ne soient pas trop impactées par la présence potentielle de ces observations, et enfin pour les derniers, l'objectif même de leurs études est de les détecter.*

Pour exemple, on peut vouloir mettre en place un système de détection d'intrusion en cybersécurité, de détection de fraudes aux cartes de crédit, traquer les variations de capteurs de tous genres pour anticiper des pannes, diagnostiquer des maladies ou bien renforcer la fiabilité des composants électroniques dans le cadre du contrôle de qualité. » [9]

Les règles de bases (normes) sont données par le GAFI qui désigne le Groupe d'action financière, un organisme intergouvernemental créé en 1989 par les Ministres de ses Etats membres. Les objectifs du GAFI sont l'élaboration des normes et la promotion de l'application efficace de mesures législatives, réglementaires et opérationnelles en matière de lutte contre le blanchiment de capitaux, le financement du terrorisme et autres menaces pour l'intégrité du système financier international. Le GAFI est donc un organisme d'élaboration des politiques qui s'efforce de susciter la volonté politique

nécessaire à la mise en œuvre des réformes législatives et réglementaires dans ces domaines. Les normes élaborées par le GAFI n'ont pas de valeur contraignante.

En tant qu'établissement de paiement, la Société pour laquelle j'ai développé l'outil est tenue de mettre en place une fonction de contrôle de conformité (complaisance) et d'en énoncer les modalités de fonctionnement. Ladite fonction contribue au bon fonctionnement du troisième niveau de contrôle interne tel que prévu par les règles applicables à la Société et en particulier les circulaires du Luxembourg : CSSF 04/155 et IML 98/143. Le Luxembourg a été parmi les premiers pays à se doter d'une loi pour lutter contre le blanchiment de capitaux. La dernière mise à jour, fondamentale, de cette législation consiste en un triptyque de lois portant la date du 27 octobre 2010. Ces lois ont été adoptées sur la base des recommandations faites par le GAFI qui, sont venues approuver la conformité du dispositif luxembourgeois avec les règles du GAFI.

Le dispositif mis en place est en premier lieu préventif. Il impose aux établissements financiers des obligations professionnelles et des règles de conduite qu'ils doivent observer à tout moment et de façon continue.

A ce titre, les établissements financiers ont notamment une obligation de vigilance à l'égard de leur clientèle et une obligation de coopération avec les autorités. Avant de nouer une relation d'affaires ou d'exécuter une transaction, ils doivent *vérifier* l'identité de leur client ou du *bénéficiaire effectif*. Par la suite, tout au long de la relation avec le client, ils doivent examiner ses transactions, notamment quant à l'origine de ses fonds. Au moindre soupçon, ils doivent de leur propre initiative informer la cellule de renseignement du parquet de Luxembourg (CRF) qui, en bloquant les transactions suspectes, peut geler les avoirs concernés. La vigilance à l'égard de la clientèle est obligatoirement renforcée vis-à-vis des clients qui sont des personnages politiques, de leur famille et de leurs proches.

Ce dispositif permanent devrait dès lors empêcher que des fonds suspects, en provenance de personnages politiques ou de leur entourage, puissent se retrouver auprès d'établissements financiers au Luxembourg, sans devoir attendre que ces personnages fassent l'objet de mesures internationales après que la situation dans leur pays d'origine aura changé.

Si de telles mesures ou sanctions internationales sont décidées au niveau politique par l'Organisation des Nations-Unies ou par l'Union Européenne, ces mesures sont

introduites au Luxembourg par le biais de règlements de l'Union Européenne directement applicables en droit national ou, en matière de lutte contre le financement du terrorisme, par l'adoption de règlements ministériels sur base d'une des trois lois du 27 octobre 2010 et du règlement grand-ducal du 29 octobre 2010. Au cas où un établissement financier aurait un client visé par une telle sanction internationale, il devrait appliquer la sanction, par exemple en gelant sans délai les avoirs du client, et en informer le Ministère des Finances. Le cas échéant, les avoirs suspects détenus auprès d'établissements financiers peuvent aussi faire l'objet de mesures décidées dans le contexte de l'entraide judiciaire internationale. C'est aussi normalement par la voie judiciaire que sera réglé le sort final de tels avoirs, dont les propriétaires légitimes devront être déterminés en Justice, à moins d'un règlement politique de la situation.

Le droit au Luxembourg comporte deux volets : un volet répressif et un volet préventif. Dans le cadre du volet préventif, les professionnels sont soumis à des règles prudentielles, assorties de sanctions en cas de non-respect qui peuvent être prononcées en dehors de toute procédure liée à un cas de Blanchiment avéré.

Le volet répressif du droit anti-blanchiment regroupe les infractions de Blanchiment et de Financement du Terrorisme. Le Code pénal réprime spécifiquement le Blanchiment et le Financement du Terrorisme. La Loi AML définit ces deux infractions par renvoi à l'article 506-1 du Code pénal et à l'article 8-1 de la Loi de 1973. Le blanchiment constitue ainsi une infraction autonome au titre de l'article 506-1 du Code pénal mais toujours conséquence d'une infraction primaire. En vertu de l'article 506-1 paragraphe 1 du Code pénal, commettent une infraction de blanchiment punie d'une peine d'emprisonnement d'un à cinq ans et d'une amende de 1.250 à 1.250.000 euros :

« 1) ceux qui ont sciemment facilité, par tout moyen, la justification mensongère de la nature, de l'origine, de l'emplacement, de la disposition, du mouvement ou de la propriété des biens visés à l'article 32-1, alinéa premier, sous 1), formant l'objet ou le produit, direct ou indirect, [de l'une ou de plusieurs infractions dont la liste est fournie par l'article 506-1 1)]

2) ceux qui ont sciemment apporté leur concours à une opération de placement, de dissimulation, de déguisement, de transfert ou de conversion des biens visés à l'article 32-1, alinéa premier, sous 1), formant l'objet ou le produit, direct ou indirect, des infractions énumérées au point 1) de cet article ou constituant un avantage patrimonial quelconque tiré de l'une ou de plusieurs de ces infractions;

3) ceux qui ont acquis, détenus ou utilisés des biens visés à l'article 32-1, alinéa premier, sous 1), formant l'objet ou le produit, direct ou indirect, des infractions énumérées au point 1) de cet article ou constituant un avantage patrimonial quelconque tiré de l'une ou de plusieurs de ces infractions, sachant, au moment où ils les recevaient, qu'ils provenaient de l'une ou de plusieurs des infractions visées au point 1) ou de la participation à l'une ou plusieurs de ces infractions. »

Ces points juridiques permettent d'appréhender l'importance de la conformité et de la lutte contre le blanchiment dans un environnement aussi surveillé.

De même, se rendent coupables de blanchiment, ceux qui commettent l'un des actes prévus à l'article 8-1 de la Loi 1973, c'est-à-dire:

« *ceux qui ont sciemment facilité par tout moyen, la justification mensongère de la nature, de l'origine, de l'emplacement, de la disposition, du mouvement ou de la propriété des biens ou revenus tirés de l'une des infractions mentionnées à l'article 8 paragraphe 1., a) et b)ceux qui ont sciemment apporté leur concours à une opération de placement, de dissimulation, de déguisement, de transfert ou de conversion de l'objet ou du produit direct ou indirect de l'une des infractions mentionnées à l'article 8 paragraphe 1., a) et b) ceux qui ont acquis, détenu ou utilisé l'objet ou le produit direct ou indirect de l'une des infractions mentionnées à l'article 8 paragraphe 1., a) et b), sachant au moment où ils le recevaient, qu'il provenait de l'une de ces infractions ou de la participation à l'une de ces infractions sont punis d'un emprisonnement d'un à cinq ans et d'une amende de 1.250 euros à 1.250.000 euros, ou de l'une de ces peines seulement.* »

1.4.1 Le Financement du Terrorisme

Comme pour le Blanchiment, la Loi AML opère un renvoi au Code pénal pour la définition de l'infraction de Financement du Terrorisme.

La lutte contre le terrorisme est l'un des deux volets qui nécessitent une attention particulière dans notre problème. Un rappel de l'environnement permet de comprendre l'attention qui est portée et la difficulté de pouvoir gérer efficacement l'ensemble des informations dans un environnement de production pour un petit établissement de paiement.

A cet égard, l'article 135-5 du Code pénal du Luxembourg dispose ainsi que :

« (1) Constitue un acte de financement du terrorisme le fait de fournir ou de réunir par quelque moyen que ce soit, directement ou indirectement, illicitement et délibérément, des fonds, des valeurs ou des biens de toute nature, dans l'intention de les voir utilisés ou en sachant qu'ils seront utilisés, en tout ou en partie, en vue de commettre ou tenter de commettre une ou plusieurs des infractions visées à l'alinéa (2) du présent article, même s'ils n'ont pas été effectivement utilisés pour commettre ou tenter de commettre une de ces infractions, ou s'ils ne sont pas liés à un ou plusieurs actes terroristes spécifiques.

…

(3) Constitue également un acte de financement du terrorisme le fait de fournir ou de réunir par quelque moyen que ce soit, directement ou indirectement, illicitement et délibérément, des fonds, des valeurs ou des biens de toute nature, dans l'intention de les voir utilisés ou en sachant qu'ils seront utilisés, en tout ou en partie, par un terroriste ou par un groupe terroriste, y compris en l'absence de lien avec un ou plusieurs actes terroristes spécifiques, même s'ils n'ont pas été effectivement utilisés par le terroriste ou le groupe terroriste.

(4) Sont compris dans le terme «fonds» des biens de toute nature, corporels ou incorporels, mobiliers ou immobiliers, acquis par quelque moyen que ce soit, et des documents ou instruments juridiques sous quelque forme que ce soit, y compris sous forme électronique ou numérique, qui attestent un droit de propriété ou un intérêt sur ces biens et les crédits bancaires, les chèques de voyage, les chèques bancaires, les mandats, les actions, les titres, les obligations, les traites et les lettres de crédit, sans que cette énumération ne soit limitative. »

La présentation du volet préventif permet de comprendre que la problématique de la recherche d'une solution de traitement informatique couvre plusieurs volets. Le premier est lié au traitement traditionnel de l'information du client ; le second est une analyse non supervisée mais intégrant les éléments de la loi afin d'évaluer la vigilance nécessaire à l'utilisation des fonds d'un client.

La Loi AML impose aux professionnels assujettis un certain nombre d'obligations visant à prévenir et détecter le Blanchiment et le Financement du Terrorisme. Ces obligations professionnelles en matière de LBC/FT sont les suivantes :

- obligation d'appliquer des mesures de vigilance à l'égard de la clientèle (article 3 de la Loi AML) ;
- obligation de conserver certains documents et informations (article 3 (6) de la Loi AML);
- obligation d'accorder une attention particulière à certaines activités et transactions (article 3 (7) de la Loi AML) ;
- obligations d'organisation interne adéquate (article 4 de la Loi AML) ;
- obligation de coopérer avec les autorités et obligation de déclaration (article 5 de la Loi AML).

De même, en vertu de l'article 4 paragraphe 2 dudit règlement, la Société doit veiller à ce que le transfert de fonds soit accompagné :

- du nom du bénéficiaire ; et
- du numéro de compte de paiement du bénéficiaire.

La connaissance du client constitue l'élément fondamental du dispositif LBC/FT et son identification doit être effectuée préalablement à l'entrée en relation d'affaires ; c'est-à-dire avant l'exécution de toute prestation de services en sa faveur, sauf s'il est nécessaire de ne pas interrompre l'exercice normal des activités ou si le risque de Blanchiment ou de Financement du Terrorisme est faible.

Selon l'article 3 (2) de la Loi AML, les mesures de vigilance consistent en:

- l'identification du client et la vérification de son identité sur base de documents, données ou informations de source fiable et indépendante;
- l'identification du Bénéficiaire Effectif et la prise de mesures raisonnables pour vérifier son identité, ainsi que, pour les personnes morales, les fiducies et les constructions juridiques similaires, la prise de mesures raisonnables pour comprendre la structure de propriété et de contrôle du client;
- l'obtention d'informations sur l'objet et la nature envisagée de la relation d'affaires;
- l'exercice d'une vigilance constante au cours de la relation.

L’article 3(2) d) de la Loi AML précise que l’exercice de vigilance constante de la relation d'affaires consiste à:

- examiner les transactions conclues pendant toute la durée de cette relation d'affaires ;
- examiner, si nécessaire, l'origine des fonds ;
- vérifier la cohérence des transactions par rapport à la connaissance du client, de ses activités commerciales et de son profil de risque ;
- mettre à jour les documents, données ou informations détenus.

1.4.2 Obligation renforcée de vigilance

II s'agit d'appliquer des mesures de vigilance renforcées dans des situations pouvant présenter, par leur nature, un risque élevé de Blanchiment et de Financement du Terrorisme et notamment:

- lorsque le client n'est pas physiquement présent aux fins d'identification ;
- lorsque le client ou le Bénéficiaire Effectif est, ou devient au cours de la relation d'affaires, une personne politiquement exposée résidant à l'étranger ou exerçant une fonction publique dans un Etat tiers ou exerçant une fonction publique pour compte d'un Etat tiers ; ou
- lorsqu’une transaction favorise l’anonymat.

Il est possible d’illustrer un exemple de transaction frauduleuse impliquant un haut fonctionnaire (Personne exposée politiquement).

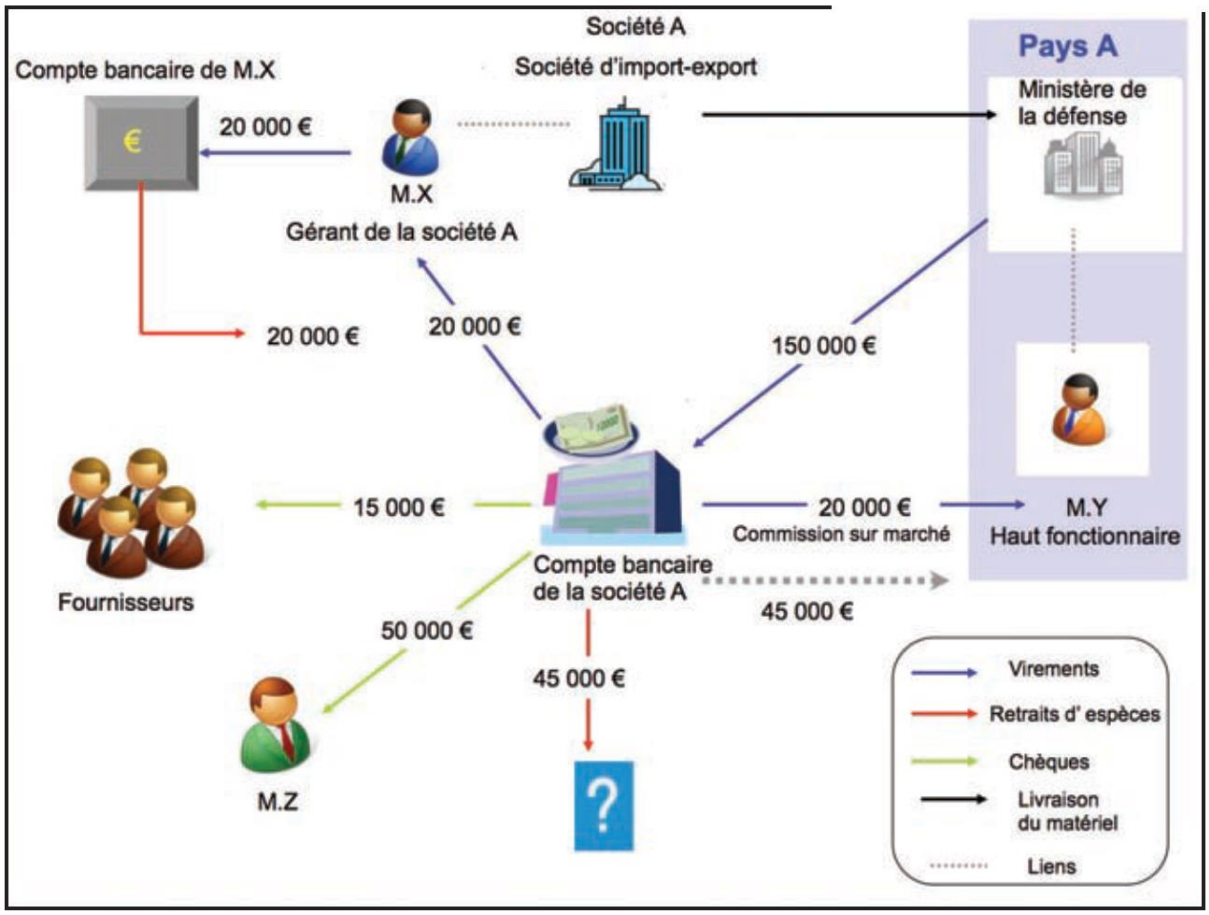


Figure 4 : Source (Tracfin)

1.4.3 Les transactions atypiques

Lorsque la première partie du KYC/AML est en place, s'en suit mise en place de la détection des transactions atypiques proprement dites.

Il est toutefois observé que la Loi anti-blanchiment précise qu'il n'appartient pas aux institutions financières de déterminer l'activité criminelle sous-jacente au blanchiment de capitaux soupçonnés. Elles ne doivent *a fortiori* pas vérifier que les éléments constitutifs des infractions pénales visées sont réunis, ni en réunir des éléments de preuve. Il suffit dès lors que leur analyse des opérations et les faits atypiques les conduisent à savoir, à soupçonner ou à avoir des raisons de soupçonner que l'une des criminalités énumérées est concernée pour qu'elles soient tenues de qualifier de suspects le fait ou l'opération atypique considérée. Dans la majorité des cas, le déclarant n'est en effet pas à même de connaître précisément l'activité criminelle sous-jacente au blanchiment de capitaux soupçonné. C'est à la l'autorité qu'il incombe de découvrir, par une analyse approfondie, le lien entre les fonds concernés, l'opération suspecte ou les faits dénoncés et l'une des formes de criminalité visées par la loi.

L'autorité joue le rôle de tri/filtre et d'enrichissement des déclarations qui lui sont adressées, permettant d'éviter que les services du parquet soient encombrés par des déclarations non pertinentes. Ceci n'empêche nullement que les déclarants peuvent faire référence à l'une ou l'autre criminalité sous-jacente lorsqu'ils savent, soupçonnent ou ont un motif raisonnable de soupçonner que les fonds blanchis sont issus de l'une ou l'autre activité criminelle.

Les termes « soupçonnent » ou « ont des motifs raisonnables de soupçonner », signifient que nous devons qualifier de suspects les fonds impliqués, l'opération concernée, ou le fait considéré si l'analyse des informations recueillies, conformément aux obligations de vigilance et en vue de l'analyse, l'amène à former une opinion de suspicion (« soupçonnent ») ou comprennent des éléments ne lui permettant pas raisonnablement d'écarter le doute (« ont des motifs raisonnables de soupçonner ») quant à la licéité de l'origine des sommes ou de l'opération ou quant à leur justification économique, juridique ou fiscale.

La Loi anti-blanchiment a ajouté à la liste des criminalités sous-jacentes la fraude sociale et la fraude informatique. Sous la notion de fraude sociale, on retrouve par exemple le travail au noir, la perception indue d'allocations, le non-respect de la règlementation relative à l'occupation de la main d'œuvre étrangère, etc.

Sous la notion de fraude informatique, et malgré que ce concept puisse déjà être englobé sous la notion d'escroquerie, la Loi anti-blanchiment vise le comportement consistant à chercher à se procurer, pour son auteur ou pour autrui, un avantage économique illégal en s'introduisant dans un système informatique, en modifiant ou effaçant des données qui y sont stockées, traitées ou transmises par un système informatique, ou en modifiant par tout moyen technologique, l'utilisation normale des données dans un système informatique.

Nous devons ainsi nous assurer de la cohérence entre l'origine et/ou la destination des fonds relatifs à une ou plusieurs opérations et les éléments de connaissance actualisée de la clientèle. Nous devons exercer une vigilance renforcée sur les transferts de fonds (virements et transmissions de fonds) en provenance ou à destination de zones géographiques considérées comme risquées en matière de terrorisme ou de financement du terrorisme ou sur les opérations effectuées dans ces zones.

Certains pays pouvant être utilisés comme pays de transit pour cacher le pays final de destination ou de provenance des fonds. Une attention particulière doit également être attachée aux schémas/motifs faisant apparaître qu'une même personne effectue sur une brève période des transferts multiples de fonds au profit de bénéficiaires localisés dans des zones géographiques à risque ou, inversement, qu'une même personne bénéficie d'un grand nombre de transferts de fonds initiés par des personnes différentes.

Le régulateur estime qu'une analyse spécifique des opérations et faits atypiques est requise à cet égard dès lors que les caractéristiques des fonds, notamment leur origine et leur destination, la nature et les caractéristiques des opérations, ou les caractéristiques des personnes impliquées dans l'opération ou la relation d'affaires, en ce compris le client,

ses mandataires ses bénéficiaires effectifs ou les contreparties des opérations présentent des liens avec les pays concernés ou des personnes ou entités connues pour leur implication dans la prolifération des armes de destruction massive.

Ces mesures complémentaires peuvent nous amener :

- à demander des informations complémentaires ou des justificatifs au client lui-même ;
- à actionner les procédures relatives au partage des informations au sein du groupe aux fins de la lutte contre le BC/FT (voir Organisation et contrôle interne au sein des groupes), en vue d'obtenir, notamment, des informations détenues par d'autres entités du groupe concernant les opérations ou relations d'affaires du même client avec ces autres entités du groupe, leur connaissance de ce même client, voire, le cas échéant, leurs éventuelles suspicions à son égard ou les éventuelles déclarations d'opérations suspectes le concernant qu'elles auraient adressées à la cellule de renseignements financiers de leur pays d'établissement ;
- à consulter des sources publiques d'informations, notamment sur internet.

A partir de cet instant ou nous connaissons le contour du KYC et de l'AML nous devons nous doter d'un dispositif d'analyse des faits ou opérations dont le caractère atypique permet une identification automatique.

Notre objectif est de pouvoir mettre en place un outil afin de considérer si on « sait, soupçonne ou avons des motifs raisonnables de soupçonner » l'opération ou le fait sont liés au blanchiment de capitaux ou au financement du terrorisme.

Lorsque cela apparaît indiqué au regard des caractéristiques du signalement, compte tenu notamment de la complexité de l'opération (transaction), du nombre d'intervenants, des montants, des fréquences, etc., ou en raison de doutes sérieux quant à la validité des informations sur lesquelles le signalement se fonde, nous avons pour objectif de réaliser une analyse en vue de vérifier que les informations directement disponibles ne

contredisent pas le caractère atypique de l'opération. Il s'agit donc d'une appréciation des circonstances sous-jacentes de l'opération. L'objectif de ce scoring d'opération est d'identifier cette transaction en partant de certains paramètres (montant, fréquences, position GPS des parties prenantes, date/heure…).

Dans la mythologie Idavoll (ou Plaines d'Ida et parfois Idavold), « Les Plaines toujours Vertes », est le lieu où les Dieux président et parlementent sur le Destin des êtres et est le centre de la cité d'Ásgard. Ce lieu a donné son nom à un module ou les transactions de notre application sont filtrées et analysées.

Ce module Idavoll-transact a pour vocation d'être un sas qui a en entrées les transactions financières et leurs paramètres (financières ou non) et en sortie une valeur variant de 0 à 1 qui permet d'identifier la qualité de la transaction. Le module Idavoll-transact est un outil de classification mais aussi un outil d'apprentissage automatique grâce à la rétropropagation. Ce module a pour vocation d'estimer, de suivre des modèles de scoring des transactions afin de détecter celles qui sont atypiques ou non.

1.5. Météo-Biz : Prévision d'activité pour les TPE

Environ 70% des entreprises sont météo-dépendantes, c'est-à-dire que leurs résultats financiers et la satisfaction de leurs clients varient selon la météo. Les mouvements sociaux telles que les grèves, difficultés liées aux transports, arrivée et départ des touristes impactent également leurs résultats au quotidien. De nombreux secteurs sont concernés : tourisme, transport, loisirs, commerce, etc…

Mon activité professionnelle repose donc sur l'analyse du lien de causalité entre :

-la météo et les activités de l'entreprise

-la grève des transports et les activités de l'entreprise

-l'heure d'arrivée et le départ des touristes

-les flux de déplacement des prospects

Météo Biz permettrait de couvrir les risques associés aux variations du climat, aux grèves ainsi que d'en exploiter les opportunités pour augmenter les ventes et favoriser

l'expérience client. Il s'agit d'un projet employant un ensemble de moyens et de méthodes de collecte et d'analyse de données dont le but est d'adapter la stratégie de la future entreprise utilisatrice à la météo et aux mouvements sociaux. Le problème est ici d'identifier les algorithmes de classification et de réseaux de neurones en vue d'exploiter des opportunités favorables et s'adapter à des situations défavorables à un fonctionnement économique optimal.

Martine Collard (2003), introduit dans son mémoire HDR 'Fouille de données, Contributions Méthodologiques et Applicatives [10] 'le domaine de la fouille de données. Elle explique de façon claire que « Restituer la compréhension est une tâche plus ambitieuse que la recherche de modèles et ouvre de nouveaux horizons de recherche. Par exemple, la manière la plus classique d'appréhender un problème de fouille de données consiste à sélectionner les tâches de modélisation et les techniques employées selon les objectifs de la personne qui analyse les données » (p. 13). Le module Idavoll-transact suit cette approche du problème en identifiant les paramètres de la modélisation (montant, fréquences, position GPS des parties prenantes, date/heure…) et des techniques utilisables dans ce contexte.

Usama Fayyad, Gregory Piatetsky-Shapiro et Padhraic Smyth en 1996 [11] nous proposent le prédicat suivant : « The value of storing volumes of data depends on our ability to extract useful reports, spot interesting event sand trends, support decisions and policy based on statistical analysis and reference, and exploit the data to achieve business,operational, or scientific goals. »

En quelques années, le domaine a très largement élargi l'étendue de ses champs d'intérêts et la manière d'appréhender ce sujet. Restituer la compréhension est une tâche plus ambitieuse que la recherche de modèles et ouvre de nouveaux horizons de recherche.

Nous pouvons classer les tâches en deux principaux groupes : description ou prédiction. Le choix des techniques dans ces deux cas dépend du type et de la structure du résultat souhaité. Parmi les différents résultats possibles, il est important de sélectionner les meilleurs. Cela passe par la mise en œuvre de différents procédés d'optimisation. Pour

l'instant, il n'existe pas encore de technique parfaite dont les performances soient meilleures, quel que soit le problème à analyser ; pour ce faire, l'analyste est obligé de procéder par une multitude d'essais et d'erreurs en applicant différentes techniques existantes jusqu'ici. Par exemple, pour réaliser la recherche d'un classifieur caractérisant au mieux les données, l'analyste pourra utiliser un algorithme d'induction d'arbre, un réseau de neurones ou encore une classification bayésienne, toutes ces techniques sans que l'on ne puisse prédire de la supériorité de l'une d'entre elles. Notre problème se résume en deux aspects : (1) d'une part la définition du meilleur classifieur ou algorithme visant à prédire l'activité économique ; (2) d'autre part la définition des fonctions de pondérations et de leurs poids individuels.

Nous avons donc travaillé sur quatre terrains différents : tout d'abord la classification des déplacements des dites visites par les touristes grâce aux données GPS, ensuite le résumé de données médicales, puis l'identification grâce a un réseau de neurone des transactions frauduleuses et enfin la prédiction d'activité économique des TPE météo dépendantes.

C'est quatre terrains ont eu pour point de départ les besoins réels d'exploitation des données dans le secteur de la petite entreprise, de l'hôpital ou de l'entreprise du secteur financier.

Leurs points communs consistent a un traitement de grand flux de données (souvent plusieurs millions de lignes) afin d'en extraire de la connaissance. Leurs différences par contre viennent de l'approche algorithmique. Dans certains cas il est possible de « perdre » de la donnée afin d'arriver au résultat « positions GPS , ou prédiction météo » mais dans d'autre cas la donnée en entrée doit figurer dans le résultat de sortie « données médicales ou financieres ».

Mais alors, forts de décennies de travail et d'expériences, bénéficiant de la progression exponentielle de la puissance de calcul, alimenté par des quantités de données (notamment satellitaires) colossales, et maintenant baignées dans la révolution de l'intelligence artificielle, pourquoi ces systèmes numériques ne fournissent-ils pas encore des prévisions de meilleure qualité. Nous tenterons d'apporter une partie de la solution dans le contexte des TPE/PME.

Les deux principaux problèmes soulevés sont le résumé de données d'une part et la rapidité de traitement d'autre part. Il n'est pas envisageable de fournir une prédiction « pour dans une heure » pour une TPE alors que le calcul lui-même dure une heure. Le résultat ne serait évidemment d'aucune utilité.

J'ai utilisé principalement des algorithmes supervisés et non supervisés en fonction de la problématique. Il faut noter que les réseaux de neurones convultifs ConvNet de Yann LeCUN dans sa publication de référence [12] ou dans son cours [13] ont été étudiés mais pas utilisés car le problème ne semble pas adapté aux réseaux de neurones convultifs.

Nous allons maintenant regarder comment ces problèmes de terrains peuvent être regardés d'un point de vue des concepts sous-jacents.

2. Corpus Théorique

Ce cadre de recherche est intrinsèquement lié à mon activité professionnelle. Les quatre thématiques que j'aborde dans ce document exploitent et adaptent toutes des méthodes liées à la classification.

2.1.Le résumé de données

Bentayeb, F., Boussaid, O., Favre, C., Ravat, F., & Teste, O. (2009) Personnalisation dans les entrepôts de données: bilan et perspectives. In EDA (pp. 7-22) [40] nous expliquent :

« La problématique est … d'inclure le processus de personnalisation par rapport à la visualisation et au résumé d'informations. Outre la définition de différentes structures de visualisation, une problématique nouvelle est l'adaptation de la structure de présentation des données en fonction des données analysées. … Grâce à cette réorganisation, les modalités d'une dimension sont agrégées selon l'ordre de leur proximité et non selon l'ordre de leur appartenance hiérarchique établi … L'objectif sera alors de prendre en compte, dans la définition et l'utilisation de structures de visualisation, des caractéristiques propres à l'utilisateur, par exemple, son niveau d'expertise. »

Cette première approche de l'adaptation de la structure des données est aussi mise en lumière par Roubens, M. (1980) [39] Analyse et agrégation des préférences : modélisation, ajustement et résumé de données relationnelles. Jorbel [44] nous décrit en utilisant l'étude de quatre problèmes, les prémices de la réduction des données aux données strictement utiles à l'analyse de notre problème. L'importance du résumé de données est perceptible dans l'utilisation CPU et mémoire afin d'obtenir un résultat dans un temps raisonnable par rapport au problème posé et des ressources informatiques disponibles.

Le jeu de données du PMSI fourni par l'ATIH (Agence de Traitement de l'information hospitalière) [41] est l'un des plus gros en France, nous l'avons présenté dans Elisabeth et Sebastien (2014) [8], suivie de la publication et Elisabeth et Sébastien (2017) [24] permettent d'appréhender les méthodes de résumé de données ainsi que leur efficacité sur des données massives, mais aussi non exemptes d'erreurs. Le papier est résumé comme ci-apres : « …

Compte tenu du volume de données importants et en croissance exponentielle produits par les établissements de santé, la fouille de données permet, grâce aux modèles exploratoires développés, une utilisation optimale de celles-ci et une compréhension de l'activité hospitalière. … »

Le résumé de donnée est parfois nécessaire pour des questions principales de performance.

Dans la publication Elisabeth, Nock, et Célimene (2013) [23] j'explique que dans le cas où il est nécessaire de fournir un résultat rapide à la demande, le résumé de données permet d'alléger le traitement au strict nécessaire. J'explique que « La croissance du volume des données à traiter ouvre la voie à de nouveaux champs d'étude. Dans certains cas, il est possible d'atténuer l'influence de la quantité de données sur les temps de réponse des traitements et algorithmes. Les traitements concernés peuvent se satisfaire d'approximations, généralement de moindre qualité que des résultats non approximatifs, mais obtenues en des temps plus acceptables.

C'est le cas de l'échantillonnage au sein d'une population ou des problèmes NPcomplets dont une solution approchée peut être obtenue par des algorithmes polynomiaux.

Le résumé de données permet de réduire un ensemble de données volumineux à un ensemble de taille plus réduite, épuré de ce que l'on considérera comme de l'information non pertinente ou non signifiante, comme du bruit. Elles sont ainsi très souvent, mais pas systématiquement, utilisées en amont des techniques supervisées ou non supervisées. Elles sont notamment très complémentaires des techniques non supervisées. »

Dans sa thèse Nock, R. (1998) [42]. explique en introduction qu' « *un bon algorithme d'apprentissage doit pouvoir trouver des formules présentant un bon compromis entre taille et adéquation aux données* ». Nous comprenons l'importance de ne pas surcharger les traitements avec des données non-pertinentes. Dans la conclusion, Elisabeth (2015) [23] l'exemple de l'utilisation au plus juste de la taille des données permet de comprendre comment une recommendation basée sur des méthodes de clustering peuvent être calculées dans des délais très courts (quasi en temps reel). « Dans le domaine de la conception de systèmes de recommandations l'effort du traitement informatique requis est minimisé par l'introduction des résumés de données. Notre approche du résumé

prend la forme de courtes séquences informatiques qui « résument » le parcours du touriste.

Ce volet applicatif a permis de mettre en lumière plusieurs approches de traitement des données. Il est l'élément de base qui permettra de proposer des recommandations aux touristes en temps réel en fonction des associations des sites « précédemment » visités, avec les sites dans lesquels son parcours s'inscrit dans le cadre du motif séquentiel.

…

Cette approche possède aussi plusieurs extensions, notamment la « clustérisations » grâce aux algorithmes k-means qui permet de créer des groupes de comportements. Ces approches couplées à une vision en temps réel selon la position et la direction du véhicule de location permet de proposer une approche géographique de l'analyse des comportements et par extension de la recommandation aux touristes. ».

2.2 Règles d'associations

Agrawal, R., Imieliński, T., & Swami, A. (1993) [5] « Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. » (pp. 207-216). Ce papier explique qu'au delà du profit, la classification par règle d'association apporte l'utilisation des règles d'association ce qui est une excellente approche aussi bien en terme de puissance de calcul nécessaire que d'adaptabilité de l'algorithme au problème. Pour Nedjema, Benlaiter (2012) [7] Apriori est un algorithme permettant de trouver des règles d'association. L'induction des règles d'association, également appelée analyse des paniers d'achats est une méthode qui a pour objectif de trouver des régularités dans le comportement des consommateurs. Ces régularités se caractérisent par des groupes de produits qui sont régulièrement achetés entre eux.

Règle d'association - Définition

Une règle d'association est une implication de la forme A ⇒ B, où A et B sont tous deux des sous-ensembles d'items de I, c'est-à-dire des itemsets. La fréquence d'un item ou d'un itemset est le nombre de transactions ou d'identifiant unique qui contient au moins une fois l'item ou l'itemset en question. Dans l'exemple de la figure 5 (paragraphe 3.1.1) , si A = {1} et que B = {2} alors la fréquence (A) = 6, la fréquence (B) = 9, et la fréquence de la règle A ⇒ B est égale à la fréquence ({1,2}) = 5. La règle A ⇒ B a un support s où s = fréquence (A ∪ B)/ n, n étant le nombre de transactions contenues de la base de données transactionnelles. Ceci peut également s'énoncer sous forme de probabilité P(A ∪ B). Dans l'exemple de la figure 5 (paragraphe 3.1.1), le support de 1 => 2 est de 50 %, car la fréquence ({1, 2}) = 5 et n = 10. La confiance, quant à elle, représente la probabilité conditionnelle.

Mc Nicholas, Murphy, T. B., & O'Regan, M. (2008) [14]

« Support, Confidence & Lift The notation P(A), P(A, B) and P(B | A) is used in the usual way and the afore-mentioned functions are defined in Table 1.

Table 1
Functions of association rules.

Function	Definition
Support	$s(A \Rightarrow B) = P(A, B)$ and $s(A) = P(A)$
Confidence	$c(A \Rightarrow B) = P(B \mid A)$
Expected Confidence	$EC(A \Rightarrow B) = P(B)$
Lift	$L(A \Rightarrow B) = c(A \Rightarrow B)/P(B) = P(A, B)/(P(A)P(B))$

Le lift represent la measure de la distance entre P (B | A) et P (B), ou de manière équivalente, la mesure dans laquelle A et B ne sont pas indépendants.

La plage de valeurs que peut prendre l'élévation d'une règle d'association A ⇒ B est limité par les valeurs respectives de P (A) et P (B) ;

max{P(A) + P(B) − 1, 1/n} P(A)P(B) ≤ L(A ⇒ B) ≤ 1 max{P(A), P(B)} , (1)

où n est le nombre de transactions, τi.

Pour Feno, (2007) [15] une mesure est dite descriptive, si elle ne change pas en cas de dilatation des données, dans le cas contraire, elle est dite mesure statistique. Donc, pour

une mesure statistique μ, la taille de données n doit intervenir dans son évaluation. Pour une mesure statistique, en fixant les quantités marginales p(X0) et p(Y 0), il est intéressant de savoir comment évaluer la règle X→Y si on augmente la taille de données n. Si une mesure varie de façon croissante avec n et admet une valeur maximale, alors elle risque de perdre son pouvoir discriminant quand n devient suffisamment grand. Il est important de mesurer la qualité des donnée mais aussi la taille de ces données. Les règles évoluent en fonction de la taille des données. Ceci est particulièrement sensible dans les dizaines de millions d'enregistrement du PMSI.

La generation des règles est beaucoup moins couteuse que la generation des motifs frequents, car il n'est plus nécessaire de faire le parcours coûteux de la base des données.

Pour générer les règles d'association, on considère l'ensemble F des motifs frequents trouves dans la phase précédente. Pour chaque motif fréquent I, on considère tous ses sous-ensembles (tous frequents d'après la propriété d'antimonotonicite). A partir de ces sous-ensembles ` frequents, on génére toutes les règles I1→I \ I1(I1 ⊂ I) telles que leurs Confiances respectives dépassent le seuil minimum de Confiance. Agrawal et Srikant ont propose dans [5] une optimisation de la generation des règles d'association. Cette optimisation est basée sur la proposition suivante. Elle permet de ne pas considérer tous les ensembles possibles des motifs frequents (Guillaume et Papon, 2013). [16]

Dans (Antonie et Zaïane, 2004) [47], les auteurs génèrent à la fois les règles positives et les règles négatives. Comme cet algorithme repose sur l'algorithme Apriori (Agrawal et Srikant, 1994), il commence par rechercher les motifs candidats, c'est-à-dire les différentes combinaisons entre les items de la base : {x1, x2}, {x1, x3}, .., {x1, x2, x3} ...), puis à partir des motifs candidats vont générer les différentes règles. Cette recherche des règles s'opère à partir des motifs candidats et non pas à partir des motifs fréquents comme dans l'algorithme fondateur Apriori.

Une règle d'association positive ou négative (RAPN) valide est une expression du type

$C1 \Rightarrow C2$ où $C1 \in \{X, X, X^{¨}\}$, $C2 \in \{Y, Y , Y^{¨} \}$, $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \emptyset$,

$C1 = X^{¨} \Leftarrow\Rightarrow C2 = Y^{¨}$, et telle que

(Ct1) : minsup 6 sup(XY) 6 maxsup,

(Ct2) : minsup¨ 6 sup(X¨Y¨),

(Ct3) : sup(C1 ⇒ C2) > minsup si (C1, C2) 6= (X, Y) et (C1, C2) 6= (X, ¨ Y¨),

(Ct4) : conf(C1 ⇒ C2) > minconf,

(Ct5) : MG(C1 ⇒ C2) > minMG,

(Ct6) : C1 ⇒ C2 est minimal au regard des motifs négatifs raisonnablement fréquents X ou Y

La contrainte (Ct6) est présente lorsque les motifs C1 et C2 sont des motifs raisonnablement fréquents négatifs X et Y, ceux-ci doivent également être minimaux, c'est-à-dire qu'il n'existe pas par exemple pour le motif X, un sous-ensemble X0 (X tel que X0 soit également raisonnablement fréquent).

Nous présentons l'algorithme avec au préalable la définition de toutes les notations utilisées.

– BD : base de données ;

– I : ensemble des items ;

– I : ensemble des négations d'items ;

– X, Y : motifs positifs ;

– X, Y : motifs négatifs ;

– X, ¨ Y¨ : conjonctions de motifs négatifs ;

– i : item ou 1-motif ;

– minsup, maxsup : seuils respectivement minimum et maximum pour le support des motifs positifs ;

– minsup : seuil minimum pour le support des motifs X¨ ;

– minconf, minMG : seuils minimum pour respectivement la confiance et la mesure MG ;

– taille(X) : nombre d'items composant un motif ;

– R : ensemble des Règles valides ;

– RF : ensemble des motifs Raisonnablement Fréquents ;

– NRFM : ensemble des motifs Négatifs Raisonnablement Fréquents Minimaux ;

– NNRFk : ensemble des k-motifs Négatifs Non Raisonnablement Fréquents ;

– CPk : ensemble des k-motifs Candidats Potentiels ;

– Ck : ensemble des k-motifs Candidats ;

– Fk : ensemble des k-motifs Fréquents X et X¨ ;

– F : ensemble de tous les motifs Fréquents ;

– s : support de la règle étudiée

– c : confiance de la règle étudiée

– m : mesure MG de la règle étudiée

L'algorithme d'extraction des RAPN (voir l'algorithme 1) commence par rechercher les motifs raisonnablement fréquents grâce à la fonction funct_RF (ligne 1).

Cette recherche est similaire à celle exposée dans (Agrawal et Srikant, 1994)[48] pour générer les motifs fréquents en rajoutant deux contraintes supplémentaires :

(1) un seuil maximal maxsup qui ne doit pas être dépassé par le support de X ;

(2) un seuil minimal minsup¨ pour le support des motifs X¨ ; ce qui permet de vérifier les contraintes (Ct1) et (Ct2).

Cette fonction sera développée dans la section 4.1 grâce à l'algorithme 2.

La fonction funct_NRFM (ligne 2 de l'algorithme 1) exposée dans l'algorithme 3 commence par initialiser l'ensemble recherché NRFM des motifs Négatifs Raisonnablement Fréquents Minimaux à l'ensemble de toutes les négations d'items de taille 1 raisonnablement fréquentes (ligne 1). Cette initialisation est rendue possible grâce à la connaissance des supports des items i calculés lors de l'exécution de la fonction funct_RF (ligne 1 de l'algorithme 1) puisque nous avons la relation suivante : sup(i) = 1 − sup(i). Puis, nous stockons dans l'ensemble NNRF1 les items négatifs non raisonnablement fréquents (i.e I \ NRFM) auquel on enlève les négations d'items i dont le support est supérieur au seuil maximal (ligne 2) car tout sur-ensemble {i, j} aura une valeur de support supérieure à celui de i ou de j. Ensuite, on génère l'ensemble C2 des motifs candidats de taille 2 à partir de l'ensemble NNRF1 des négations d'items non raisonnablement fréquentes (ligne 3). Le processus suivant va être réitéré jusqu'à ce que

l'on n'arrive plus à générer de candidats (Ck 6= ∅) (lignes 4 à 16). On commence par initialiser l'ensemble NNRFk des motifs X ayant un support inférieur à minsup (car si le support de X est supérieur à maxsup alors tout sur-ensemble XY aura un support encore plus élevé) à l'ensemble vide (ligne 5). On parcourt tous les motifs candidats X (ligne 6) afin de détecter ceux qui sont raisonnablement fréquents (lignes 7 et 8). Si ce n'est pas le cas, on s'assure que X n'a pas un support supérieur à maxsup (par conséquent on teste si son support est inférieur à minsup) (ligne 10) et on le stocke dans NNRFk (ligne 11) comme motif pouvant générer au niveau supérieur (k + 1) un motif candidat (ligne 15).

Pasquier et Lakhal (2000) [43] proposent une nouvelle sémantique basée sur la connexion de Galois pour le problème de l'extraction de règles d'association. Utilisant les opérateurs de fermeture de la connexion de Galois, ils dénissent les itemsets fermés qui forment le treillis des itemsets fermés et les itemsets fermés fréquents.

Ce sont ces notions de règles d'associations que nous utilisons dans le volet Géotourisme pour identifier les interactions entre deux sites, c'est-à-dire le fait qu'un touriste visite deux sites touristiques. Puis, dans PMSI nous les utilisons pour faire la relation lors du passage de la tarification en dotation globale à la tarification à l'activité. En 2020 suite à la crise du Covid-19, la tarification en dotation globale est de nouveau d'actualité.

2.3 Motifs séquentiels

> *« Determine whether sequential pattern mining is effective for identifying temporal relationships between medications and accurately predicting the next medication likely to be prescribed for a patient. »* *(*Wright, Wright, McCoy et Sittig, 2015) [17].

Les motifs séquentiels ont étés utilisés en bio-informatique notamment pour l'identification de séquences temporelle. Notre approche a été de les utiliser dans le cadre de la temporalité de la succession des visites des sites touristiques (volet géotourisme), mais aussi dans l'analyse du PMSI (même si on s'éloigne de la question de la bio informatique citée) , mais plutôt dans le volet de l'extraction de données dans des flots de données très importants

En termes formels, soit I = { i 1 , i 2 ,..., i m } un ensemble d'items, par exemple (metformine, simvastatine , venlafaxine).

Soit la séquence s , notée 〈s 1, s 2,..., sn〉 une liste ordonnée temporellement d'ensembles d'items, par exemple 〈(metformine, simvastatine, venlafaxine), (aspirine, glipizide), (hydrochlorothiazide, insuline)〉.

Soit a une autre séquence notée 〈(metformine), (glipizide), (insuline)〉.

La séquence a est appelée une sous-séquence de la séquence s puisque (metformine) ⊆ (metformine, simvastatine, venlafaxine) et (aspirine) ⊆ (aspirine, glipizide) et (insuline) ⊆ (hydrochlorothiazide, insuline).

Une séquence de données est une liste de transactions avec le même ID de séquence (c'est-à-dire toutes les transactions appartenant à un patient). Le support de la séquence a est la fraction de séquences de données qui contiennent une sous-séquence. Par exemple, les deux séquences de données du tableau 1 contiennent la séquence (metformine, insuline) mais seulement 1 sur 2 contient la séquence (metformine, glipizide , insuline) donc s'il s'agissait de l' ensemble de données complet, le support de (metformine, insuline) serait de 1 et le support de (metformine, glipizide, insuline) serait de 0,5. La tâche de l'exploration séquentielle de modèles est d'identifier les séquences fréquentes, où fréquent est défini comme ayant un support au-dessus d'un seuil défini par

l'utilisateur. Dans cet article, nous ferons référence aux séquences fréquentes extraites de séquences de données en tant que modèles séquentiels extraits, et nous nous référerons à la séquence de données d'un patient (l'historique ordonné dans le temps de tous les médicaments prescrits à ce patient) comme une séquence d'un patient.

> *"Tracing symptoms over a patient history may indicate symptoms that tend to go together or regularly follow each other in thrombosis patients, and so analysis was performed using the association rules and CaPri algorithms. The first tells what tend to co-occur, the second uses timing information to indicate what tends to follow another symptom. »*
> Jensen, S., & SPSS, U. (2001,) [18]

L'utilisation des motifs séquentiels est aussi active de puis une 20e d'années dans le cadre de la recherche de séquences pouvant identifier des thromboses. La relation entre ces deux étapes du processus d'exploration de données itérative, j'inclus les deux dans la même section pour la brièveté de ce document. Beaucoup de nettoyage et de reformatage des données ont été effectués pour rendre l'analyse possible et plus interprétable, y compris le reformatage des dates en un type cohérent, l'extraction numérique des informations issues de champs numériques contenant quelques symboles (c'est-à-dire <5000 comme entrée), etc. Un aspect heureux de Clémentine est que toutes les manipulations sont clairement visibles dans le flux, contenu dans la série de nœuds menant du (des) nœud (s) de source de données à les nœuds de sortie, qui peuvent être un tableau, un graphique ou une série de modèles. Dans Clémentine, comme un atelier d'exploration de données, un enregistrement visuel du processus d'exploration de données est auto-entretenu.

Garofalakis, M. N., Rastogi, R., & Shim, K. (1999, September). SPIRIT: Sequential pattern mining with regular expression constraints. In VLDB (Vol. 99, pp. 7-10). [19]

Les champs numériques ont été regroupés individuellement en faible / moyen / élevé, en utilisant des histogrammes. Les bandes ont été créées pour avoir approximativement le même nombre d'individus dans chacun. Où les champs numériques ont été donnés contenant des symbols exceptions (telles que <31,2), le «<» a été supprimé et la (valeur –1) a été arbitrairement utilisé à sa place. Cela a des problèmes inhérents mais a été utilisé comme un général raisonnable guider. En l'absence de résultat, un «vide» a été attribué au champ.

Un réseau de neurones de rétro-propagation a été exécuté sur toutes les variables possibles du laboratoire examens, pour voir quels résultats étaient les plus révélateurs de la présence / absence de thrombose.

Lorsque les variables numériques contenaient des exceptions symboliques, les nombres ont été extraits afin que toutes les variables possibles soient traitées comme des nombres. Les variables symboliques ont été laissés tels que présentés à l'origine pour être utilisés dans le modèle.

”Les motifs séquentiels ont été introduits par Agrawal et Srikant (1995)[BIB] et peuvent être considérés comme une extension du concept de règle d’association en prenant en compte la temporalité associées aux itemsets. La recherche de motifs séquentiels consiste à extraire des ensembles d’items couramment associés au cours du temps. Dans le contexte du «panier de la ménagère », un motif séquentiel peut par exemple être : «60% des clients achètent une télévision, puis achètent plus tard un lecteur de dvd » Masseglia Florent et Teisseire M. et poncelet Pascal (2004) [20](page 2).

J’ai eu l’opportunité d’échanger avec Florent Masseglia à Sophia sur son papier et notamment la question des motifs séquentiels. Son approche pragmatique m’a permis d’optimiser l’algorithme développé pour Géotourisme. L’effort réalisé dans l’optimisation de l’implémentation de l’algorithme impacte le temps de réponse donc le fait que l’outil puisse devenir utilisable avec des délais de réponse acceptable dans le cadre contrain du résumé de données des positions GPS arrivant au fil de l’eau.. Les résultats dans le papier. Elisabeth, Erol, Richard Nock, and Fred Célimene. "2013 [22]” résultent de son approche des motifs séquentiels sans utiliser de “lift”.

Cette approche a permis un traitement en quasi temps reel pour la classification d’un nouveau parcours d’un nouveau touriste entrant dans le système.

Les méthodes que nous avons proposées et expérimentées sur des jeux de données réels permettent d’extraire des motifs représentatifs à partir de données issues de capteurs. Les différents types de représentations des données sont suffisamment adaptables pour s’accorder avec les diverses caractéristiques des données que l’on peut rencontrer, comme nous l’avons fait pour les données ferroviaires. Ainsi, dans certains domaines, des relevés sont enregistrés très fréquemment et décrivent pourtant des comportements qui évoluent peu avec le temps. Il est alors nécessaire de mettre l’accent sur l’agrégation de relevés pour résumer les données (en modifiant le paramètre minSim). De même, dans le

cas où un nombre de capteurs trop élevé entraînerait un volume de données trop grand et trop redondant pour permettre l'extraction de connaissances utiles, l'agrégation de capteurs en adoptant un certain niveau d'abstraction est adéquate. Cette adaptabilité s'étend également aux connaissances recherchées. Par exemple, dans certains domaines, les simples valeurs mesurées par les capteurs ne permettent pas d'obtenir des connaissances satisfaisantes, car elles sont trop dépendantes de conditions externes. C'est le cas avec les données ferroviaires, où les températures mesurées sur certains composants sont dépendantes à la fois du comportement du train, mais aussi de facteurs tels que la température externe. Ne disposant pas d'informations à propos de ce facteur dans nos données ferroviaires, les températures mesurées sont peu caractéristiques 5. En revanche, les variations de ces températures jouent un rôle important. Cet exemple montre la nécessité d'obtenir des connaissances complémentaires pour répondre aux besoins propres au diagnostic de pannes.

Elisabeth, Erol, Richard Nock, and Fred Célimene. 2013 [22]

"Extraction of sequential patterns [2] and [5] make possible the discovery of temporal relations between 2 sites. »

Dans cet article, nous avons appliqué un ensemble d'algorithmes d'exploration de données à l'aide de données collectées à partir de GPS de suivi installés dans des voitures de tourisme de location. Organisations touristiques et les agences pourraient se pencher sur ces applications pour trouver le meilleur moyen d'extraire connaissances de leurs propres systèmes de base de données. Les entreprises de suivi GPS peuvent également trouver des idées pour améliorer l'utilisation de leurs données collectées.

Les motifs séquentiels utilisées dans le volet Géotourisme pour identifier les interactions entre deux sites, c'est-à-dire le fait qu'un touriste visite deux sites touristiques et dans PMSI pour faire la relation lors du passage de la tarification en dotation globale à la tarification à l'activité. Suite à la crise du Covid-19, la tarification en dotation globale est de nouveau d'actualité en 2020.

L'exploration séquentielle de modèles est une technique d'exploration de données utile pour identifier les relations temporelles entre les médicaments. À partir de simples séquences à deux éléments, les voies de traitement médicamenteux peuvent être visualisées en fonction des lignes directrices pour un traitement médicamenteux par étapes. Ces relations temporelles sont utiles pour faire des prédictions sur le médicament

qu'un prescripteur est susceptible de choisir ensuite lors du traitement d'une maladie évolutive telle que le diabète. Des travaux futurs sont nécessaires pour optimiser l'utilisation de l'exploration séquentielle de modèles pour détecter les relations temporelles entre les éléments du dossier médical et améliorer les soins aux patients.

2.4 Kmeans

« KMA is the most popularly used algorithm to find a partition that minimizes SE measure » (Krishna, et Murty, 1999). [45].

L'algorithme Kmeans permet de façon simplifiée de pouvoir implémenter des kmeans. L'apport de mes travaux est principalement sur la performance dans le cadre de notre problématique. L'initialisation des items dans les clusters n'est plus aléatoire mais est liée dans le cadre de Géourisme à la positition de la voiture. Cette initiatialisation minimise drastiquement les calculs liés aux déplacements des items afin de trouver le « plus proche ».

Comme en GA, GKA maintient une population de solutions codées. Le la population est initialisée au hasard et évolue au fil des générations; la population de la génération suivante est obtenue en appliquant la génétique opérateurs sur la population actuelle. L'évolution a lieu jusqu'à ce qu'un la condition de fin est atteinte. Les opérateurs génétiques utilisés dans GKA sont la sélection, la mutation basée sur la distance et l'opérateur Kmeans.

Dans cette section, nous expliquons GKA en spécifiant les schémas de codage et d'initialisation et l'opérateur génétique.

Il a été montré en utilisant la théorie des chaînes de Markov finies que les algorithmes génétiques canoniques convergent vers l'optimum global [14].

Nous prouvons la convergence globale de GKA dans le même sens en dérivant des conditions sur les paramètres de GKA qui garantissent la convergence. Considérons le processus fP (t) gt 0, où P (t) représente la population maintenue par GKA à la génération t. L'espace d'état de ce processus est l'espace de toutes les populations possibles S et les états peuvent numéroté de 1 à jSj. Comme mentionné précédemment, l'espace d'état est limité aux populations contenant des chaînes légales, c'est-à-dire des chaînes représentant des partitions avec K clusters non vides. De la définition de GKA, P (t + 1) peut être déterminé complètement par P (t), c'est-à-dire,

Pr{P(t) = ptjP(t -1) = pt-1; ; P(0) = p0}

= PrfP(t) = ptjP(t 1) = p

Cette publication est la plus importante concernant Géotourisme, elle détaille l'utilisation des motifs séquentiels, kmeans, et règles d'association. Cette publication explique grâce aux différents résultats l'utilisation des algprithmes avec un groupe fermé d'utilisateurs. Elisabeth (Mai 2015) [22]. Cette publication explique que ces travaux visent à trouver des solutions au problème du volume de données, de l'explosion combinatoire notamment pour les règles d'associations et les motifs séquentiels. Il est nécessaire de faire un premier résumé de données. Ce résumé de donné consiste à définir ce qu'est un "arrêt" et à le coder en terme informatique de la façon suivante :

- numéro automatique d'insertion de ligne (numauto = ex 5)

- numéro de boitier gps embarqué (numbeepway = ex 344691000067372)

- l'ip publique du boitier gps (ip = ex 193.251.163.1)

- la date et l'heure de la position collectée (date heure = ex 2009-12-22 00 :00 :3)

- la latitude de la position collectée (latitude = ex 1436.0512)

- l'orientation NS de la position collectée (ns = ex N)

- la longitude de la position collectée (longitude = ex 06056.1889)

- l'orientation EW de la position collectée (ew = ex W)

- la hauteur de la position collectée (hauteur = ex 138)

- la vitesse de la position collectée (vitesse = ex 16.01)

- des informations complémentaires de la position collectée (info =; ;20.00) (cumul en km depuis le dernier démarrage)

Une « trame » peut donc avoir la forme suivante :

5 344691000067372 193.251.163.1 2009-12-22 00 :00 :37 1436.0512 N 06056.1889 W 138 16.01 ; ;20.00

Un trajet étant la succession des arrêts entre le premier arrêt à partir de la première location jusqu'au retour au parking du loueur. Un premier traitement a été réalisé afin de générer pour les arrêts collectés (position longitude, latitude) une table de succession des arrêts pour chaque trajet.

Il est possible après de manipuler ces données très simplement avec les langages de programmations traditionnels.

La publication E Elisabeth, N Sébastien (2017) [8], est une synthèse des travaux sur le PMSI, il met en évidence la difficulté et les orientations qui ont été retenues pour les résumés de données et l'extraction de données.

Ces travaux se poursuivent avec une série de données encore plus importantes notamment en y ajoutant aux données MCO les données de la psychiatrie PSY, de l'hospitalisation à domicile HAD et de soin de suite et de réadaptation jusqu'au 31 Décembre 2019. Il existe peu de recherche sur l'efficience économique des établissements de santé réalisé par des informaticiens. Ces travaux ont en général une approche médicale, ou économie de la santé. Il existe peu de littérature sur les approches non supervisées. L'approfondissement de ces travaux est en cours grace aux données accessibles avec plus de profondeur (2009 à 2019) et dans les 3 autres domaines HAD, PSY, SSR en plus du MCO traité.

2.5 Logique floue

Le projet MétéoBiz a besoin d'une fonction permettant d'activer les neurones de son reéseau de neurones. Plusieurs approches ont été explorées afin d'identifier une fonction permettant de service de neurone d'activation.

L'analyse de stabilité est expliquée dans Bühler, H. (1994) [35].

Il fixe les bases du réglage par logique floue de manière consolidée et présente des résultats inédits, en particulier en ce qui concerne l'analyse de stabilité.

Les modèles en logique floue ont été utilisés dans la gestion de la prediction météo dans les papiers, Hello, G. (2002) [46].

Dans sa these Hello montre que les modèles de prévision numérique sont extrêmement "sensibles à la précision de leurs conditions initiales. Les dépressions météorologiques et les tempêtes restent ainsi difficilement prévisibles en dépit du réalisme des modèles et de leurs états initiaux. …Cette voie, appelée l'observation adaptative, … La prévisibilité d'une dépression est matérialisée par des champs de sensibilité, le vecteur gradient d'une propriété scalaire caractéristique d'une tempête prévue autour de 36 ou 48 heures."

Prise en compte de la dynamique associée aux dépressions des latitudes moyennes dans la détermination des conditions initiales des modèles météorologiques (Doctoral dissertation, Toulouse 3) [28] et Monbet, V. (2009) [29]

Notre approche est de ne pouvoir l'utiliser que dans la fonction d'activation du neurone du réseau de neurone de kohonen que nous verrons dans le point 2.6. D'autres utilisations de la modélisation avec des outils en logiques floues ont donnés de très bons résultats dans le domaine de la prédiction météo.

Exemple: Daoud, A. B. (2010)[31] et Lackner, J. R., & Dizio, P. (1994)[33] et Lynch, P. (2008)[27]

Dans ces trois papiers utisent une approche pragmatique pour la prévision météorologique. La logique floue est utilisée sous la forme d'équation simplifiée permettant d'avoir des seuils rapidements calculés.

Notre approche nous permet d'utiliser de façon efficace (en réduisant les calculs), donc en approchant des temps de réponses compatibles avec la problématique dans le projet Météo Biz.
La logique floue est un type de modélisation qui s'intéresse à la prédiction d'une variable catégorielle *Y* « subjective » au sens où elle n'est pas objectivable: elle dépend de l'observateur (l'individu est « grand », « moyen » ou « petit »). Ce cadre sort de la statistique classique dans lequel la valeur de la variable *Y* est objectivable (« l'individu mesure 176 cm »). L'application de la logique floue revient à tenter d'appliquer un raisonnement proche de la pensée humaine:

- Les variables prédictives (comme la variable à prédire) sont catégorielles avec des modalités subjectives (« grand », « petit ») et non pas de données objectivables (176 cm). Ces variables catégorielles sont appelées « variables linguistiques ». Dans le cadre statistique usuel, la variable continue initiale (ici la taille en cm) peut être discrétisée pour donner des intervalles distincts, par exemple : « petit < 170cm < moyen < 180cm <grand ». La logique floue vise à prendre en compte les incertitudes qui existent au voisinage des seuils (due en partie à des principes de subjectivité).
- Une donnée peut appartenir à plusieurs modalités d'une même variable (un individu de 165 cm peut être considéré comme petit mais aussi comme moyen). Les classes définies ne partitionnent donc pas l'ensemble des possibles car elles peuvent se recouper.
- La logique floue intègre un ensemble de règles permettant d'attribuer (d'une manière logique) une sortie à une entrée.

La logique floue permet donc d'intégrer des systèmes experts dans des processus automatisés. Ce point constitue à la fois une force et une faiblesse de la logique floue. Le graphique de véracité suivant montre qu'un individu de 162 cm peut être considéré en logique floue comme étant petit à 60% et moyen à 40%. Au-delà de cette différence de principe, elle intègre également une prise en compte des interactions différentes de celle du monde probabiliste en redéfinissant les opérateurs logiques.

Son fonctionnement peut se résumer en trois grandes étapes:

- La **fuzzification** transforme les variables chiffrées en variables floues (aussi appelées variables linguistiques) en leur associant des lois de véracité (la variable taille est divisée en modalités « un individu de taille 162 cm est « petit » à 60%, « moyen » à 40% et « grand » à 0% »). Ce procédé s'apparente à la définition de lois a priori en statistiques bayésiennes, avec dans cette exemple une loi a priori (0,6 ; 0,4 ; 0). La différence dans ce cadre est que la somme des véracités n'est pas tenue de valoir 1.
- L'**inférence floue** construit les règles (et les résultats) basées sur les variables linguistiques, attribution d'une véracité à chaque règle, puis agrégation des règles pour obtenir un résultat (linguistique) unique.
- La **defuzzification** passe d'un résultat linguistique à un résultat chiffré.

2.6 Réseaux de neurones de kohonen et (Réseaux de neurones convolutifs non utilisés)

Teuvo Kohonen a publié plusieurs essais et une quantité impressionnante de plus de 300 articles. Sa contribution importante qui a dirigé notre travail concerne la carte auto adaptative dite « carte de Kohonen » (Kohonen, 1982) [49] ; Kohonen et Honkela, 2007) [50].

Les réseaux Kohonen permettent d'apporter une solution à la problématique de Météo-biz dans cette première approche. La combinaison de neurones d'activation utilisées a travers la logique floue et des réseaux de Kohonen (en rétropropagation) apportent une solution satisfaisante pour cette première étude.

Plus récemment, j'ai étudié avec beaucoup d'intérêt comment utiliser les réseaux Convultifs décrits dans les travaux de Yann LeCun; (LeCun, Yann, et al., 1998 [46] ; LeCun, et Bengio, 1995 [47]). Les réseaux Kohonen permettent d'apporter une solution à la problématique de Météo-biz dans cette première approche. La combinaison de neurones d'activation utilisées a travers la logique floue et des réseaux de Kohonen (en rétropropagation) apportent une solution satisfaisante pour cette première étude.

Etant persuadé que la performance des réseaux convultifs peut être une approche performante pour le sujet MéteoBIz. Il serait nécessaire d'étudier son adaptation au problème Météo-biz. L'objectif étant d'évaluer la possibilité d'avoir des valeurs d'entrée compatibles avec la structure des convultifs. Cette problématique sera développé dans une thèse CIFRE commanditée par notre société BEEPWAY et dirigée par l'Institut des Mines Télécom en 2020-2023.

Les réseaux de neurones convolutifs désignent une sous-catégorie de réseaux de neurones : ils présentent donc toutes les caractéristiques listées ci-dessus. Cependant, les CNN sont spécialement conçus pour traiter des images en entrée. Leur architecture est alors plus spécifique : elle est composée de deux blocs principaux.

- Le premier bloc fait la particularité de ce type de réseaux de neurones, puisqu'il fonctionne comme un extracteur de features. Pour cela, il effectue du template

matching en appliquant des opérations de filtrage par convolution. La première couche filtre l'image avec plusieurs noyaux de convolution, et renvoie des "feature maps", qui sont ensuite normalisées (avec une fonction d'activation) et/ou redimensionnées.

- Ce procédé peut être réitéré plusieurs fois : on filtre les features maps obtenues avec de nouveaux noyaux, ce qui nous donne de nouvelles features maps à normaliser et redimensionner, et qu'on peut filtrer à nouveau, et ainsi de suite. Finalement, les valeurs des dernières feature maps sont concaténées dans un vecteur. Ce vecteur définit la sortie du premier bloc, et l'entrée du second.
- Le second bloc n'est pas caractéristique d'un CNN : il se retrouve en fait à la fin de tous les réseaux de neurones utilisés pour la classification. Les valeurs du vecteur en entrée sont transformées (avec plusieurs combinaisons linéaires et fonctions d'activation) pour renvoyer un nouveau vecteur en sortie. Ce dernier vecteur contient autant d'éléments qu'il y a de classes : l'élément i représente la probabilité que l'image appartienne à la classe i. Chaque élément est donc compris entre 0 et 1, et la somme de tous vaut 1. Ces probabilités sont calculées par la dernière couche de ce bloc (et donc du réseau), qui utilise une fonction logistique (classification binaire) ou une fonction softmax (classification multi-classe) comme fonction d'activation.

Comme pour les réseaux de neurones ordinaires, les paramètres des couches sont déterminés par rétropropagation du gradient : l'entropie croisée est minimisée lors de la phase d'entraînement. Mais dans le cas des CNN, ces paramètres désignent en particulier les features des images.

3 Géotourisme

Géotourisme a pour objectif de fournir une méthode pour la mise en place d'un système de recommandation basée sur la collecte de données de positionnement GPS de véhicules loués pas les touristes dans le cadre d'un apprentissage des déplacements de créer des classes représentants les caractères de ces déplacements.

3.1 Géotourisme - volet applicatif

Un loueur de voitures a accepté d'installer 12 systèmes GPS dans ses véhicules destinés à la location touristique en Martinique. Les véhicules ont roulés pendant plus de 18 mois. Les positions ont été collectées, et un résumé de données (arrêts des véhicules) a été conservé. Des points d'intérêt (environ 540) ont été géocodés (association site/point géographique).

Cet aspect de notre recherche concerne le développement d'outil permettant la modélisation non-supervisé des parcours touristiques. L'inclusion d'appareil mobile, aujourd'hui appelés « appareils connectés », dans notre étude permet une utilisation et une extension des algorithmes informatiques hors du champ traditionnel de l'internet. Dans cette situation, la fouille de données se propose de donner les outils et/ou techniques nécessaires pour l'extraction de ces connaissances. Deux classes de motifs se sont alors avérées très utiles et simultanément utilisées dans la pratique, à savoir d'une part les *itemsets* fréquents et d'autre part les règles d'association. Un *itemset* est une conjonction d'items relatifs au contexte d'extraction alors qu'une règle d'association est une expression causale avec parties prémisses, conséquentes et probabilistes ayant une fréquence ou support, une force ou confiance des co-occurrences entre les items de la prémisse et ceux de la conclusion. Toutefois, l'ensemble de tous les itemsets fréquents et de toutes les règles valides (par rapport aux mesures de support et de confiance) extrait à partir des contextes réels est généralement de taille importante, dont une bonne partie est redondante. Pour pallier à cette situation, un nombre important de travaux proposent d'extraire seulement un sous-ensemble représentatif, appelé représentation concise.

Nous détaillons ici, les notions relatives à ces deux classes de motifs et donnons un aperçu des principales approches permettant de réduire la taille des ensembles extraits. Le cadre de cet exposé concerne l'extraction des connaissances à partir des données (ECD).

Le but d'un processus d'ECD est d'extraire des connaissances qui sont non triviales, potentiellement utiles et significatives. Plus précisément, nous investiguons donc ces deux classes de motifs qui se sont avérées très utiles dans l'ECD : les itemsets fréquents et les règles d'association. L'extraction de connaissances dans les bases de données, également appelé *data mining*, désigne le processus non trivial permettant d'extraire des informations et des connaissances utiles qui sont enfouies dans les bases de données, les entrepôts de données (data warehouses) ou autres sources de données.

Nous traitons des problèmes de la génération efficace des règles d'association, de la pertinence et de l'utilité des règles d'association extraites. Une règle d'association est une implication conditionnelle entre ensembles d'attributs binaires appelés items. Dans l'ensemble des travaux existants, l'extraction de règles d'association est décomposée en deux sous-problèmes qui sont la recherche des ensembles fréquents d'items et la génération des règles d'association à partir de ces ensembles.

Le premier sous-problème, dont la complexité est exponentielle dans la taille de la relation et qui nécessite de parcourir (donc de réaliser des boucles) à plusieurs reprises celle-ci, constitue la phase la plus coûteuse en termes de temps d'exécution et d'espace mémoire.

Nous proposons une nouvelle approche pour le problème de l'extraction des règles d'association. Les résultats expérimentaux présentés dans le point 3.1.2 démontrent que ces algorithmes permettent de réduire les temps d'extraction et l'espace mémoire nécessaire dans le cas de jeux de données constitués de données denses ou corrélées. Nous proposons d'améliorer la pertinence et l'utilité des règles d'association extraites en limitant l'extraction à des bases pour les règles d’association. Nous proposons également des algorithmes efficaces de génération de ces bases.

3.1.1 Les règles d'associations

Les règles d'associations utilisent des méthodes de génération combinatoires et engendrent un nombre élevé de règles qui sont difficilement exploitables. Plusieurs approches de réduction de ce nombre ont été proposées comme l'usage de mesure de qualité, le filtrage syntaxique par contraintes, la compression par les bases représentatives ou génériques. Cependant, ces approches n'intègrent pas l'expert dans le déroulement du processus limitant ainsi l'aspect interactif du processus. En effet, l'expert ne sait pas toujours initialement quelle connaissance il souhaite obtenir. Nous analysons l'activité cognitive de l'expert dans différents processus de recherche de règles d'association et nous montrons que dans ces approches, l'expert n'intervient pas durant les tâches du processus. Pour accroître cette interactivité avec l'expert, il est nécessaire que celui-ci soit au coeur du processus afin de répondre à l'un des objectifs de l'ECD.

L'utilisation d'une interface graphique adaptée s'avère donc nécessaire pour que l'expert puisse interagir de manière optimale avec le processus. L'efficacité de cet algorithme a été montrée sur un problème réel de marketing (référence) faisant intervenir des experts du monde touristique. Dans cet exemple, l'efficacité est expliquée par la pertinence résultat. En effet la représentation graphique des règles d'associations montre clairement des groupes des différents types de touristes par types de visites (musées ou plages ou commerce de ville). Cet algorithme est aussi rapide en termes d'exécution avec une consommation très modérée de ressources informatique (cpu, mémoire).

L'intérêt des règles d'associations est qu'elles sont faciles à interpréter. La méthode est issue de l'apprentissage non supervisé. Elle est basée sur des calculs élémentaires, elle est très coûteuse en temps et elle marche pour des découvertes de faits fréquents, elle peut produire des règles triviales et inutiles.

Les règles d'association ont été, initialement utilisées par Agrawal [5] en analyse de données puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données organisées selon des schémas relationnels de grandes tailles (Simon et Napoli, 1999). Elles ont été, par la suite, appliquées à la fouille de textes (Feldman et

Dagan, 1995 ;Toussaint et al., 2000). Soit une base de données transactionnelle où chaque transaction est une liste d'items (achats par un client lors d'une visite). Beaucoup de mesures de qualité existent dans la littérature notamment en dans la thèse de Feno (2007) à la Réunion- France [15] . La mesure de la qualité est dans notre approche liée à la finesse du choix du « support » et de la « confiance » que nous expliquons plus bas dans le texte. Par support et confiance, on désigne précisément les algorithmes d'extraction qui recherchent de façon exhaustive les règles d'association dont le support et la confiance dépassent des seuils fixés au préalable par l'utilisateur, notés *min supp* et *min conf* .

La recherche de règles d'association intéressantes est un thème privilégié de l'extraction des connaissances à partir des données. Les algorithmes du type Apriori fondés sur le support et la confiance des règles ont apporté une solution élégante au problème de l'extraction de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes. Dans le cadre des données GPS, nous pouvons simplement supprimer les sites peu visités en réglant finement le support et la confiance.

Il faut disposer d'autres mesures venant compléter le support et la confiance. Dans la pratique chaque enregistrement est une transaction alors que les différents champs correspondent aux données susceptibles de composer la transaction.

Dans la mesure où l'on s'intéresse à la présence-absence de chaque article dans les différentes transactions, on associe à chaque article l'acte d'achat correspondant, appelé item, qui est une variable booléenne.

Soit $I = \{i1, i2, \dots, in\}$ l'ensemble des items pouvant faire partie d'une base de données transactionnelle de D.

Soit *T* une transaction ; *T* est donc un sous-ensemble de $I : T \subseteq I$

Une règle d'association est de la forme $X \Rightarrow Y$ ou :

$$X \subset I, Y \subset I et X \cap Y = \emptyset$$

Sur l'ensemble des transactions, on obtient une matrice booléenne de dimensions *n* et *p*. A un ensemble d'articles, on associe la conjonction des actes d'achat correspondant, ou itemset, qui est aussi une variable booléenne.

À partir de la matrice booléenne qui indique les articles présents dans chaque transaction, on veut extraire des règles si un touriste client va sur le site A et sur le site B. Il est probable qu'il aille aussi sur le site C.

Une règle d'association est ainsi une expression *r* du type $A \rightarrow B$, ou l'antécédent A et le conséquent B sont des itemsets qui n'ont pas d'items communs.

Dans la mesure où le nombre de règles d'association possibles croît exponentiellement avec le nombre d'items, il est capital de pouvoir se limiter à l'extraction des règles les plus intéressantes. Il faut pour cela être capable de définir celles-ci et de les identifier, puis il faut les valider. Les algorithmes doivent être affinés en les liants aux critères de support et de confiance. Ils parcourent les itemsets afin de rechercher les itemsets fréquents, donc ceux dont le support dépasse *min supp*, pour en déduire les règles d'association dont la confiance dépasse *min conf*. Dans notre étude cette approche est essentielle car étant donné les volumes de données et leur nature géographique, une représentation qui ne prendrait pas en compte des seuils serait simplement illisible.

Sylvie Guillaume et Pierre-Antoine Papon dans leur papier « Étude comparative d'extraction de règles d'association positives et négatives et optimisations » (2013) [16] expliquent l'algorithme d'extraction de RAPN (Règles d'Association Positives et Négatives) reposant sur l'algorithme fondateur Apriori. Cela nous permet d'appréhender l'importance d'affiner les paramètres support et confiance.

Cette méthode est appropriée car deux types de recherches sont possibles ; d'une part les itemsets fréquents, d'autre part les itemsets non fréquents.

L'algorithme fondateur « Apriori », (Agrawal et Srikant, 1994) procède en deux temps : premièrement on recherche les itemsets fréquents, ceux dont le support dépasse minsupp, en balayant le treillis des itemsets dans sa largeur et en calculant les fréquences par comptage dans la base, ce qui impose une passe sur la base; ensuite pour chaque itemset fréquent X, dont la confiance dépasse le seuil *min conf* .

Au point de départ de l'algorithme, il faut fixer un seuil de support minimal pour que seules les règles d'association avec un support plus grand ou égal à ce seuil soient générées. Il génère tous les sous-ensembles de k items potentiellement fréquents à partir des sous-ensembles des (k-1) items fréquents ; il élague tous les sous-ensembles de k items qui ne peuvent être fréquents. L'algorithme fait un k-ième passage dans la base de données pour calculer le support des sous-ensembles de k items générés et retenus.

Les règles déduites des itemsets fréquents ont nécessairement une confiance supérieure au seuil de support, dans la mesure où *Supp (A ! B) < Conf (A ! B)*. L'efficacité de Apriori diminue en présence de données denses ou fortement corrélées. Toute la difficulté de l'extraction des fréquents consiste à identifier la bordure entre itemsets fréquents et itemsets non-fréquents dans le treillis des itemsets [4]. La recherche peut se faire en largeur ou en profondeur. Dans chaque cas, on peut procéder par comptage direct de la fréquence de chaque itemset dans la base, ou procéder par intersection des deux itemsets qui constituent l'itemset candidat.

// Recherche des « beepways » appartenant aux flottes de $tab_flottes_recommandation

$ s q l = "SELECT DISTINCT ('nom_jeux ') FROM ' jeux_arret ' " ;

D'un point de vue pratique, l'extraction de toutes les règles valides se fait en deux étapes tout d'abord détermination des *itemsets* fréquents c'est-à-dire le support puis on pratique une dérivation des règles d'association valides, c'est-à-dire de confiance.

La première étape est généralement la plus coûteuse, car elle nécessite des accès au contexte d'extraction alors que la deuxième ne nécessite aucun.

Implémentation (Règle) R : Site touristique A -> Site touristique B

Dans ces jeux de données il est possible de connaitre la relation entre deux sites si elle existe. Cette existence est pondérée par un support et une confiance dans le calcul des règles d'associations.

Etape 1 : On fixe les paramètres et le jeu de donné résultant du résumé de données (liste des arrêts géo référencés)

Etape 2 : On calcule la matrice (Lignes = parcours ; Sites = Identifiant ID du site touristique)

Exemple : dans le parcours 8 le touriste a été sur le site

ID :12 (Habitation Clément) Rhumerie

Puis sur le site

ID : 16 (Les Rails de la Canne à Sucre)

Lignes\Sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1															
2		1														
3		1														
4			1													
5	1															
6		1														
7												1				
8												1				1
9																
10																
11			1			1										

Etape 3 : On calcule les supports et la confiance minimale afin d'éliminer les règles peu fréquentes.

Support X U X` :	Total AR X X`	Total AR X` X	AR X X`	AR X` X
1 2 = 3 => Support : 0.023	{1 }=>{2} = 0.333	{2 }=>{1} = 0.231	{1 }=>{2} = 0.333	{2 }=>{1} = 0.231
1 3 = 2 => Support : 0.015	{1 }=>{3} = 0.222	{3 }=>{1} = 0.069	{1 }=>{3} = 0.222	
1 15 = 1 => Support : 0.008	{1 }=>{15} = 0.111	{15 }=>{1} = 0.5		
1 16 = 1 => Support : 0.008	{1 }=>{16} = 0.111	{16 }=>{1} = 0.25		
1 23 = 3 => Support : 0.023	{1 }=>{23} = 0.333	{23 }=>{1} = 0.111	{1 }=>{23} = 0.333	
1 24 = 1 => Support : 0.008	{1 }=>{24} = 0.111	{24 }=>{1} = 0.143		
1 27 = 1 => Support : 0.008	{1 }=>{27} = 0.111	{27 }=>{1} = 0.25		
1 44 = 1 => Support : 0.008	{1 }=>{44} = 0.111	{44 }=>{1} = 0.111		
1 52 = 2 => Support : 0.015	{1 }=>{52} = 0.222	{52 }=>{1} = 0.182	{1 }=>{52} = 0.222	{52 }=>{1} = 0.182
1 57 = 1 => Support : 0.008	{1 }=>{57} = 0.111	{57 }=>{1} = 0.111		
1 80 = 7 => Support : 0.053	{1 }=>{80} = 0.778	{80 }=>{1} = 0.056	{1 }=>{80} = 0.778	
1 124 = 1 => Support : 0.008	{1 }=>{124} = 0.111	{124 }=>{1} = 0.167		
1 129 = 1 => Support : 0.008	{1 }=>{129} = 0.111	{129 }=>{1} = 0.25		
1 134 = 2 => Support : 0.015	{1 }=>{134} = 0.222	{134 }=>{1} = 0.286	{1 }=>{134} = 0.222	{134 }=>{1} = 0.286
1 135 = 1 => Support : 0.008	{1 }=>{135} = 0.111	{135 }=>{1} = 0.167		
1 137 = 1 => Support : 0.008	{1 }=>{137} = 0.111	{137 }=>{1} = 0.083		
1 140 = 1 => Support : 0.008	{1 }=>{140} = 0.111	{140 }=>{1} = 0.333		
1 179 = 1 => Support : 0.008	{1 }=>{179} = 0.111	{179 }=>{1} = 0.5		
1 187 = 2 => Support : 0.015	{1 }=>{187} = 0.222	{187 }=>{1} = 0.133	{1 }=>{187} = 0.222	
1 188 = 2 => Support : 0.015	{1 }=>{188} = 0.222	{188 }=>{1} = 0.4	{1 }=>{188} = 0.222	{188 }=>{1} = 0.4
1 197 = 1 => Support : 0.008	{1 }=>{197} = 0.111	{197 }=>{1} = 0.5		
1 200 = 1 => Support : 0.008	{1 }=>{200} = 0.111	{200 }=>{1} = 0.037		
1 209 = 1 => Support : 0.008	{1 }=>{209} = 0.111	{209 }=>{1} = 0.167		
1 217 = 2 => Support : 0.015	{1 }=>{217} = 0.222	{217 }=>{1} = 0.286	{1 }=>{217} = 0.222	{217 }=>{1} = 0.286
1 218 = 2 => Support : 0.015	{1 }=>{218} = 0.222	{218 }=>{1} = 0.222	{1 }=>{218} = 0.222	{218 }=>{1} = 0.222

Figure 5 : Note : La figure (copie écran) fait apparaître toutes les données

Les Colonnes AR XX '& AR X'X permettent de « voir » les valeurs des règles. Certaines colonnes sont vides car le support ou la confiance est inférieur au seuil donné en paramètre.

3.1.3 Les motifs séquentiels A-> B

Introduits dans Agrawal (1995), la recherche de motifs séquentiels consiste ainsi à extraire des enchaînements d'ensembles d'items, couramment associés sur une période de temps bien définie. Le motif séquentiel met en évidence des associations inter-transactions, contrairement à celle des règles d'association qui extrait des combinaisons intra-transactions.

Les travaux récemment menés (lesquels ?) ont montré que toutes les approches qui visent à extraire l'ensemble des motifs séquentiels deviennent cependant inefficaces dès que le support minimal spécifié par l'utilisateur est trop bas ou lorsque les données sont fortement corrélées. En effet, dans ce cas, et plus encore que pour les itemsets, les recherches sont pénalisées par un espace de recherche trop important.

Pour essayer de gérer au mieux ces problèmes et l'introduction de complexités spatiales et temporelle, deux grandes tendances se distinguent à l'heure actuelle. Dans le premier cas, les propositions comme PrefixSPAN Pei et al. (2004)[51] ou SPADE Zaki (2001) [52] se basent sur de nouvelles structures de données et une génération de candidats efficaces. Les approches de la seconde tendance considèrent l'extraction d'une représentation condensée Mannila et Toivonen (1996) [53].

Depuis plus de vingt ans de nombreux travaux Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015) [17] Jensen, S., & SPSS, U. (2001) [18] Garofalakis, M. N., Rastogi, R., & Shim, K. (1999) [19] en extraction d'information et en fouille de textes appliquées au domaine médical et biomédical ont vu le jour. Deux tâches sont particulièrement explorées correspondant aux deux requêtes mentionnées précédemment : la première est la reconnaissance d'entités nommées de type biologique (noms de gènes, protéines, fonctions biologiques, etc.) et la deuxième concerne l'identification et le typage de relations entre entités biologiques précédemment reconnues.

Ces approches ont fait leurs preuves pour des applications traitant les transactions financières, les suivis de navigations sur le web, les données musicales, la sécurité

informatique etc. Aujourd'hui les firewalls utilisent des approches intégrant de la classification afin d'identifier les tentatives d'intrusion. La lutte contre la fraude à la Carte bancaire utilise le délai entre deux utilisations et la distance et avec la péréquation de celle-ci permet de scorer si la carte bancaire a été dérobée ou non.

A l'inverse de la problématique d'extraction des itemsets, les travaux sur l'extraction des motifs séquentiels, de par leur complexité, sont rares et de nombreux axes de recherche restent encore à étudier (Masseglia, Teisseire et Poncelet, 2004) [20].

L'équipe TATOO - Fouille de données environnementales au sein du LIRMM développe un thème de recherche sur l'extraction des motifs séquentiels et ses applications à des gros volumes de données comme les données médicales (garantir le respect de la vie privée pour les données), les données multidimensionnelles (fouille de cube), les données du web (web sémantique). Le problème de l'extraction de motifs séquentiels peut sembler proche de celui de l'extraction de règles d'association. Ce rapprochement s'avère cependant très fragile en raison d'un élément clé qui est propre à l'extraction de motifs séquentiels : la temporalité. Cette notion permet à la fois de distinguer à l'intérieur des enregistrements un ordre d'apparition mais aussi de regrouper certains éléments.

Si les règles d'association s'appliquent à des données de type itemsets (et permettent l'extraction de règles intra-transaction), la recherche de motifs séquentiels s'applique à des données de type séquences d'itemsets (et permet donc l'extraction de règles inter-transactions). « *Introduits dans [AGR 95b] et largement étudiés dans [MAS 02], les motifs séquentiels peuvent être vus comme une extension de la notion de règles d'association intégrant diverses contraintes temporelles. La recherche de tels motifs consiste ainsi à extraire des enchaînements d'ensembles d'items, couramment associés sur une période de temps bien spécifiée. En fait, cette recherche met en évidence des associations d'inter-transactions, contrairement à celle des règles d'association qui extrait des combinaisons intra-transactions. Par exemple, des motifs séquentiels peuvent montrer que 60 pour cent des gens qui achètent une télévision, achètent un magnétoscope dans les deux ans qui suivent.* » Masseglia, Teisseire et Poncelet (2004, p. 2) [20].

Ce problème, posé à l'origine dans des contextes de marketing, est important dans des domaines divers (détection de fraudes), la finance, ou encore la médecine (identification des symptômes précédant les maladies, données de facturations médicales RSS/RSF). La prise en compte importante de la temporalité dans les enregistrements à étudier permet

une plus grande précision dans les résultats, mais implique aussi un plus grand nombre de calculs et de contraintes.

Définition

Soit D une base de données de transactions de clients où chaque transaction T est composée de :

-Un identifiant du véhicule noté *Cid*

-Le temps utilisé, notée temps

-Un ensemble d'items 'itemset 'intervenants dans la transaction i_t

-Une transaction constitue, pour un touriste C, l'ensemble des items arrêts effectués par C à lors d'un même parcours (de la récupération du véhicule à sa restitution).

Une transaction s'écrit sous la forme d'un ensemble : id-client, id-date, itemset.

Dans notre exemple la fin de séquence est caractérisée par le retour du véhicule au parking du loueur. La rupture de séquence intervient donc à ce niveau.

le parcours réalisé est donc : 52,257,1,23,1,3,1,80 (Parking de remise du véhicule)

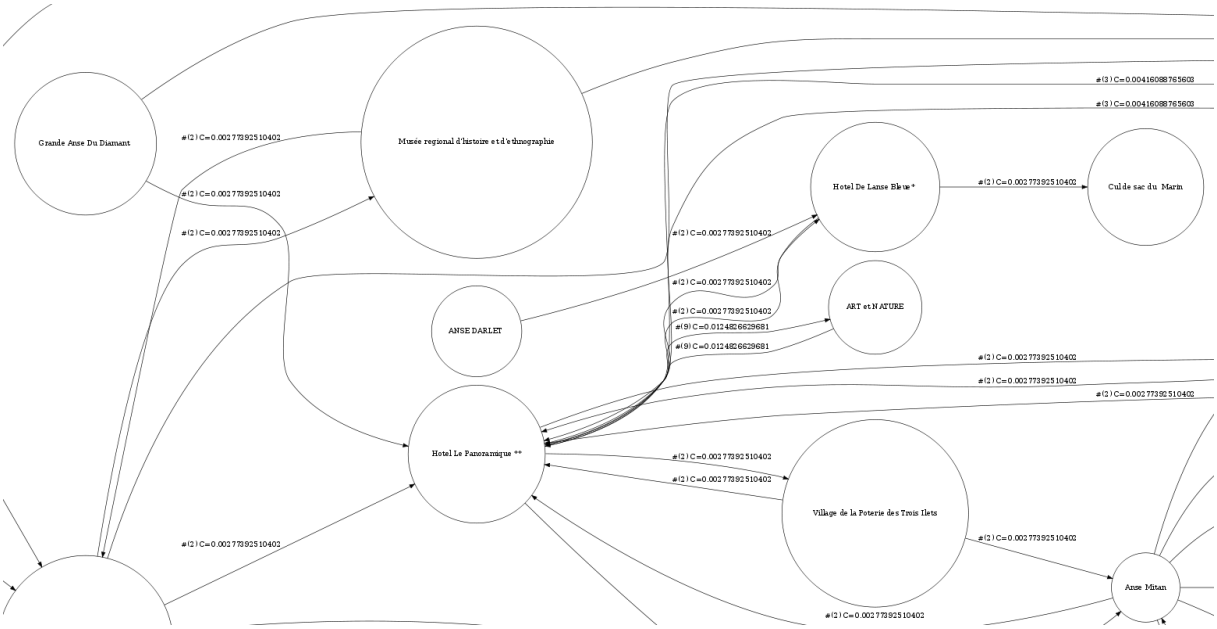


Figure 6 : Representation graphique des règles d'associations

Dans cette séquence on note la relation entre 52 : "ART et Nature" AND 257 : "Hotel le Panoramique".

Nous pouvons proposer deux types de représentations. La carte géographique des Motifs conv(A → B) permet une représentation visuelle des comportements « arrêts touristiques ». Il est alors possible de créer un graphe orienté de relations intersites.

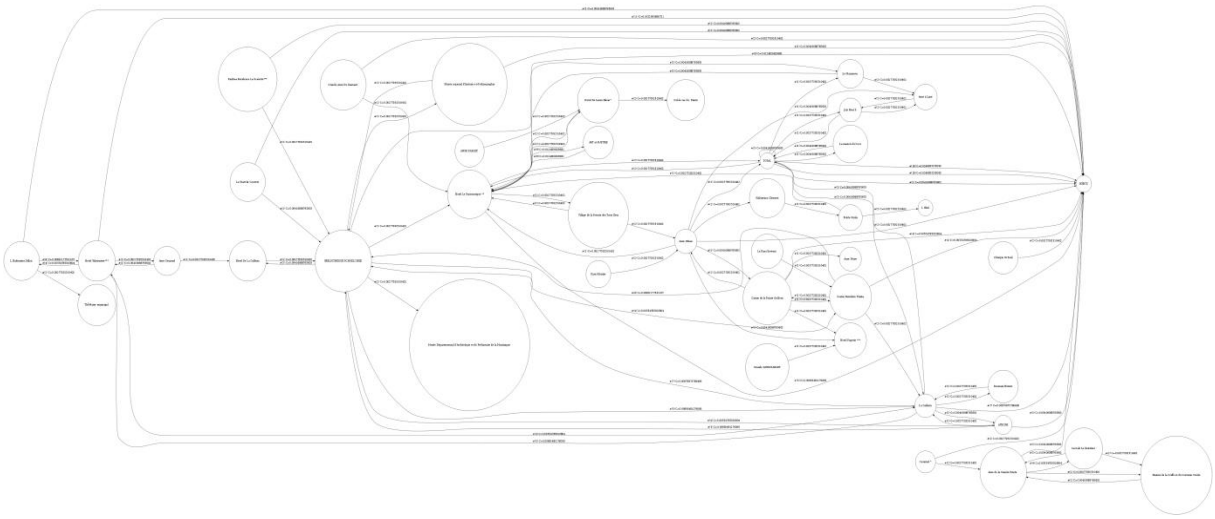


Figure 7:Réprésentaion Ensemble des motifs séquentiels

Nous pouvons explorer l'ensemble des motifs séquentiels permettant de retracer les parcours des touristes de toute l'ile. Ce graphe orienté est la base de travail pour les recommandations.

Nous pouvons proposer deux types de présentations. La carte géographique des motifs permet une représentation visuelle des comportements arrêts touristiques.

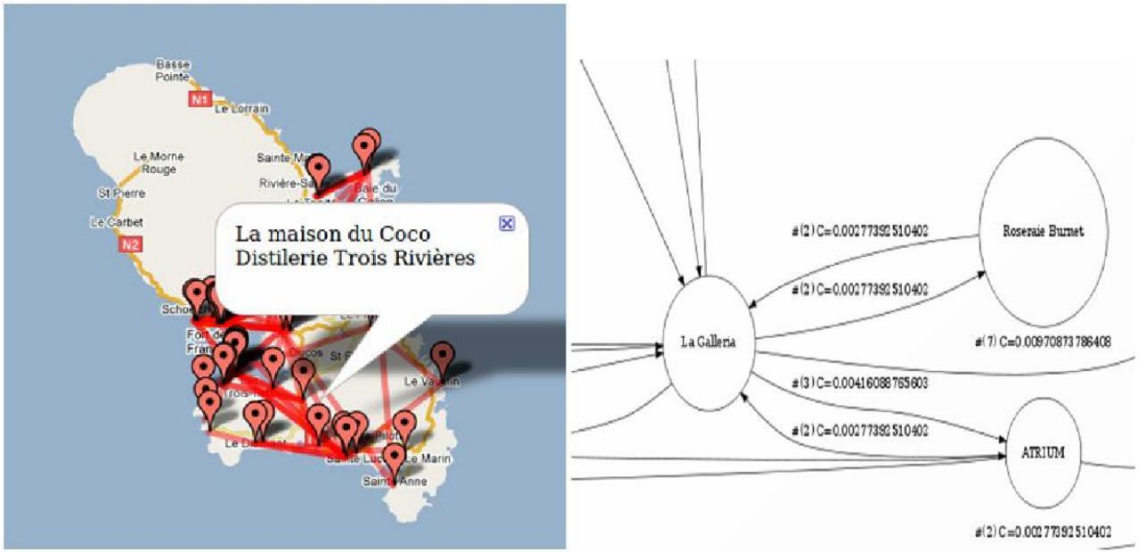


Figure 8: Réprésentation géographique des liaisons inter-sites

Ce dernier graphe permet de mettre en évidence simplement et intuitivement les liaisons entre les différents sites touristiques.

3.1.2 Les barycentres géographiques

Le barycentre est un point qui permet de résumer un ensemble géométrique sur lequel sont réparties des valeurs numériques. Ces valeurs représentent des poids pour déterminer le point d'équilibre d'un élément non axe. Mathématiquement, le barycentre s'obtient en annulant une relation vectorielle. Cette notion généralise la construction du milieu d'un segment ou du centre de gravité d'un triangle, on parle dans ces cas d'isobarycentre.

Dans notre étude l'objectif est de calculer à l'intérieur d'un Q motif. Les valeurs numériques des sommets sont les points GPS du motif séquentiel et les poids pondérant le barycentre et le support de chaque item du motif séquentiel. La finalité est de pouvoir dans le cadre d'une recommandation, ou l'on a réussi à attribuer -en temps réel- un véhicule roulant à un motif séquentiel, de pouvoir lui recommander une activité la plus proche géographiquement en sachant que l'ensemble des autres ont un poids équivalent dans son motif séquentiel.

On arrive alors à affiner la recommandation à l'intérieur d'un motif en intégrant la notion de proximité géographique. Le point vert représentant la recommandation proposée.

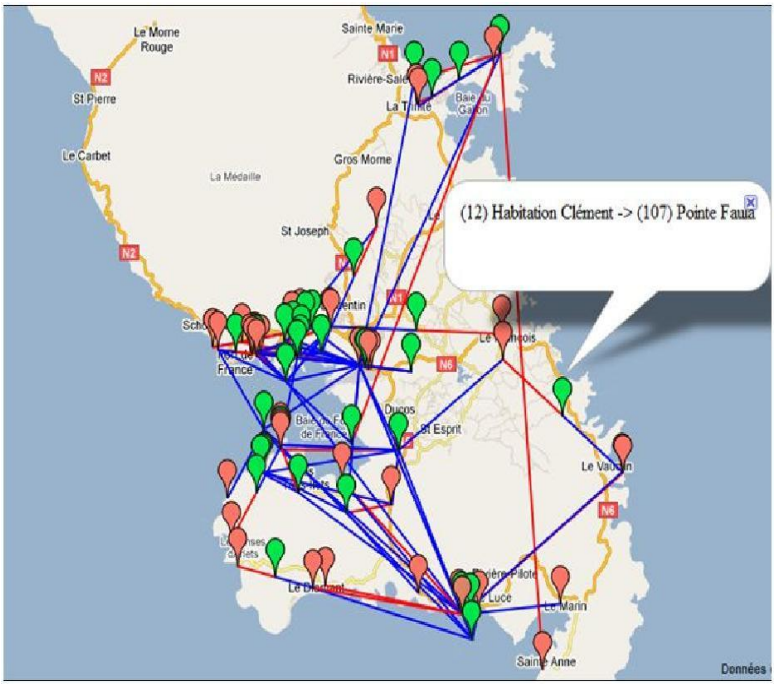


Figure 9: Réprésentation graphique de recommandation à l'intérieur d'un motif

3.1.3 Les k-means pour géotourisme

Les k means est une des techniques de classification non supervisée (clustering) les plus utilisées. Etant donné un entier K, K-means partitionne les données en K groupes, ou "clusters", ou "classes". Ces classes ne se chevauchent pas. Ce résultat est obtenu en positionnant K "centroïdes" dans les zones ou les éléments ou "item" sont les plus présents. Chaque item est alors affecté au centroïde le plus proche "Distance Minimale". Chaque classe contient donc les observations qui sont plus proches de son centroïde que des autres.

Les centroïdes sont positionnés par un traitement qui a chaque itération les "bougent" et les amènent progressivement dans leur position finale stable.

La grande popularité de K-means vient de :

1. Sa simplicité conceptuelle.
2. Sa rapidité et ses faibles exigences en taille mémoire.

Cependant, elle présente également de certains défauts :

- L'utilisateur doit choisir *a priori* la valeur de K, le nombre de classes. Ce choix peut se faire par simple examen visuel dans le cas de données bidimensionnelles, mais il n'en est pas de même pour des données de dimension supérieure. Il n'existe en général pas d'indication claire sur le nombre approprié de classes, et un "mauvais choix" pour la valeur de K conduira alors à une typologie sans rapport avec la réalité.
- Pour une valeur donnée de K, les classes obtenues dépendent beaucoup de la configuration initiale des prototypes, ce qui rend l'interprétation des classes difficiles.

- K-means est une technique objective, ce qui veut dire qu'elle minimise la valeur d'un certain critère numérique. C'est donc une technique d'optimisation. Comme c'est souvent le cas en optimisation, l'algorithme K-means s'arrête lorsqu'il ne peut plus faire baisser la valeur du critère. Cependant, il est tout à fait possible qu'une autre configuration des prototypes conduise à des valeurs encore plus faibles du critère.

Une méthode spécialement adaptée à la segmentation d'images couleurs a été développée par Lucchese (2001) [21], mais elle sera pas utilisée dans notre étude.

Le principe est repris dans notre traitement et prend une forme simple

> formater les donnes ( )

> l ' initialisation ( ) // par défaut initialisé de f aléatoirement

> Analyse des résultats ( )

// Réitération jusqu'a ce que les classes soient stables

> Distance entre les classes ( )

> La recherche des milieux ( )

> Composition des classes ( )

> Moyennes par classe ( )

3.1.4 Le projet industriel – Géotourisme

Le projet vise à équiper des véhicules de location d'un système de géolocalisation GPS, et de collecter le comportement des touristes sur un territoire. Dans un premier temps un message quotidien sera envoyé à chaque véhicule le matin afin de sensibiliser le touriste aux dangers des routes et pour lui souhaiter « Une bonne journée au nom de la collectivité ». Le touriste pourra ensuite recevoir une recommandation issue de l'heure et de son positionnement géographique, et une recommandation issue de la base de

connaissances. Aucune action du touriste ne sera requise afin de préserver le principe non intrusif.

Ce projet de recherche appliquée (entreprise/laboratoire publique) aboutira à la protection de la propriété intellectuelle de deux méthodes (le résumé de données et la recommandation touristique) conformément aux objectifs visés par ce projet. Il permettra aussi un rapprochement les travaux théoriques et le monde industriel (l'entreprise BEEPWAY.COM) afin de valoriser rapidement les travaux de recherches. Beaucoup de sites Web d'e-commerce offrent de nombreux services et sans l'appui du système, filtrant les produits non pertinents, comparant des solutions de rechange, choisir la meilleure option peut être difficile ou impossible pour l'utilisateur connecté au Web par un dispositif mobile. Ces systèmes sont souvent des systèmes de recommandations.

L'objectif d'un système de recommandations est d'aider les utilisateurs à faire leurs choix dans un domaine où ils disposent de peu d'informations pour trier et évaluer les alternatives possibles. L'abondance de contenus sur internet notamment depuis l'essor du contenu généré par les utilisateurs rend plus complexes la navigation et la découverte de contenu pertinent. Il existe certes des moteurs de recherches mais ils ne sont pas adaptés à la tâche. Les nouvelles interfaces permettent d'améliorer les interfaces utilisateurs, de classer le contenu et favorise la décision d'un utilisateur quant à la pertinence d'un contenu.

Des systèmes de recommandations manuels élaborés aident aussi à la découverte de nouveaux produits ou services, dans un cadre strictement supervisé. C'est la caractéristique principale des approches traditionnelles des systèmes de recommandations : les systèmes rassemblent des préférences d'utilisateur en interrogeant explicitement l'utilisateur. Le système exploite les préférences acquises pour activer l'algorithme spécifique de recommandation. Bien que ces préférences tendent à être stables, l'approche a plusieurs inconvénients. D'abord, les utilisateurs doivent avoir assez de connaissances au sujet du domaine pour rendre leurs préférences explicites selon le modèle de produit (par exemple, les attributs du produit).

En second lieu, les préférences incertaines ou inachevées peuvent devenir claires pendant que les utilisateurs agissent avec le système et comprennent mieux ce qu'ils veulent et quels produits sont disponibles, ce qui signifie qu'ils ne peuvent pas être demandés avant que le système ne fournisse quelques recommandations. Troisièmement, beaucoup d'utilisateurs sont peu disposés à indiquer leurs préférences jusqu'à ce qu'ils reçoivent un certain bénéfice du système.

Il existe plusieurs approches dans la littérature : certaines basées sur le contenu et fondées sur l'apprentissage automatique de profils utilisateurs, et d'autres dites de filtrage collaboratif, fondées sur des techniques de fouille de données. Il existe aussi une troisième voie prise notamment par l'Inria et l'équipe AxIS (dans le projet UA/MIDAS), qui poursuit le développement d'une approche hybride de calcul de recommandations basée sur l'analyse du contenu visité et centrée fouille de données où les comportements passés d'un groupe d'utilisateurs sont utilisés pour calculer les recommandations.

CHOIX DE LA QUANTITE DE VEHICULES

Il aurait été possible de créer un échantillon de véhicules, mais cela aurait masqué deux pans importants de l'étude. Si on décide de choisir entre l'échantillonnage probabiliste et l'échantillonnage non probabiliste, ce choix tient à une hypothèse de base au sujet de la nature de la population étudiée. Dans le cas de l'échantillonnage probabiliste, chaque unité (triplet : loueur / voiture / type de touriste) a une chance d'être sélectionnée. Dans un échantillonnage non probabiliste, on suppose que la distribution des caractéristiques à l'intérieur de la population est égale. Dans le projet comment être assuré que l'unité choisie sera une exacte, de plus sans fonction "apriori" l'échantillonnage devient aléatoire. La « randomisation » (le fait de prendre des éléments aléatoirement) est une caractéristique du processus de sélection, plutôt qu'une hypothèse au sujet de la structure de la population. Cela viendrait à choisir aléatoirement des loueurs, de voitures et des touristes les utilisant. Généraliser des comportements touristiques sur des aléas en cascades n'est pas possible. Comment avoir un résultat exact ou proche de l'exactitude si la population utilisée est basée sur des probabilités aléatoires. Dans notre cas spéfifique il n’existe pas encore de référentiel. Après cette étude nous pourrons créer un premier référentiel pour des territoires plus grands ou différents.

Nous aurions pu faire le choix de l'échantillonnage non probabiliste. On choisit arbitrairement des unités, il n'existe aucune façon d'estimer la probabilité pour une unité quelconque d'être incluse dans l'échantillon. Cette méthode en question ne fournit aucunement l'assurance que chaque unité aura une chance d'être incluse dans l'échantillon, on ne peut estimer la variabilité de l'échantillonnage ni identifier le biais possible. Aucune mesure de fiabilité de l'échantillon dans notre domaine de recherche n'est possible. Dans des domaines ayant une temporalité passée et un ordonnancement faible (ex : assurance, produit nanciers) il est utilisé et peut donner de bons résultats. Dans le cadre de prédictions/recommandations sur des comparaisons passées et potentielles futures, un échantillon non probabiliste ne permettrait pas non plus de fournir l'assurance que les estimations ne dépasseront pas un niveau acceptable d'erreur. Les statisticiens hésitent à utiliser les méthodes d'échantillonnage non probabiliste, parce qu'il n'existe aucun moyen de mesurer la précision des échantillons en découlant.

Nous aurions pu tester un type d'échantillonnage parmi les plus courants comme l'échantillonnage de commodité ; volontaire ; au jugé mais peut être que seul l'échantillonnage par quotas aurait donné un résultat. Mais cette méthode nécessite une population de base. Or le champ de l'étude est nouveau et aucune population n'existe pour mettre en place ce type d'échantillon. Notre étude vise à créer un premier échantillon pour ce type de recherche, plusieurs laboratoires de France sont utilisateurs de nos échantillons notamment l'INRIA Sophia (Institut national de recherche en informatique et automatique) dans le cadre du projet MIDAS. MIDAS (Mining Data Streams) a été reconnu par l'ANR au titre de son programme non thématique MDCO. L'EURL RD-GEO, a fait l'objet de travaux de recherche dans le cadre d'un projet reconnu par l'ANR (Agence Nationale de la Recherche), au titre de son programme "Masses de Données et Connaissances" (projet MIDAS, ANR-07-MDCO-008). Ce projet ayant pour responsable local Pr Richard Nock [25] [26], porte sur l'étude d'algorithmes de fouille de données dans les flots de données.

Deux publications ont pu être validées sur cette thématique. D'une part Elisabeth, Nock, et Célimene, (2013) « Demonstrator of a tourist recommendation system » [22] et Elisabeth « Fouille de données spatio-temporelle : Application à un système de modélisation des déplacements touristiques » (2015) [23].

Nous avons la chance en Martinique d'être dans un territoire fermé ou la population est petite et non volatile (les véhicules ne sortent *a priori* pas du champ de l'étude i.e. : l'ile). Cela permet alors un domaine de recherche « non ouvert » et exhaustif même si le territoire est petit. Ce sont des conditions idéales pour notre étude. Il est important de transformer nos spécificités en chances pour la recherche. Le projet à cheval sur deux iles (Martinique et Guadeloupe) permettra aussi de créer un modèle de comparaison utilisable pour d'autres territoires.

PRINCIPE DE COMMUNICATION GPRS

Le système GPS préconisé dans le projet Géotourisme est un système embarqué automobile, alimenté par la batterie du véhicule. Le boîtier est composé d'une antenne GPS réceptrice des positions, en position horizontale pour avoir une vision vers le ciel, d'une antenne GSM /GPRS et un transmetteur GSM 900/1800Mhz compatible avec le réseau de l'opérateur GSM. Le boîtier est aussi composé d'un calculateur, d'une mémoire et d'un écran servant afficher les recommandations au touriste. Le boîtier permet de positionner le véhicule grâce à la réception des signaux GPS et de transmettre cette position via le réseau GSM/GPRS. Le mode d'emploi précise que le boîtier est déjà pré-équipé d'une puce GSM. Sans celle-ci le boîtier ne sert à rien. Le système contient une mémoire interne de quelques kilos octets qui permet de stocker environ 8100 positions dans le cas où le réseau GSM de l'opérateur serait indisponible dans la zone. Cette mémoire est déchargée une fois que le réseau GSM/G-PRS de l'opérateur redevient disponible.

Dans le cas où nous devrions nous passer du réseau GSM avec un autre boîtier (mais aujourd'hui tous les boîtiers de géolocalisation fonctionnent sur ce principe, il faudrait télécharger par un moyen physique (par exemple en branchant une clef USB manuellement), chaque véhicule toutes les 34 heures d'utilisation. Ceci est évidement matériellement impossible c'est la raison pour laquelle les réseaux datas GSM/GPRS sont utilisés pour le transport des données. Le cout, la souplesse (aucune intervention humaine), la rapidité (information de position toutes les 15 secondes) sont sans égales par rapport à un procédé manuel.

Par ce même réseau il est possible de connaître l'état du boîtier et planifier les interventions si nécessaires donc de minimiser le temps passé au diagnostic et à la prévention des pannes.

Dans le projet Géotourisme il est prévu que l'affordance perceptible (telle que définit par James Jerome Gibson, Référence), donc l'intelligence du système à suggérer sa propre utilisation soit très forte. Il faut donc interagir et c'est d'ailleurs la nature d'une IHM (interface homme machine)- avec l'utilisateur de façon non supervisée et en temps réel. Sans réseau GSM donc transfert de données, le boîtier ne fonctionnerais pas, mais il serait impossible a moins de se rendre physiquement en n de parcours du touriste et de recommander au touriste une activité pour une prochaine période. Ceci qui est logistiquement irréalisable et demanderait des interventions humaines pour chaque touriste. Le projet est basé sur les technologies d'aujourd'hui économes en moyen humain. La recommandation doit être fondamentalement en temps réel sinon elle perd sa nature. Par analogie lorsque l'on achète un ouvrage sur un site internet et que celui-ci propose un ouvrage analogue (principe de recommandation), la recommandation n'arrive pas par voie postale. Elle est en temps réel et dans la continuité de l'interaction entre la machine (site web du commerçant). Les systèmes de recommandations sont aujourd'hui orientés vers le e-marketing et le eCRM. Mais on doit imaginer les extensions des modèles et un élargissement des utilisations dans la santé, la recherche d'information, analyse des usages, la personnalisation et le tourisme.

La principale problématique dans le domaine de la conception de systèmes de recommandations est de produire des recommandations de qualité tout en minimisant si possible l'effort requis de la part des producteurs et des consommateurs. Ce problème est particulièrement prégnant dans l'utilisation de dispositifs mobiles.

Pour le surmonter, les chercheurs (qui ?) ont proposé des méthodologies pour dériver des préférences d'utilisateur en analysant le comportement de navigation de l'utilisateur sur un dispositif mobile. Bien que ces approches exigent considérablement moins d'effort de l'utilisateur, elles doivent interpréter l'action de l'utilisateur (par exemple, cliquant sur un hyperlien) et la traduire en préférences. Les préférences implicitement indiquées tendent malheureusement à être imprécises et bruitées.

Une autre approche qui a récemment suscité beaucoup d'intérêt consiste en un dialogue structuré et cyclique entre l'humain et l'ordinateur comme l'explique la méthode Diane+ (conçue initialement pour la spécification du dialogue homme-machine) Tarby (1993) [54]. Dans les systèmes conversationnels de recommandation, à chaque cycle d'interaction, le système peut par exemple demander à l'utilisateur une préférence, ou lui proposer un produit. L'utilisateur répond à la question du système ou critique sa proposition. Ainsi, le système peut poser une question sélective pour affiner des préférences de l'utilisateur quand le système n'a trouvé aucun article (ou trop d'articles). En répondant, l'utilisateur peut indiquer, enlever, ou modifier quelques conditions de sorte que le système puisse rechercher un meilleur ensemble de résultat. Ces applications fonctionnent néanmoins la plupart du temps sur PDAs, et pas sur téléphones portables. C'est la première spécificité de notre cadre d'étude.

La plupart des approches fondées sur la fouille de données sont principalement des approches statistiques où l'ordre d'occurrence d'événements dans l'historique n'est pas pris en compte lors du calcul de recommandation, or dans la modélisation des parcours et des durées cet ordre est une composante très importante, et une autre limitation dans notre cadre d'étude.

Conclusion

Le projet Géo-tourisme prend comme base un service de géo localisation. Ce service vise à assurer le suivi GPS des flottes.

- de véhicules de transport pour la maîtrise de la chaine du froid
- de pompes à béton dans le but de maîtriser l'apport d'eau pour la norme NF
- de véhicules de commerciaux
- de yoles sur l'eau pour un suivi en temps réel sur internet
- des taxi-co (véhicule à 9 places) pour la régulation du transport urbain
- d'ambulances

Dans le domaine de la conception de systèmes de recommandations, l'effort du traitement informatique requis est minimisé par l'introduction des résumés de données. Notre

approche du résumé prend la forme de courtes séquences informatiques qui « résument » le parcours du touriste.

Ce volet applicatif a permis de mettre en lumière plusieurs approches de traitement des données. Il est l'élément de base qui permettra de proposer des recommandations aux touristes en temps réel en fonction des associations des sites « précédemment » visités, avec les sites dans lesquels son parcours s'inscrit dans le cadre du motif séquentiel.

Cette approche possède aussi plusieurs extensions, notamment la « clustérisation » grâce aux algorithmes k-means qui permet de créer des groupes de comportements. Ces approches couplées à une vision en temps réel selon la position et la direction du véhicule de location permet de proposer une approche géographique de l'analyse des comportements et par extension de la recommandation aux touristes.

Dans la recommandation nous prenons en compte de façon très fine la notion de positionnement géographique, mais aussi la donnée temporelle. Nous apportons une solution à la question de la modélisation des déplacements touristiques en intégrant le lieu et le temps.

4 MCO et fouille de données

Les établissements de santé publics et privés disposent de volumes importants d'informations quantifiées et standardisées sur leur activité médicale, au travers du Programme de Médicalisation des Systèmes d'information (PMSI) qui s'inscrit dans la réforme hospitalière avec comme ambition l'optimisation de l'organisation de l'offre de soins et la réduction des inégalités de ressources entre les établissements de santé. Ces réformes se sont accompagnées de l'instauration de la T2A, qui vise à une répartition plus équitable des moyens financiers entre établissements assurant des missions de service public.

4.1 Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie : l'apport de la fouille de données

Le PMSI est apparu à l'initiative du directeur des Hôpitaux de l'époque, Jean de Kervasdoué, en 1982, qui s'est inspiré du système de maîtrise de dépenses de santé élaboré aux États-Unis par le professeur Fetter. En offrant une description précise de l'activité médicale de chaque établissement de santé, le PMSI permettre de mesurer les coûts des différentes pathologies en groupant les malades par famille de pathologies (GHM). Le GHM est permis par un classement des malades au moyen des résumés standards de sortie (RSS). En donnant la possibilité de comparer l'activité et les coûts respectifs des établissements, le PMSI permet de réduire les inégalités de dotation budgétaire ainsi mises en évidence, car sur la base des séjours classés par GHM, il est calculé un indice synthétique d'activité qui, valorisée, permet de calculer les dotations budgétaires des établissements de santé.

4.1.1 Visualisation des données en 3D interactives

Pour la visualisation des données en 3D interactive, nos travaux s'appuient sur le couplage de deux outils, Mathematica et Wolfram SystemModeler.

Mathematica est un puissant moteur de calcul symbolique et numérique. Il est doté de capacités graphiques et de visualisations exceptionnelles, un langage robuste de

programmation et de développement, un environnement intuitif pour l'édition scientifique et technique. Mathematica est le logiciel de calcul de référence dans le monde de l'enseignement et de la recherche.

Wolfram SystemModeler offre une nouvelle approche à la modélisation et la simulation de systèmes complexes. Le couplage avec Mathematica permet de bénéficier d'une suite logicielle fonctionnelle pour notre étude.

Le premier module développé permet pour une année de visualiser et de comparer pour deux établissements les durées de séjours (*Figure 10*). Ce module paramétrable permet de sélectionner un AGE, une durée de séjours MAXIMUM et afficher pour DEUX établissements les durées de séjour moyennes : les diagnostics principaux sont cumulés

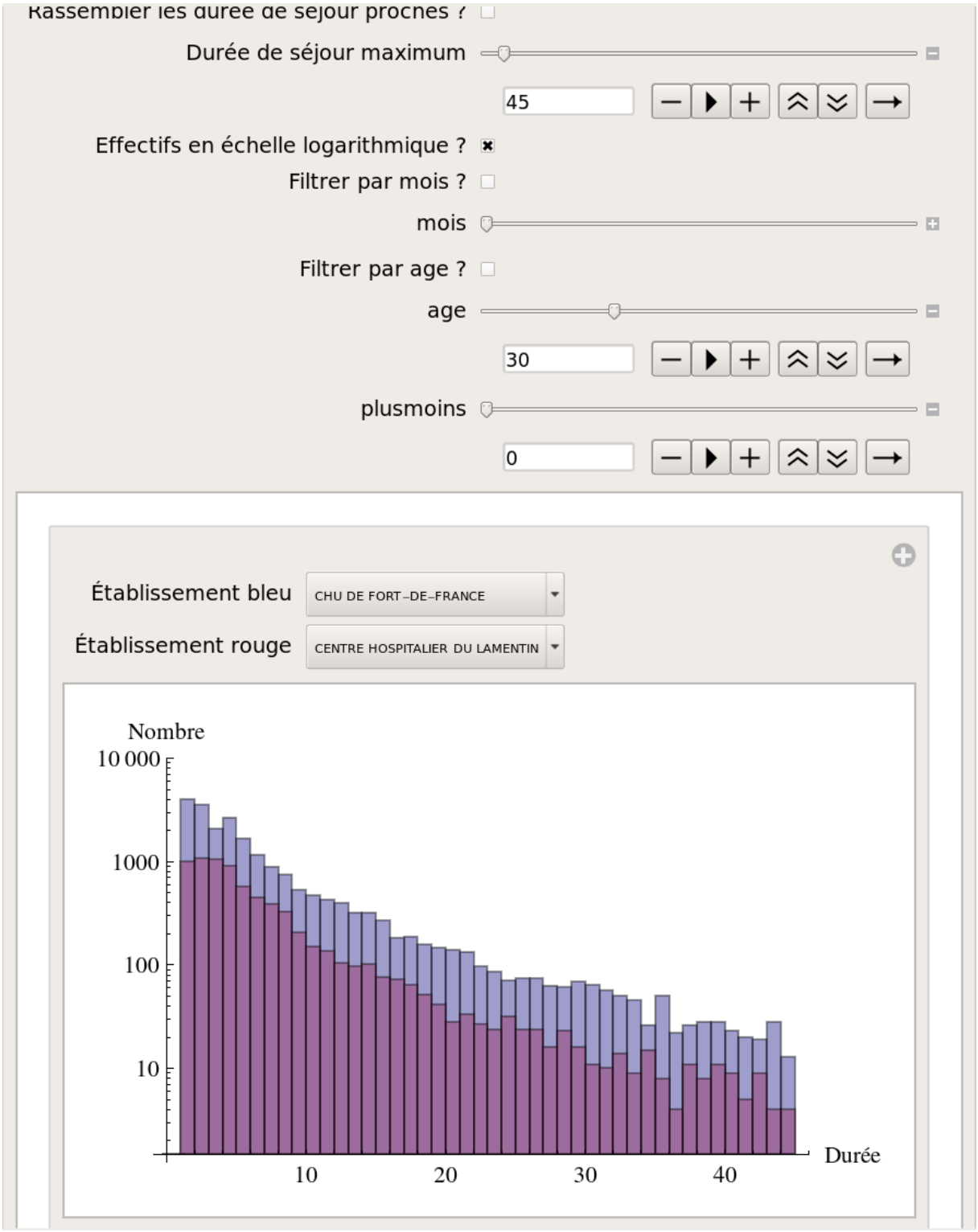


Figure 10 – Comparaison des durées de séjours entre le Centre Hospitalier du Lamentin et le CHU de Fort-de-France en 2007

Le deuxième module développé permet de visualiser avec une granularité mensuelle les durées de séjours pour un établissement et pour une année donnée (*Figure 11*). Il est possible de ne prendre en compte que des durées maximales.

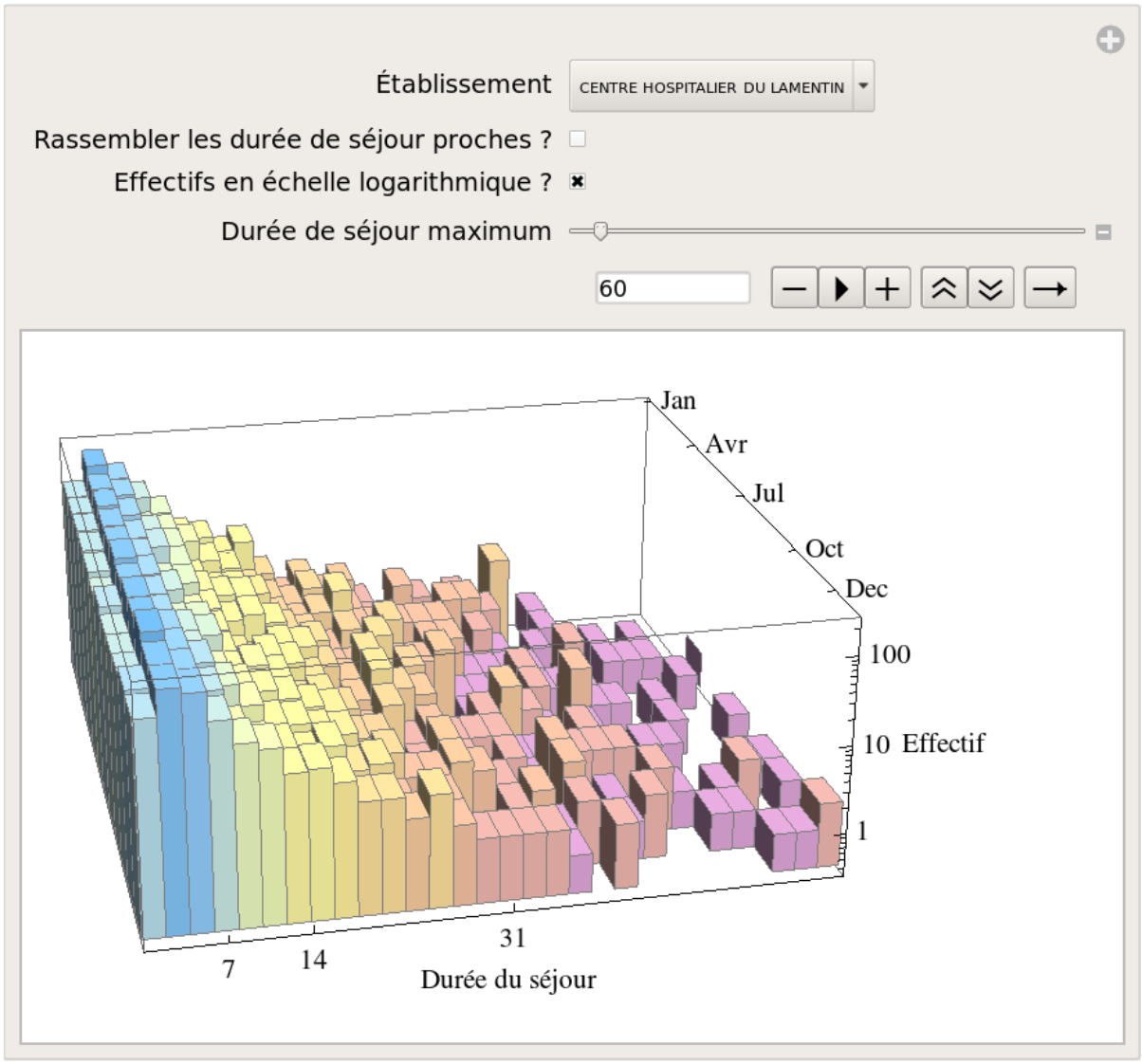


Figure 11 – Liste des durées de séjour du Centre hospitalier du Lamentin en 2007

Le troisième module permet de comparer côte à côte les durées de séjour de deux établissements (*Figure 12*). Les comparaisons visuelles sont facilitées. Il est possible d'appliquer un filtre pour les durées de séjours maximales, pour un mois donné et pour un âge donné. Le module fonctionne pour une année donnée. Les échelles sont différentes,

l'utilisateur doit donc être vigilant lors de l'analyse des résultats. Le graphique est interactif et un passage de la souris sur un histogramme affiche l'effectif.

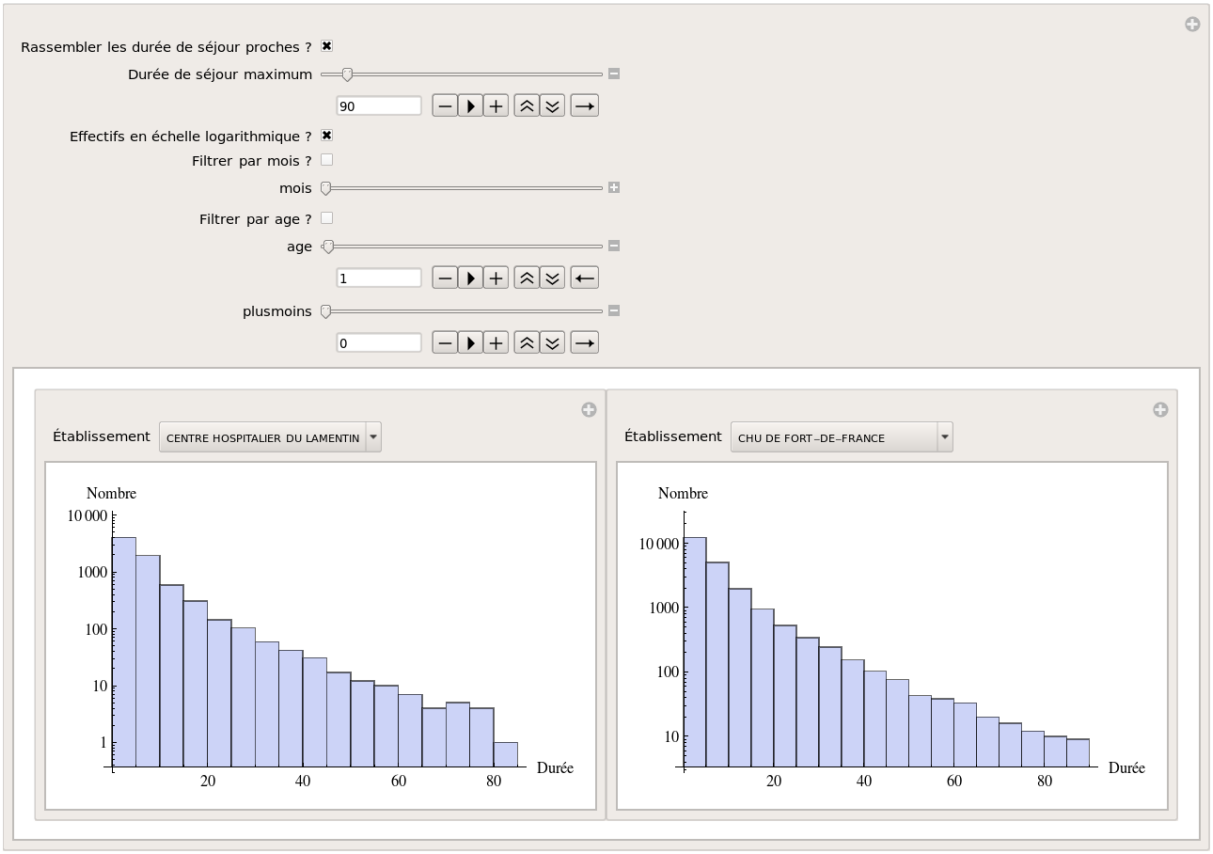


Figure 12 – Comparaison des durées de séjour entre le Centre hospitalier du Lamentin et le CHU de Fort-de-France en 2007

4.1.2 Analyse des Diagnostics principaux et des diagnostics associés

L'outil suivant permet après avoir sélectionné une année puis un établissement de lister, d'afficher sous forme graphique et d'exporter les diagnostics principaux et les diagnostics associés en fonction de leur effectif (*Figure 13a et 13b*).

Dans un souci d'optimisation de l'utilisabilité de l'outil, celui-ci offre la possibilité à l'utilisateur, s'il le souhaite, d'exporter toutes ces données au format Excel (personnalisation des analyses et des présentations...).

Analyse DP/DAS(s) 2008

CENTRE HOSPITALIER DU LAMENTIN - Martinique

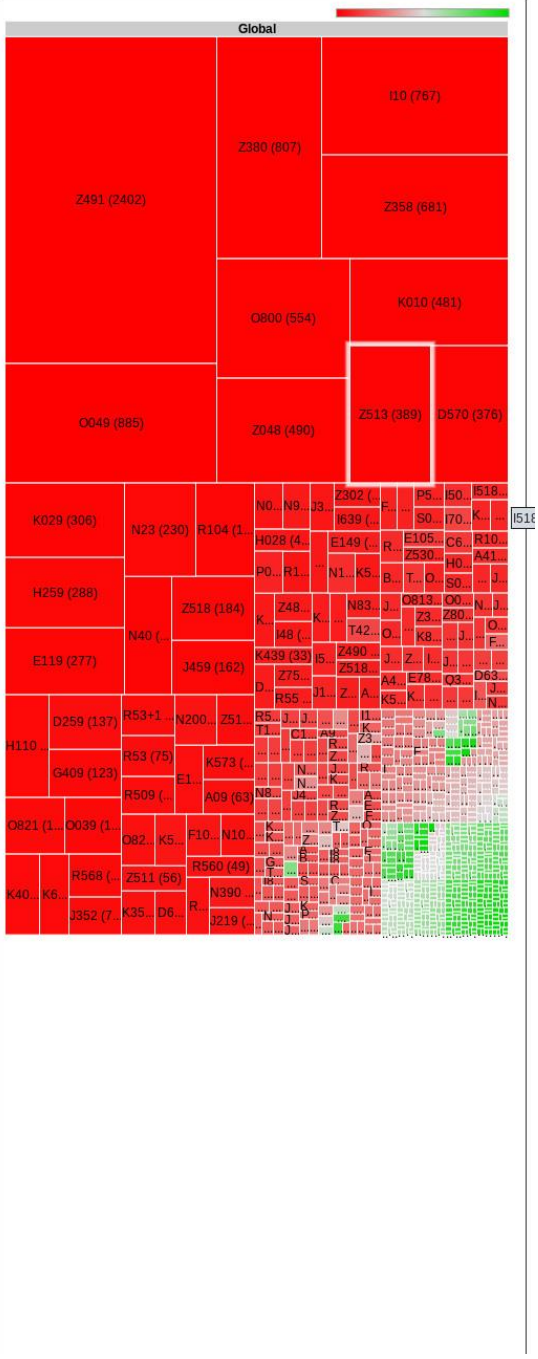
Diagnostic principal	Diagnostic principal (Nombre d'occurences)
Z491 Dialyse extra-corporelle O049 Avortement médical complet ou sans précision, sans complication Z380 Enfant unique, né à l'hôpital I10 Hypertension essentielle (primitive) Z358 Surveillance d'autres grossesses à haut risque O800 Accouchement (unique) spontané par présentation du sommet Z048 Examen et mise en observation pour d'autres raisons précisées K010 Dents incluses Z513 Transfusion sanguine (sans mention de diagnostic) K029 Carie dentaire, sans précision D570 Anémie à hématies falciformes [anémie drépanocytaire] avec crises N40 Hyperplasie de la prostate E119 Diabète sucré non insulino-dépendant, sans complication R104 Douleurs abdominales, autres et non précisées H259 Cataracte sénile, sans précision J459 Asthme, sans précision H110 Ptérygion Z518 Autres formes précisées de soins médicaux N23 Colique néphrétique, sans précision G409 Épilepsie, sans précision O039 Avortement spontané complet ou sans précision, sans complication O821 Accouchement (unique) par césarienne d'urgence K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène) K625 Hémorragie de l'anus et du rectum R568 Convulsions, autres et non précisées J352 Hypertrophie des végétations adénoïdes R53+1 Malaise R53 Malaise et fatigue R509 Fièvre, sans précision Z515 Soins palliatifs D259 Léiomyome de l'utérus, sans précision E109 Diabète sucré insulino-dépendant, sans complication K573 Diverticulose du côlon, (sans perforation ni abcès) A09 Autres gastroentérites et colites d'origine infectieuse ou non précisée O820 Accouchement (unique) par césarienne programmée D649 Anémie, sans précision K359 Appendicite aiguë, sans précision N10 Néphrite tubulo-interstitielle aiguë R101 Douleur localisée à la partie supérieure de l'abdomen Z511 Séance de chimiothérapie pour tumeur R560 Convulsions fébriles J219 Bronchiolite (aiguë), sans précision N049 Syndrome néphrotique, sans précision N200 Calcul du rein F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance R11 Nausées et vomissements J329 Sinusite (chronique), sans précision K566 Occlusions intestinales, autres et sans précision Z302 Stérilisation H028 Autres affections précisées des paupières I639 Infarctus cérébral, sans précision D069 Carcinome in situ du col de l'utérus, sans précision K590 Constipation K429 Hernie ombilicale, (sans occlusion ni gangrène) I48 Fibrillation et flutter auriculaires N390 Infection des voies urinaires, siège non précisé N948 Autres affections précisées des organes génitaux de la femme et du cycle menstruel Z488 Autres soins de contrôle chirurgicaux précisés K439 Autres hernies de la paroi abdominale antérieure, (sans occlusion ni gangrène) D509 Anémie par carence en fer, sans précision Z7588 Autres difficultés liées aux installations médicales et autres soins de santé, non classées ailleurs K859 Pancréatite aiguë, sans précision N47 Hypertrophie du prépuce, phimosis et paraphimosis R55 Syncope et collapsus (sauf choc)	

Figure 13a – Diagnostics principaux du Centre Hospitalier du Lamentin en 2008

Analyse DP/DAS(s) 2008

CHU DE FORT-DE-FRANCE - Martinique

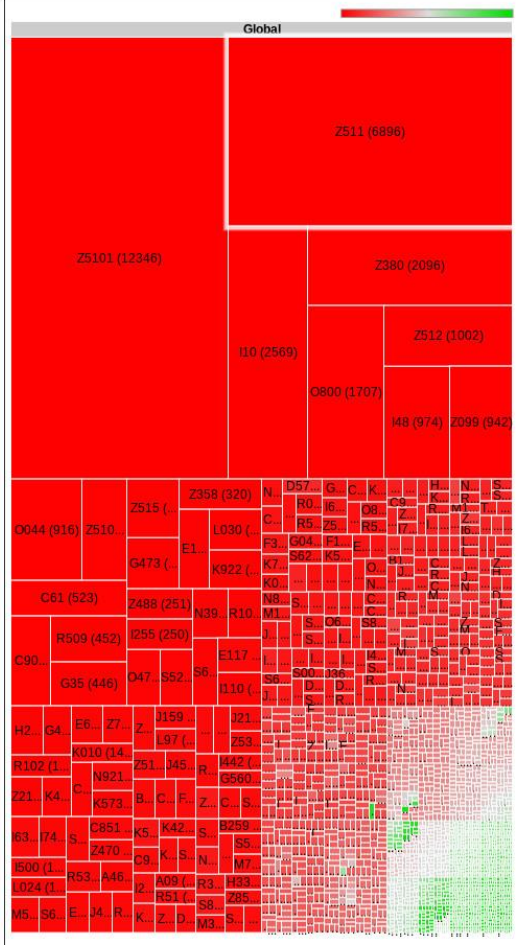
Diagnostic principal	Diagnostic principal (Nombre d'occurences)
Z5101 Séance d'irradiation	
Z511 Séance de chimiothérapie pour tumeur	
I10 Hypertension essentielle (primitive)	
Z380 Enfant unique, né à l'hôpital	
O800 Accouchement (unique) spontané par présentation du sommet	
I48 Fibrillation et flutter auriculaires	
Z512 Autres formes de chimiothérapie	
Z099 Examen de contrôle après traitement pour une affection, sans précision	
O044 Avortement médical incomplet, sans complication	
Z5100 Séance de préparation à une irradiation	
C61 Tumeur maligne de la prostate	
R509 Fièvre, sans précision	
C900 Myélome multiple	
G473 Apnée du sommeil	
Z358 Surveillance d'autres grossesses à haut risque	
G35 Sclérose en plaques	
Z515 Soins palliatifs	
L030 Phlegmon des doigts et des orteils	
K922 Hémorragie gastro-intestinale, sans précision	
E107 Diabète sucré insulino-dépendant, avec complications multiples	
Z488 Autres soins de contrôle chirurgicaux précisés	
O470 Faux travail avant 37 semaines entières de gestation [Menace d'accouchement prématuré]	
S5250 Fracture fermée de l'extrémité inférieure du radius	
R104 Douleurs abdominales, autres et non précisées	
S610 Plaie ouverte de(s) doigt(s) (sans lésion de l'ongle)	
I255 Myocardiopathie ischémique	
I110 Cardiopathie hypertensive, avec insuffisance cardiaque (congestive)	
N390 Infection des voies urinaires, siège non précisé	
H259 Cataracte sénile, sans précision	
G409 Épilepsie, sans précision	
R102 Douleur pelvienne et périnéale	
Z21 Infection asymptomatique par le virus de l'immunodéficience humaine [VIH]	
K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène)	
E660 Obésité due à un excès calorique	
Z751 Sujet attendant d'être admis ailleurs, dans un établissement adéquat	
C504 Tumeur maligne du quadrant supéro-externe du sein	
K573 Diverticulose du côlon, (sans perforation ni abcès)	
L024 Abcès cutané, furoncle et anthrax d'un membre	
M511 Atteintes d'un disque lombaire et d'autres disques intervertébraux avec radiculopathie	
S663 Lésion traumatique du muscle et du tendon extenseurs d'un autre doigt au niveau du poignet et de la main	
I743 Embolie et thrombose des artères des membres inférieurs	
S611 Plaie ouverte de(s) doigt(s) avec lésion de l'ongle	
Z470 Soins de contrôle impliquant l'enlèvement d'une plaque et autre prothèse interne (orthopédique) de fixation	
A46 Érysipèle	
E105 Diabète sucré insulino-dépendant, avec complications vasculaires périphériques	
I500 Insuffisance cardiaque congestive	
E117 Diabète sucré non insulino-dépendant, avec complications multiples	
J450 Asthme à prédominance allergique	
R100 Syndrome abdominal aigu	
Z018 Autres examens spéciaux précisés	
C851 Lymphomes à cellules B, sans précision	
J159 Pneumopathie bactérienne, sans précision	
R53+1 Malaise	
L97 Ulcère du membre inférieur, non classé ailleurs	
K010 Dents incluses	
J459 Asthme, sans précision	
B182 Hépatite virale chronique C	
C509 Tumeur maligne du sein, sans précision	
F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance	
I633 Infarctus cérébral dû à une thrombose des artères cérébrales	
D500 Anémie par carence en fer secondaire à une perte de sang (chronique)	
K802 Calcul de la vésicule biliaire (sans cholécystite)	

Figure 13b – Diagnostics principaux du CHU de Fort-de-France en 2008

Il est possible d'afficher et d'exporter les diagnostics associés en cliquant sur les diagnostics principaux. Ainsi dans notre exemple il est possible de quantifier les naissances pour l'année 2008 (*Figure 14*).

Analyse DP/DAS(s) 2008

CENTRE HOSPITALIER DU LAMENTIN - Martinique

Diagnostic principal	Diagnostic associé (Nombre d'occurences)
Z491 Dialyse extra-corporelle	Z370 Naissance unique, enfant vivant
O049 Avortement médical complet ou sans précision, sans complication	(510 occurence(s))
Z380 Enfant unique, né à l'hôpital	
I10 Hypertension essentielle (primitive)	Z370 Naissance unique, enfant vivant
Z358 Surveillance d'autres grossesses à haut risque	O601 Travail prématuré spontané avec accouchement prématuré
O800 Accouchement (unique) spontané par présentation du sommet	(40 occurence(s))
Z048 Examen et mise en observation pour d'autres raisons précisées	
K010 Dents incluses	Z371 Naissance unique, enfant mort-né
Z513 Transfusion sanguine (sans mention de diagnostic)	O601 Travail prématuré spontané avec accouchement prématuré
K029 Carie dentaire, sans précision	(2 occurence(s))
D570 Anémie à hématies falciformes [anémie drépanocytaire] avec crises	
N40 Hyperplasie de la prostate	Z370 Naissance unique, enfant vivant
E119 Diabète sucré non insulino-dépendant, sans complication	O692 Travail et accouchement compliqués d'une autre forme d'enchevêtrement du cordon, avec compression
R104 Douleurs abdominales, autres et non précisées	(1 occurence(s))
H259 Cataracte sénile, sans précision	
J459 Asthme, sans précision	
H110 Ptérygion	Z374 Naissance gémellaire, jumeaux morts-nés
Z518 Autres formes précisées de soins médicaux	O601 Travail prématuré spontané avec accouchement prématuré
N23 Colique néphrétique, sans précision	(1 occurence(s))
G409 Épilepsie, sans précision	
O039 Avortement spontané complet ou sans précision, sans complication	
O821 Accouchement (unique) par césarienne d'urgence	
K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène)	
K625 Hémorragie de l'anus et du rectum	
R568 Convulsions, autres et non précisées	
J352 Hypertrophie des végétations adénoïdes	
R53+1 Malaise	
R53 Malaise et fatigue	
R509 Fièvre, sans précision	
Z515 Soins palliatifs	
D259 Léiomyome de l'utérus, sans précision	
E109 Diabète sucré insulino-dépendant, sans complication	
K573 Diverticulose du côlon, (sans perforation ni abcès)	
A09 Autres gastroentérites et colites d'origine infectieuse ou non précisée	
O820 Accouchement (unique) par césarienne programmée	
D649 Anémie, sans précision	
K359 Appendicite aiguë, sans précision	
N10 Néphrite tubulo-interstitielle aiguë	
R101 Douleur localisée à la partie supérieure de l'abdomen	
Z511 Séance de chimiothérapie pour tumeur	
R560 Convulsions fébriles	
J219 Bronchiolite (aiguë), sans précision	
N049 Syndrome néphrotique, sans précision	
N200 Calcul du rein	
F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance	
R11 Nausées et vomissements	
J329 Sinusite (chronique), sans précision	
K566 Occlusions intestinales, autres et sans précision	
Z302 Stérilisation	
H028 Autres affections précisées des paupières	
I639 Infarctus cérébral, sans précision	
D069 Carcinome in situ du col de l'utérus, sans précision	
K590 Constipation	
K429 Hernie ombilicale, (sans occlusion ni gangrène)	
I48 Fibrillation et flutter auriculaires	
N390 Infection des voies urinaires, siège non précisé	
N948 Autres affections précisées des organes génitaux de la femme et du cycle menstruel	
Z488 Autres soins de contrôle chirurgicaux précisés	
K439 Autres hernies de la paroi abdominale antérieure, (sans occlusion ni gangrène)	
D509 Anémie par carence en fer, sans précision	
Z7588 Autres difficultés liées aux installations médicales et autres soins de santé, non classées ailleurs	
K859 Pancréatite aiguë, sans précision	

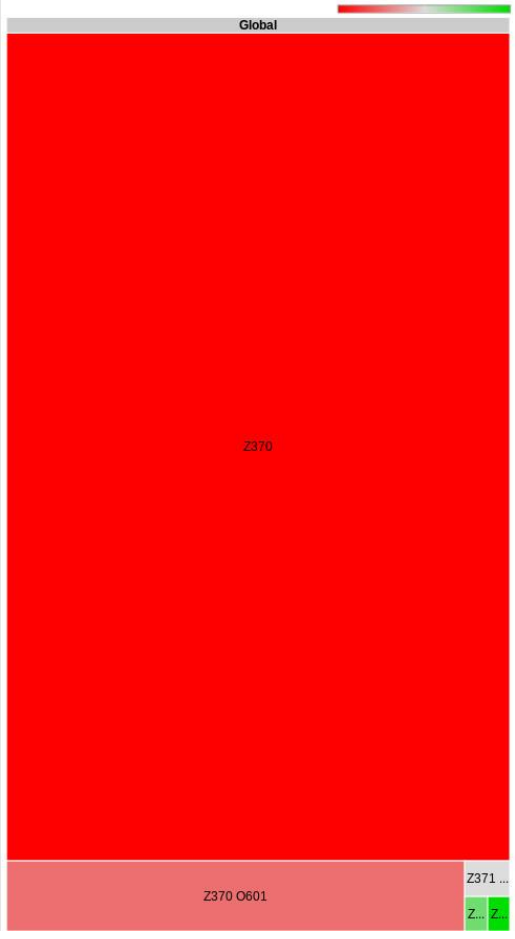


Figure 14 – Diagnostics associés pour le Centre Hospitalier du Lamentin en 2008 pour le diagnostic principal « Accouchements »

Cette même analyse est faite pour le CHU de Fort-de-France pour l'année 2008 (*Figure 15*).

L'outil permet donc des comparaisons entre établissements sur des prises en charge similaires.

Analyse DP/DAS(s) 2008

CHU DE FORT-DE-FRANCE - Martinique

Diagnostic principal	Diagnostic associé (Nombre d'occurences)
Z5101 Séance d'irradiation	Z370 Naissance unique, enfant vivant
Z511 Séance de chimiothérapie pour tumeur	(1512 occurence(s))
I10 Hypertension essentielle (primitive)	
Z380 Enfant unique, né à l'hôpital	Pas de diagnostique associé
O800 Accouchement (unique) spontané par présentation du sommet	(72 occurence(s))
I48 Fibrillation et flutter auriculaires	
Z512 Autres formes de chimiothérapie	O471 Faux travail à ou après la 37ème semaine entière de gestation
Z099 Examen de contrôle après traitement pour une affection, sans précision	Z370 Naissance unique, enfant vivant
O044 Avortement médical incomplet, sans complication	(39 occurence(s))
Z5100 Séance de préparation à une irradiation	
C61 Tumeur maligne de la prostate	Z358 Surveillance d'autres grossesses à haut risque
R509 Fièvre, sans précision	(37 occurence(s))
C900 Myélome multiple	
G473 Apnée du sommeil	Z340 Surveillance d'une première grossesse normale
Z358 Surveillance d'autres grossesses à haut risque	Z370 Naissance unique, enfant vivant
G35 Sclérose en plaques	(26 occurence(s))
Z515 Soins palliatifs	
L030 Phlegmon des doigts et des orteils	Z371 Naissance unique, enfant mort-né
K922 Hémorragie gastro-intestinale, sans précision	O358 Soins maternels pour d'autres anomalies et lésions foetales
E107 Diabète sucré insulino-dépendant, avec complications multiples	(présumées)
Z488 Autres soins de contrôle chirurgicaux précisés	(8 occurence(s))
O470 Faux travail avant 37 semaines entières de gestation [Menace d'accouchement prématuré]	
S5250 Fracture fermée de l'extrémité inférieure du radius	Z340 Surveillance d'une première grossesse normale
R104 Douleurs abdominales, autres et non précisées	O48 Grossesse prolongée
S610 Plaie ouverte de(s) doigt(s) (sans lésion de l'ongle)	Z370 Naissance unique, enfant vivant
I255 Myocardiopathie ischémique	(6 occurence(s))
I110 Cardiopathie hypertensive, avec insuffisance cardiaque (congestive)	
N390 Infection des voies urinaires, siège non précisé	Z370 Naissance unique, enfant vivant
H259 Cataracte sénile, sans précision	O757 Accouchement par voie vaginale après une césarienne
G409 Épilepsie, sans précision	O48 Grossesse prolongée
R102 Douleur pelvienne et périnéale	(4 occurence(s))
Z21 Infection asymptomatique par le virus de l'immunodéficience humaine [VIH]	
K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène)	Z358 Surveillance d'autres grossesses à haut risque
E660 Obésité due à un excès calorique	O48 Grossesse prolongée
Z751 Sujet attendant d'être admis ailleurs, dans un établissement adéquat	O700 Déchirure obstétricale du périnée, du premier degré
C504 Tumeur maligne du quadrant supéro-externe du sein	(2 occurence(s))
K573 Diverticulose du côlon, (sans perforation ni abcès)	
L024 Abcès cutané, furoncle et anthrax d'un membre	O364 Soins maternels pour mort intra-utérine du foetus
M511 Atteintes d'un disque lombaire et d'autres disques intervertébraux avec radiculopathie	N751 Abcès de la glande de Bartholin
S663 Lésion traumatique du muscle et du tendon extenseurs d'un autre doigt au niveau du poignet et de la main	Z371 Naissance unique, enfant mort-né
	(1 occurence(s))
I743 Embolie et thrombose des artères des membres inférieurs	
S611 Plaie ouverte de(s) doigt(s) avec lésion de l'ongle	
Z470 Soins de contrôle impliquant l'enlèvement d'une plaque et autre prothèse interne (orthopédique) de fixation	
A46 Érysipèle	
E105 Diabète sucré insulino-dépendant, avec complications vasculaires périphériques	
I500 Insuffisance cardiaque congestive	
E117 Diabète sucré non insulino-dépendant, avec complications multiples	
J450 Asthme à prédominance allergique	
R100 Syndrome abdominal aigu	
Z018 Autres examens spéciaux précisés	
C851 Lymphomes à cellules B, sans précision	
J159 Pneumopathie bactérienne, sans précision	
R53+1 Malaise	
L97 Ulcère du membre inférieur, non classé ailleurs	
K010 Dents incluses	
J459 Asthme, sans précision	
B182 Hépatite virale chronique C	
C509 Tumeur maligne du sein, sans précision	
F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance	
I633 Infarctus cérébral dû à une thrombose des artères cérébrales	
D500 Anémie par carence en fer secondaire à une perte de sang (chronique)	
K802 Calcul de la vésicule biliaire (sans cholécystite)	
Z538 Acte non effectué pour d'autres raisons	

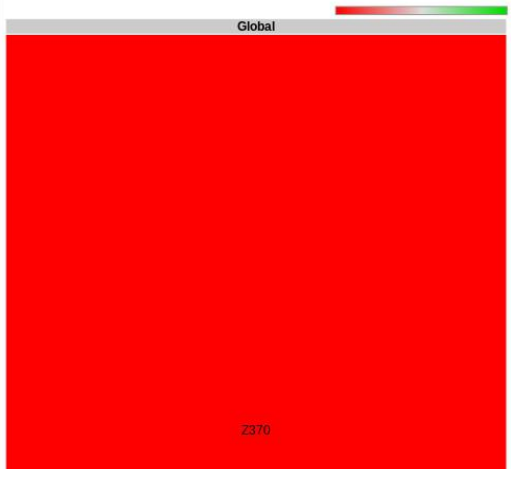


Figure 15 – Diagnostics associés pour le CHU de Fort-de-France en 2008 pour le diagnostic principal « Accouchements »

4.1.3. Durées de séjours des patients

Ce module permet d'extraire, à partir de résumés de données, les durées de séjours différenciés pour tous les diagnostics ou pour un diagnostic principal spécifique (*Figure 16*).

CEREGMIA - PREG Santé (Chaire PMSI) : Durées de séjours

2007

Guadeloupe

C.H.U. DE POINTE-A-PITRE/ABYMES

Tous les DP

Exécutée en 0.000869 secondes.

Age	Durée moyenne Janvier	Durée moyenne Février	Durée moyenne Mars	Durée moyenne Avril	Durée moyenne Mai	Durée moyenne Juin	Durée moyenne Juillet	Durée moyenne Août	Durée moyenne Septembre	Durée moyenne Octobre	Durée moyenne Novembre	Durée moyenne Décembre
0	H:11.6349 F:9.2254	H:7.7581 F:10.7826	H:17.6000 F:8.4386	H:7.2545 F:19.9250	H:9.0167 F:11.1951	H:18.1364 F:9.6735	H:10.5000 F:9.4615	H:9.9467 F:12.7451	H:11.8947 F:14.0862	H:8.1176 F:13.5000	H:5.4203 F:7.7833	H:12.6129 F:10.1000
1	H:1.8333 F:2.7826	H:2.3077 F:1.7619	H:2.3600 F:2.0500	H:1.4783 F:1.9375	H:20.0500 F:1.9444	H:2.2593 F:4.8261	H:3.7143 F:1.8235	H:1.1250 F:1.6500	H:2.0952 F:2.0588	H:2.6452 F:2.2800	H:2.0333 F:2.2000	H:2.5500 F:1.6087
2	H:2.0769 F:2.4667	H:1.7143 F:6.3750	H:1.8800 F:2.6111	H:3.3333 F:5.6500	H:3.9048 F:2.0667	H:1.2632 F:3.4118	H:2.6000 F:3.7895	H:2.0500 F:1.9375	H:1.4211 F:2.6364	H:2.5926 F:0.8000	H:2.3636 F:1.5625	H:1.4706 F:2.9412
3	H:1.2500 F:1.4000	H:2.5556 F:2.4444	H:1.4615 F:2.7333	H:2.7500 F:0.7500	H:0.8750 F:0.8750	H:2.1667 F:3.8889	H:0.5000 F:2.9286	H:1.2857 F:3.0833	H:1.3750 F:1.7143	H:1.3077 F:2.2308	H:2.3333 F:8.6667	H:1.7143 F:1.7222
4	H:2.0625 F:4.2000	H:1.8667 F:3.5000	H:1.6667 F:1.3333	H:4.2857 F:2.8333	H:15.7273 F:1.5556	H:1.9091 F:2.0000	H:1.8889 F:1.3333	H:1.6364 F:1.2857	H:2.7778 F:2.5556	H:3.6667 F:2.9000	H:3.5714 F:1.2500	H:1.2308 F:1.8750
5	H:1.3571 F:2.2500	H:3.1111 F:1.2308	H:1.2000 F:0.7500	H:1.4444 F:1.7778	H:3.7778 F:1.5556	H:2.0000 F:1.3333	H:1.1875 F:4.1429	H:6.6000 F:1.9375	H:1.6667 F:1.4444	H:1.0000 F:1.1429	H:3.9000 F:3.8333	H:0.7143 F:6.5556
6	H:2.3333 F:3.2000	H:0.5000 F:9.4000	H:3.4545 F:2.6667	H:2.0833 F:0.7500	H:2.8000 F:1.1000	H:2.5714 F:1.3333	H:5.5000 F:2.1667	H:2.1333 F:0.8000	H:1.6000 F:2.0000	H:1.6923 F:2.2000	H:1.9000 F:6.1111	H:2.6667 F:1.5000
7	H:2.2000 F:2.6000	H:4.9000 F:2.0833	H:1.6250 F:1.8571	H:2.1667 F:11.1429	H:1.1667 F:1.7000	H:1.5000 F:1.7500	H:2.0000 F:1.6000	H:1.4000 F:2.5714	H:3.2857 F:1.8571	H:2.7333 F:3.2667	H:2.7143 F:3.4286	H:2.2857 F:4.4444
8	H:3.4000 F:2.0000	H:2.1429 F:1.5000	H:2.4286 F:2.1250	H:3.1818 F:1.0000	H:7.8333 F:13.5714	H:1.0000 F:2.0000	H:2.0000 F:1.2000	H:2.7778 F:3.0000	H:3.3077 F:2.7500	H:2.2727 F:2.7143	H:1.8889 F:7.0000	H:2.9091 F:2.8571
9	H:9.3333 F:2.5833	H:2.5714 F:3.0000	H:2.7000 F:0.7143	H:7.6667 F:1.2222	H:2.8750 F:2.0000	H:1.9000 F:7.8889	H:1.3636 F:3.2500	H:3.0000 F:3.5714	H:3.0000 F:1.7500	H:2.9000 F:3.0000	H:5.8571 F:1.8750	H:0.5000 F:2.6000
10	H:3.0833 F:3.7778	H:0.8333 F:3.0000	H:3.3333 F:3.8571	H:1.3750 F:1.5000	H:1.6154 F:2.5000	H:4.4444 F:1.2000	H:1.5000 F:1.2000	H:1.2000 F:1.2000	H:2.2500 F:6.0000	H:4.6667 F:2.5000	H:2.5714 F:4.2000	H:7.5000 F:3.2000
11	H:3.7000 F:1.4286	H:1.6667 F:4.3750	H:2.0000 F:5.3000	H:5.4286 F:1.0000	H:2.5000 F:1.8000	H:4.5000 F:5.8333	H:2.0000 F:5.2222	H:5.3636 F:3.0000	H:1.6667 F:2.1667	H:3.7273 F:2.1818	H:4.7500 F:2.0000	H:0.5000 F:14.6667
12	H:5.2727 F:1.8571	H:2.1111 F:2.5000	H:2.5714 F:1.0000	H:1.4000 F:2.5556	H:1.3000 F:3.3333	H:1.5000 F:3.0000	H:2.2000 F:5.7143	H:2.4615 F:2.6000	H:0.8889 F:0.6000	H:4.6429 F:1.5000	H:2.5455 F:2.8333	H:3.0000 F:2.0000
13	H:2.3636 F:1.5000	H:7.0000 F:2.2500	H:1.7143 F:1.0000	H:2.0000 F:3.3333	H:7.6364 F:2.3333	H:1.9091 F:2.3750	H:3.3333 F:1.6667	H:2.2000 F:5.0000	H:3.6667 F:2.6667	H:5.6667 F:3.0000	H:4.4000 F:4.0000	H:1.8889 F:5.0000
14	H:2.3125 F:1.8750	H:1.8750 F:3.2000	H:1.7778 F:7.5000	H:0.8333 F:2.5000	H:2.6923 F:3.2727	H:4.3333 F:2.5000	H:3.8000 F:2.2000	H:2.7778 F:3.1250	H:1.1250 F:2.1429	H:2.1111 F:2.7143	H:2.4000 F:2.1111	H:3.8571 F:4.4545
15	H:2.0000 F:2.1111	H:1.7500 F:3.6250	H:3.1429 F:3.6667	H:0.8750 F:1.5714	H:1.5714 F:3.1667	H:1.6000 F:3.1429	H:6.1429 F:1.3750	H:1.3750 F:3.5556	H:3.5385 F:2.5714	H:6.8750 F:1.1111	H:1.5000 F:2.6000	H:3.0833 F:2.5714
16	H:5.4000 F:4.7273	H:1.0000 F:5.8000	H:6.5000 F:1.8750	H:2.2857 F:2.7500	H:3.4444 F:3.0000	H:5.6000 F:2.0000	H:4.7778 F:2.4444	H:2.0000 F:4.0000	H:1.5000 F:2.5833	H:1.0000 F:2.8182	H:4.4286 F:3.9167	H:1.5000 F:2.7500
17	H:2.8889 F:4.8750	H:3.3333 F:4.6000	H:2.6000 F:4.4167	H:1.6000 F:1.9000	H:4.2000 F:1.7778	H:2.0000 F:3.2727	H:3.5714 F:3.1765	H:3.8889 F:5.0000	H:2.7778 F:2.1667	H:6.4167 F:1.9000	H:5.0909 F:2.7273	H:1.8571 F:3.7647
18	H:2.3333 F:4.5600	H:16.2000 F:5.3684	H:1.3333 F:4.2500	H:2.4286 F:2.8000	H:5.3333 F:3.0909	H:2.8571 F:4.2727	H:1.5000 F:3.8750	H:2.0556 F:3.0000	H:3.6667 F:2.7143	H:9.5556 F:4.4375	H:1.0000 F:3.0000	H:4.8571 F:6.3846
19	H:3.9231 F:3.6000	H:3.6667 F:3.0769	H:3.0000 F:3.0000	H:5.6250 F:5.2143	H:5.5455 F:2.3333	H:0.7500 F:3.4118	H:4.2667 F:7.2105	H:3.3333 F:2.9231	H:3.9286 F:4.0000	H:4.6000 F:4.4286	H:2.3333 F:4.1538	H:1.5000 F:4.2632
20	H:3.5714	H:6.1818	H:4.2857	H:12.2500	H:2.3333	H:8.9000	H:1.6667	H:5.7500	H:2.5000	H:1.8750	H:6.5000	H:3.5714

Figure 16 – Durées de séjours Hommes et Femmes tous diagnostics principaux confondus pour le CHU de Pointe-à-Pitre en 2007

4.1.4. Durées de séjours des patients hors Région

Ce module permet d'extraire, à partir de résumés de données, les durées de séjours

Ce module permet d'afficher, pour une année et pour un établissement, les diagnostics principaux des patients pris en charge non originaire de la région d'implantation de l'établissement (*Figure 17*).

Au regard de l'organisation de l'offre de soins, ce module permet de conduire des analyses sur le « taux de fuite » pour les patients résidents français. L'indicateur « Taux de fuite » correspond au rapport entre le nombre de séjours de la zone géographique sélectionnée pris en charge en dehors de cette zone sur le nombre total de séjours issus de la zone sélectionnée (une zone géographique qui ne contient aucun établissement hospitalier présente un taux de fuite de 100 %). En outre, ce module est une des bases à un module destiné à quantifier le tourisme médical. Il est aussi important pour générer la cartographie des diagnostics principaux/Origine géographique des patients.

2007

Guadeloupe

C.H.U. DE POINTE-A-PITRE/ABYMES

Tous les DP

Exécutée en 0.000975 secondes.

Age	Durée moyenne Janvier	Durée moyenne Février	Durée moyenne Mars	Durée moyenne Avril	Durée moyenne Mai	Durée moyenne Juin	Durée moyenne Juillet	Durée moyenne Août	Durée moyenne Septembre	Durée moyenne Octobre	Durée moyenne Novembre	Durée moyenne Décembre
0	H:8.1892 F:7.3590	H:6.1068 F:7.5244	H:11.6383 F:6.1800	H:5.9286 F:12.4189	H:7.3953 F:7.6667	H:12.3148 F:6.7976	H:7.8019 F:6.7579	H:7.7478 F:9.4250	H:8.8977 F:10.4667	H:6.4375 F:9.2697	H:4.5565 F:6.4457	H:9.4300 F:7.3645
1	H:1.8000 F:2.3103	H:2.0000 F:1.4231	H:2.2692 F:1.9524	H:1.4167 F:1.7778	H:19.0952 F:1.9474	H:1.9063 F:4.0714	H:3.6207 F:1.7778	H:1.0000 F:1.5714	H:2.0000 F:1.9474	H:2.5143 F:2.0000	H:1.8684 F:2.1176	H:2.5455 F:1.6087
2	H:1.8235 F:2.4667	H:1.5000 F:5.1000	H:1.5484 F:2.0435	H:2.9583 F:5.1364	H:3.9048 F:1.6842	H:1.1304 F:2.7619	H:2.4375 F:3.4286	H:1.7826 F:1.6000	H:1.3810 F:2.6364	H:2.3750 F:0.8750	H:2.2917 F:1.5294	H:1.4762 F:2.8333
3	H:1.2500 F:1.4000	H:2.0909 F:2.4444	H:1.1875 F:2.7333	H:2.3333 F:0.7500	H:0.7778 F:0.7778	H:1.9286 F:3.6000	H:0.3333 F:2.5625	H:1.2000 F:2.5333	H:1.3529 F:1.6000	H:1.2353 F:2.0000	H:2.3684 F:7.3636	H:1.6957 F:1.7368
4	H:2.0625 F:3.0000	H:1.7647 F:3.1111	H:1.3636 F:1.0000	H:3.3333 F:2.4286	H:15.7273 F:1.4000	H:1.9091 F:2.0000	H:1.8889 F:1.0909	H:1.5000 F:1.2857	H:2.3636 F:2.7000	H:3.4000 F:2.7273	H:3.4000 F:1.3000	H:1.2667 F:1.8750
5	H:1.3333 F:2.2500	H:2.8000 F:1.2308	H:1.2500 F:0.7333	H:1.4444 F:1.4545	H:3.6000 F:1.4000	H:2.0000 F:0.8000	H:1.1667 F:2.9000	H:4.7143 F:1.9375	H:1.5000 F:1.4444	H:1.0000 F:1.1429	H:3.9000 F:3.8333	H:0.7143 F:6.5556
6	H:2.1333 F:3.2000	H:0.3750 F:9.4000	H:2.9231 F:2.6667	H:1.9231 F:0.6000	H:2.4706 F:1.1000	H:2.4000 F:1.3333	H:4.4000 F:1.8000	H:2.0625 F:0.6667	H:1.6364 F:2.0000	H:1.6923 F:2.2000	H:1.9000 F:6.1111	H:2.6667 F:1.5000

Figure 17 – Durées de séjours Hommes et Femmes tous diagnostics principaux confondus pour le CHU de Pointe-à-Pitre en 2007

Conclusion

Nous nous sommes attachés dans cet article à présenter des outils établis à partir de clusters au service de l'amélioration de la connaissance du client au sens générique, de la prédiction de ses comportements et de l'optimisation de l'offre proposée dans un environnement spécifique qu'est l'hôpital. Nous avons tenté de mettre en évidence l'importance de la fouille de données comme un puissant levier dans la connaissance du client, au bénéfice duquel l'hôpital créé un service (prise en charge, au sens global). Nous avons identifié des outils, qui proposé aux spécialistes des données du domaine (chercheurs, établissements de santé, ARS, DHOS, etc.), permettent d'optimiser l'offre de soins.

Cette recherche n'a cependant pas permis d'intégrer à ce stade l'articulation entre connaissance de l'activité et performance des établissements au sens rapport qualité-coût des prestations délivrées par les établissements. Il conviendra de proposer des outils qui prendront en compte cette deuxième composante du PMSI, en mettant l'accent sur la valorisation des séjours et l'incidence financière du parcours patient en se basant notamment sur les algorithmes de groupage utilisés par l'ATIH. D'autres outils mettant l'accent sur l'optimisation des dépenses seront également étudiés (consommation médicamenteuses…).

5 Analyse de transactions financieres

La Loi AML (Anti-Money Laundering) impose aux professionnels assujettis un certain nombre d'obligations visant à prévenir et détecter le Blanchiment et le Financement du Terrorisme.

5.1 Hemdal : Lutte contre le Blanchiment – Idavoll-Transact

Les quarante recommandations du Groupe d'action financière sur le blanchiment des capitaux (GAFI), c'est-à-dire des normes internationales qui encadrent la lutte contre " l'argent sale " ont été révisées en juin 2003. Depuis le 11 septembre 2001, une profonde transformation des services financiers a été mise en place. Cette transformation implique pour nous une part algorithmique et logicielle, et autre part organisationnelle. Le premier volet de la transformation vise à identifier les transactions liées aux activités malveillantes, lutter contre le " financement du terrorisme ". La gestion de cette obligation s'est inspirée des acquis de la lutte anti-blanchiment et a conduit à rapprocher les services de contrôle de l'appréciation des risques informatique.

La société Uniskip a un processus par lequel il faut passer afin de devenir client. Ensuite pour chaque transaction un contrôle que nous allons détailler est réalisé.

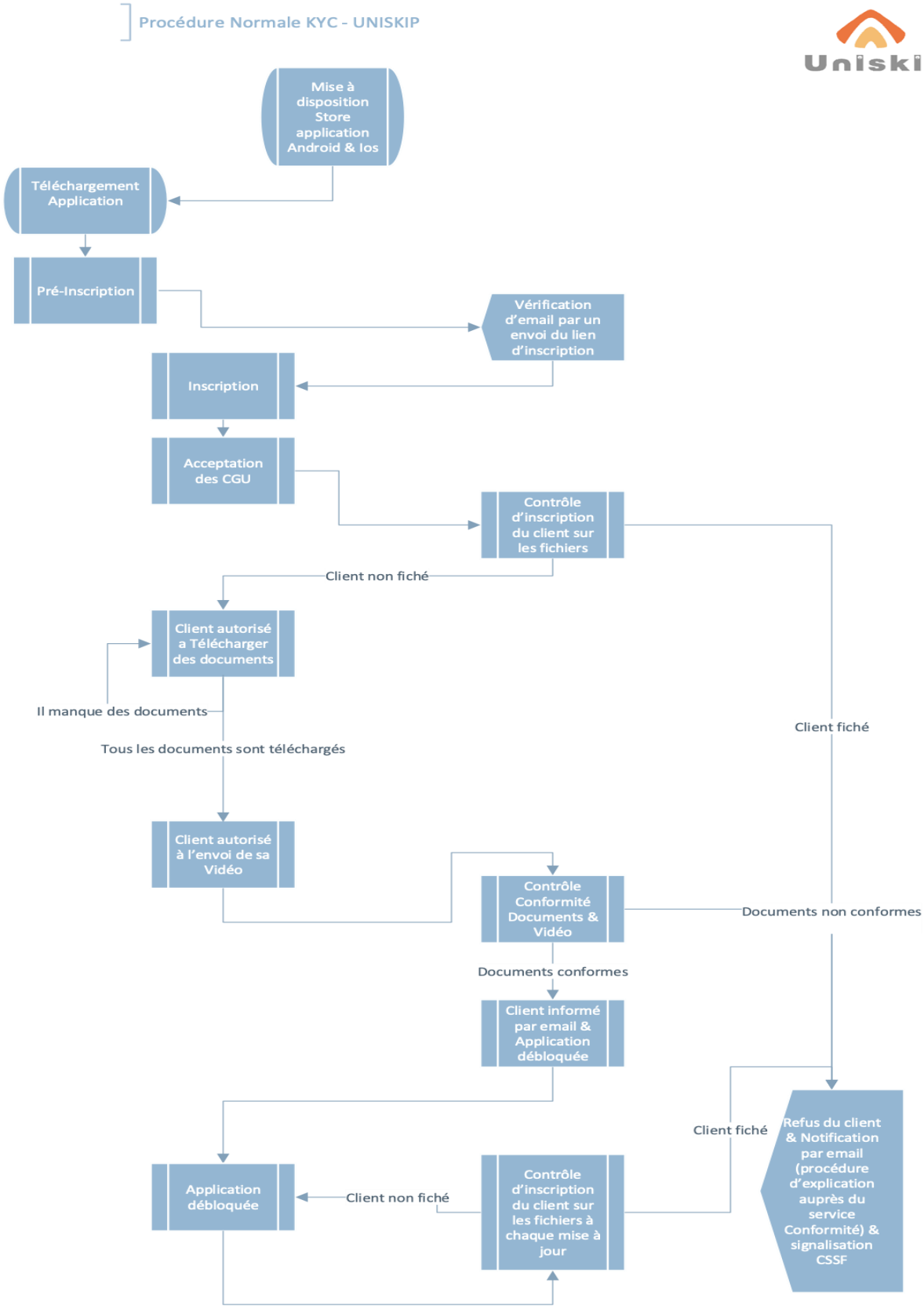


Figure 18 : Procédure d'acceptation d'un nouveau client

Nous nous basons sur des règles d'identification des clients et de conservation des documents. Cette étape est importante car elle permet de vérifier qu'un utilisateur peut être accepté dans le système avant d'utiliser le module Hemdal. Certains algorithmes de reconnaissance faciale ont été utilisés mais ne seront pas développés ici car il y a eu peu de recherche mais plutôt de l'optimisation de l'utilisation du réseau CNN afin d'avoir un résultat rapide.

1. UNISKIP ne détiendra pas de comptes anonymes, ni de comptes sous des noms manifestement fictifs : l'entreprise est tenue (par des lois, des règlements, des accords entre autorités de contrôle et institutions financières, ou par des accords d'autodiscipline entre institutions financières) d'identifier, sur la base d'un document officiel ou d'une autre pièce d'identité fiable, ses clients habituels ou occasionnels, et d'enregistrer leur identité, lors de l'entrée en relation ou lors des transactions (en particulier dans le cadre d'ouverture de compte, transactions fiduciaires, ou de transactions importantes en espèces).

Afin de satisfaire aux exigences d'identification concernant les personnes morales, UNISKIP prendra les mesures suivantes, le cas échéant:

- (i) vérifier l'existence et la structure juridique du client en obtenant de celui-ci ou à partir d'un registre public, ou bien grâce à ces deux sources, une preuve de la constitution en société comprenant des renseignements concernant le nom du client, sa forme juridique, son adresse, les dirigeants et les dispositions régissant le pouvoir d'engager la personne morale;
- (ii) vérifier que toute personne prétendant agir au nom du client est autorisée à le faire et identifier cette personne.

2. UNISKIP prend des mesures raisonnables pour obtenir des informations sur l'identité véritable des personnes dans l'intérêt desquelles un compte est ouvert ou une transaction est effectuée, s'il y a le moindre doute sur le fait que ces clients pourraient ne pas agir pour leur propre compte, par exemple dans le cas de sociétés de domicile (c'est-à-dire des institutions, des sociétés, des fondations, des fiducies, etc. qui ne se livrent pas à des opérations commerciales ou industrielles, ou à toute autre forme d'activité commerciale, dans le pays où est situé leur siège social).

3. UNISKIP conserve pendant au moins cinq ans toutes les pièces nécessaires se rapportant aux transactions effectuées, à la fois nationales et internationales, afin de leur permettre de répondre rapidement aux demandes d'informations des autorités compétentes. Ces pièces doivent permettre de reconstituer les transactions individuelles (y compris les montants et les types d'espèces en cause, le cas échéant) de façon à fournir, si nécessaire, des preuves en cas de poursuites pour conduite criminelle. UNISKIP conserve une trace écrite de la justification d'identité de ses clients (par exemple, copie ou enregistrement des documents officiels comme les passeports, les cartes d'identité, les permis de conduire, ou des documents similaires), les livres de comptes et la correspondance commerciale pendant cinq ans au moins après la clôture du compte.

4. UNISKIP apporte une attention particulière à toutes les opérations complexes, importantes, et à tous les types inhabituels de transactions, lorsqu'elles n'ont pas de cause économique ou licite apparente. l'arrière-plan et l'objet de telles opérations devraient être examinés, dans la mesure du possible ; les résultats de cet examen devraient être établis par écrit, et être disponibles pour aider les autorités de contrôle, de détection et de répression, les commissaires aux comptes et les contrôleurs internes ou externes.

5. [……]

6. [……]

7. UNISKIP, en cas de situation litigieuse nécessitant une déclaration de soupçon aux autorités compétentes, n’avertira pas ses clients de la procédure mise en œuvre.

8. UNISKIP déclarant ses soupçons doit se conformer aux instructions en provenance des autorités compétentes.

9. UNISKIP a mis au point un programme de lutte contre le blanchiment de capitaux, qui comprend :

(i) des politiques, des procédures et des contrôles internes, y compris la désignation de personnes responsables au niveau de la direction générale, et des procédures adéquates lors de l'embauche des employés, de façon à s'assurer qu'elle s'effectue selon des critères exigeants ;

(ii) [……]

(iii) [……]

Mesures pour faire face aux problèmes des pays dépourvus totalement ou partiellement de dispositifs de lutte contre le blanchiment de capitaux

10. [……].

11. UNISKIP devrait porter une attention particulière à ses relations d'affaires et à ses transactions avec les personnes physiques et morales, y compris les sociétés ou les institutions financières, résidant dans les pays qui n'appliquent pas ou trop peu les présentes recommandations. Lorsque ces transactions n'ont pas de cause économique ou licite apparente, leur arrière-plan et leur objet devraient être examinés dans la mesure du possible. Les résultats de cet examen devraient être établis par écrit, et être disponibles pour aider les autorités de contrôle, de détection et de répression, les commissaires aux comptes et les contrôleurs internes ou externes.

J'ai donc du développer le Module Idavoll-transact lié à la gestion des transactions afin d'en faire un outil d'apprentissage automatique, permettant d'estimer et suivre des modèles de scoring des transactions afin de détecter les transactions atypiques.

Le Module Idavoll-transact est divisé en deux modules :

- Sélection des Prédicats et affinage des variables (manuel ou automatique)
- Sélection des Prédicats permettant de délimiter les variables utilisées à partir de l'ensemble de caractéristiques liées à l'Uniskiper.

Les variables sont de trois types que nous détaillons dans le 5.1.1

Les prédicats ont deux formes, une forme figée par le gestionnaire, et une forme qui évolue en fonction des transactions réalisées et qui évolue dans le cadre d'un apprentissage automatique.

Le module crée un classement des variables sur la base de trois mesures. Grâce à ces trois types de variables, le module identifie les caractéristiques qui ont un impact important sur le caractère de la transaction et son atypique. L'ensemble des variables peut être comparé à une matrice ayant un nombre de dimensions inconnus. Chaque dimension de la matrice est corrélation entre toutes les variables. Un poids pour chaque variable est défini par l'utilisateur dans la première forme des prédicats (forme figée). Un second poids est calculé en fonction des transactions réelle. L'écart type entre ces deux poids

permet de calculer la distance (ou correlation) entre l'apprentissage automatique issu des transactions et la forme figée.

Le module Sélection des Prédicats permet de délimiter les variables utilisées à partir de l'ensemble des caractéristiques liées à l'Uniskiper (utilisateur Uniskip).

Le contrôle de la transaction passe par un premier contrôle qui vise à valider que l'utilisateur a un compte actif (non bloqué par le service conformité), et qu'il a bien passé toutes les étapes de la conformité en fonction de son profil conformément aux procédures internes (Type A €<250 € mois ou Normal).

Le contrôle passe ensuite en revue le profil AML client Age [tranche], origine géographique du client (Pays/Région ou ville), ancienneté, type d'utilisateur, nombre de contacts Uniskiper, et vérifie que celui-ci a bien de droit de réaliser la transaction en fonction des seuils AML correspondants (montant, position gps de la transaction, destination géographique de la transaction, motif de la transaction, type de transaction), montant moyen autorisé sur 12 mois, fréquence des transactions autorisées dernière minute, fréquence des transactions autorisées dernière heure, fréquence des transactions autorisées dernier jour, fréquence des transactions autorisées dernière semaine, fréquence des transactions autorisées dernier mois, fréquence des transactions autorisées dernière année, fréquence des transactions autorisées dernières depuis le début la relation, délai entre approvisionnement et dépense.

Le contrôle passe enfin dans un moteur d'intelligence artificielle (utilisant un réseau de neurones). Dans un premier temps le système ne sera pas bloquant et apprendra de la base complète des utilisateurs. Le système établit une classification automatique des transactions en fonction des utilisateurs afin de proposer un état en temps réel des paramètres des transactions actuelles et passées. Cela nous permet d'ajuster au mieux les paramètres du premier sas. Lorsque le moteur d'intelligence artificielle aura suffisamment appris des comportements des Uniskiper, nous pourrons ajuster en temps réel les paramètres du premier sas en le rendant un peu plus strict ou un peu plus souple.

5.1.1 Détection des transactions atypiques par apprentissages automatiques : une approche multidimensionnelle

Notre approche est liée à l'utilisation de critères multiples. Ces critères sont les dimensions de notre environnement de recherche. Nous sélections des Prédicats et ensuite nous réalisons un affinage des variables (manuelle ou automatiquement en utilisant la rétropropagation).

Nous pouvons résumer le processus de paiement par le schéma suivant :

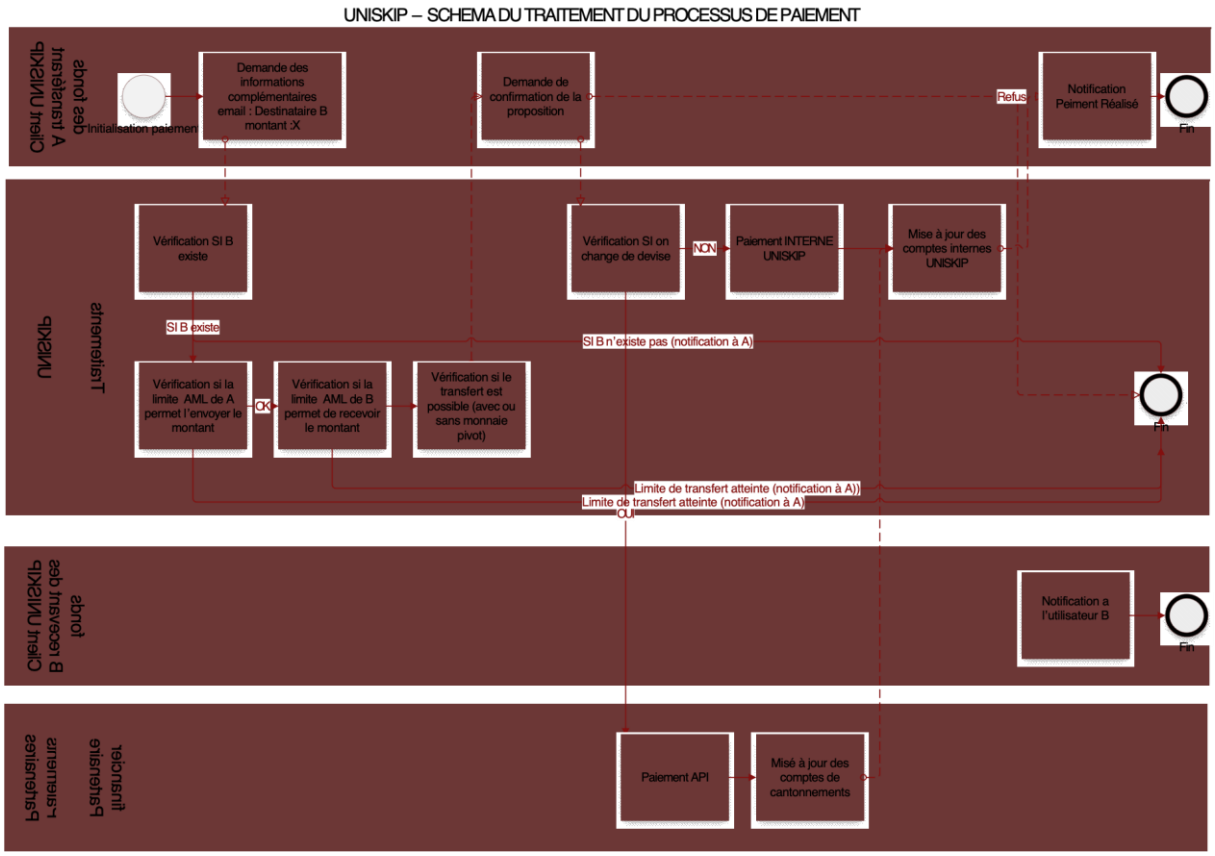


Figure 19 : Process global de paiement Uniskip

Le module Sélection des Prédicats permet de délimiter les variables utilisées à partir de l'ensemble des caractéristiques liées à l'Uniskiper (utilisateur Uniskip).

Nos variables sont de trois types :

- le premier type des variables est « relativement » figé et fortement lié à l'utilisateur (exemple Age, Origine géographique, ancienneté, type d'utilisateur, nombre de contact uniskiper,)
- le second type de variables est lié à la transaction (montant, moyenne montant, origine géographique de la transaction, destination géographique de la transaction, motif)
- le troisième type de variable est fortement lié à la temporalité (fréquence des transactions (dernière minute, dernière heure, dernier jour, dernière semaine, dernier mois, dernière année), délai entre approvisionnement et dépense.

Les prédicats que nous avons définis ont deux formes, une forme figée par le gestionnaire, et une forme qui évolue en fonction des transactions réalisées et qui évolue dans le cadre d'un apprentissage automatique.

Il n'y a pas de nécessité identifiées de détecter des motifs de hauts niveaux sur l'ensemble des données avec des méthodes types Réseaux convultifs ConvNet présenté par LeCun 1995 [12] en raison de la structure de données qui ne semble pas s'y prêter. Nous pouvons schématiser le processus dans lequel nous nous situons (Compliance IA) figure[19].

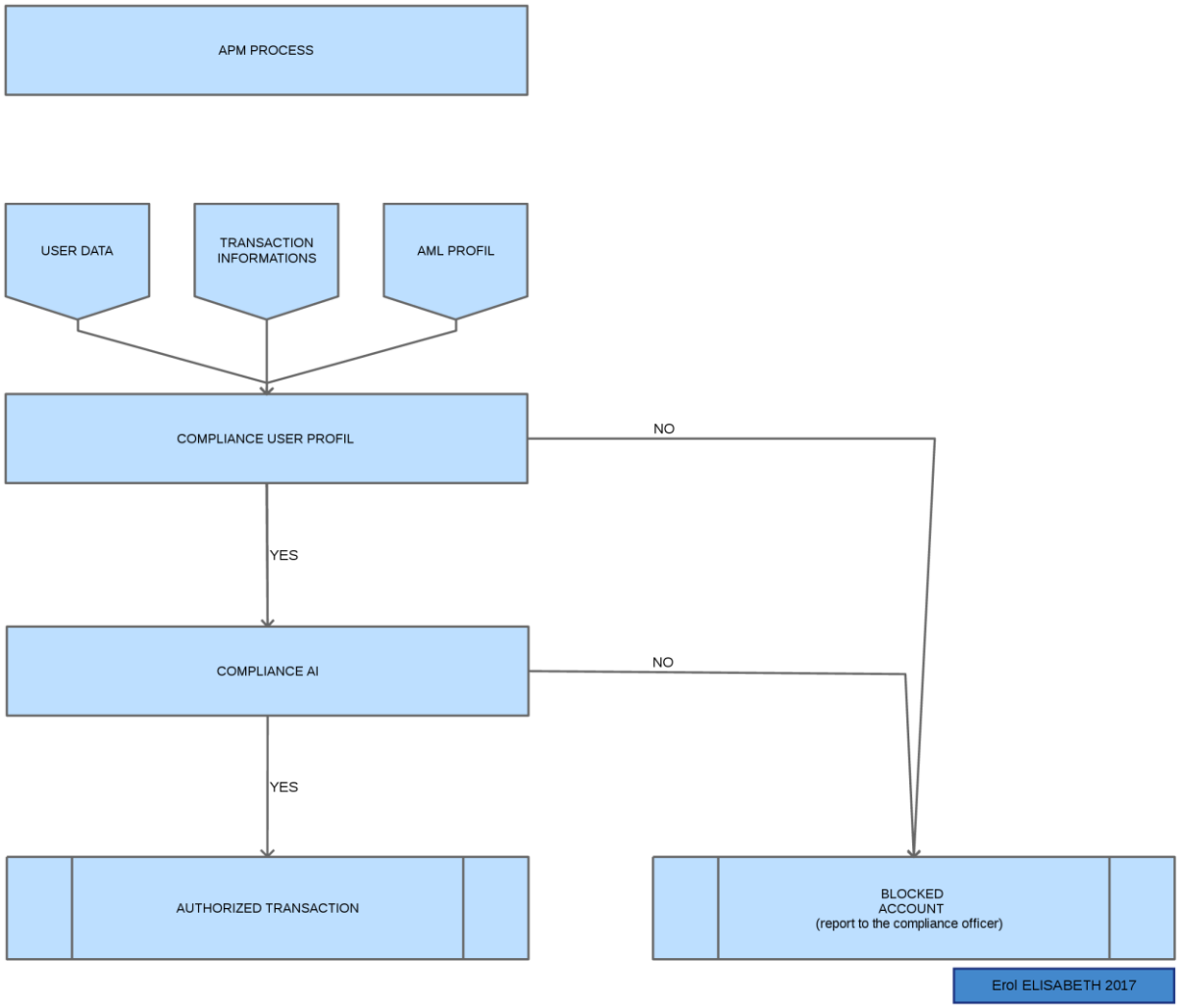


Figure 19 Process d'acceptation d'une transaction

Dans la gestion des transactions atypiques, notre objectif est d'arriver à un résultat qui vise à accepter ou non la transaction. Notre système vise aussi l'apprentissage automatique dans le même cycle, le résultat de cet apprentissage automatique est aussi versé dans un système permettant d'alimenter la base de connaissance du client (scoring ou cartographie de comportement client).

Chaque détecteur permettra d'alimenter le réseau de façon individuel. Un poids (0 à 1) sera donné à chaque détecteur. La fin du cycle du réseau de neurones prendra plus ou moins fortement en compte le poids de chaque neurone intermédiaire afin de valider ou non la transaction.

Dans notre étude il n'est pas nécessaire de répéter la détection du noyau à plusieurs reprises, car les caractéristiques des poids sont discrètes et ne peuvent pas être répétées dans l'entrée.

Exemple : le montant de la transaction ne se répète pas deux fois dans l'entrée. Ce type de répétition pourrait prendre son sens dans la détection d'images ou il y a plusieurs visages.

L'entraînement du réseau s'est fait grâce à une liste de transactions réalisées par un GUF (groupe d'utilisateurs fermés). Ces utilisateurs réalisent des transactions.

La dernière couche du réseau est la réponse OUI ou NON « la transaction est-elle atypique ? ». Si on enlève la dernière couche et qu'on la stocke comme le résultat des données en entrées et que l'on pondère les poids et couches intermédiaires, nous avons un système qui apprend de l'expérience des entrées.

La première étape essentielle est principalement l'initialisation de l'algorithme. Nous n'utilisons pas de méthode aléatoire puisque nous avons les prédicats qui vont borner et orienter le traitement.

Dans notre approche d'apprentissage il faut pouvoir mettre en valeur les transactions « éloignées » donc peu fréquentes mais pourtant avec des variantes importantes (notamment fréquence, montant) représentants des transactions valides.

Il est nécessaire que l'apprentissage des transactions atypiques se fasse avec des transactions pas forcément « éloignées » mais dont les caractères sont quand même proches des transactions sûres. C'est-à-dire que les montants, positions géographiques, fréquences soient dans les bornes fixées par les règles.

Cette approche nous permet de mettre en place un apprentissage des transactions qui aboutit à une carte de distribution de la probabilité des éléments de la transaction qui se présente.

Nous pouvons nous baser sur la représentation structurelle d'un neurone artificiel qui calcule la somme de ses entrées puis cette valeur passe à travers la fonction d'activation pour produire sa sortie.

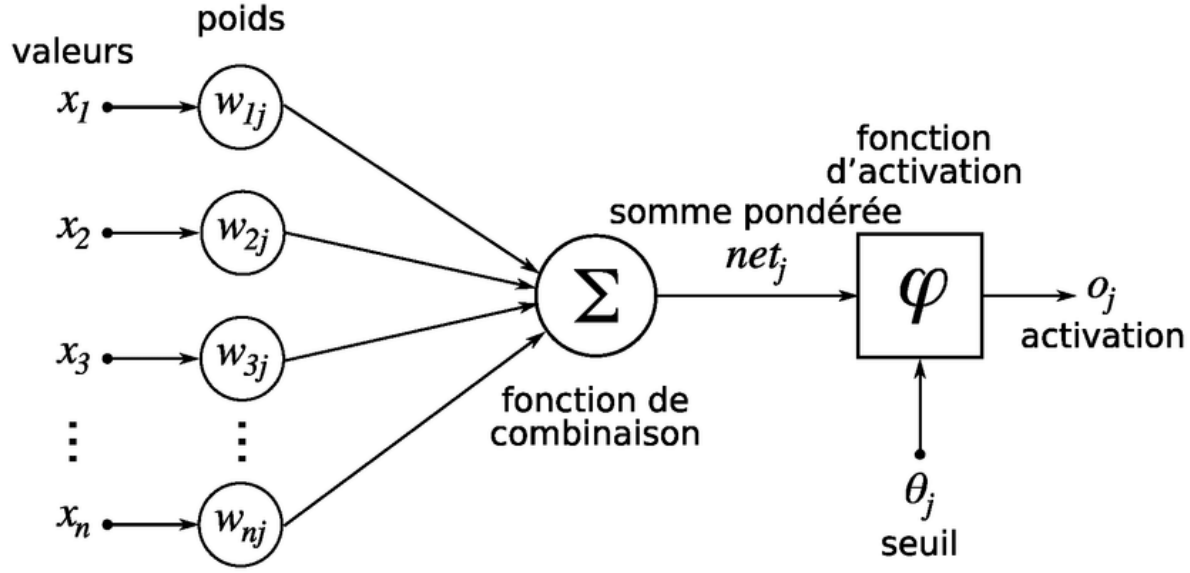


Figure 20 Structure d'un neurone artificiel. Le neurone

Notre objectif est alors de produire un réseau de neurones qui permet d'une part d'obtenir l'activation du neurone de sortie (Oui la transaction est sure même atypique, ou NON la transaction n'est pas sure).

Les informations suivantes sont confidentielles et ne doivent pas être publiées.

Nos entrées sont identifiées de la façon suivante :

X1 => Seuil de dépense plausible dans 1 journée : Montant max cash_out

Profil utilisateur (vient de user & aml_profil_user)

X2 => Age [tranche]

X3 => Origine géographique du client (Pays/Région ou ville)

X4 => Ancienneté (mois)

X5 => Type d'utilisateur

X6 => Nombre de contacts uniskiper

Transaction

X7 => Montant

X8 => X Y de la transaction (position GPS)

X9 => Destination géographique de la transaction

X10 => Code du motif

X11 => Type de transaction : [paiement entré, paiement sortie]

Aml_profil_transaction

X12 => % du solde du compte

X13 => Nombre de transactions entre les 2 utilisateurs (fréquences)

X14 => Montants de transactions entre les 2 utilisateurs

Aml_profil_user_ia

X15 => Montant (permet d'intégrer la rétropropagation en neurone d'entrée, ce n'est pas le montant de la transaction mais le montant calculé d'une telle transaction)

X16 => Moyenne montant sur 12 mois

X17 => Fréquence des transactions dernière minute

X18 => Fréquence des transactions dernière heure

X19 => Fréquence des transactions dernier jour

X20 => Fréquence des transactions dernière semaine

X21 => Fréquence des transactions dernier mois

X22 => Fréquence des transactions dernière année

X23 => Fréquence des transactions depuis le début la relation

X24 => Délai entre approvisionnement- et dépense

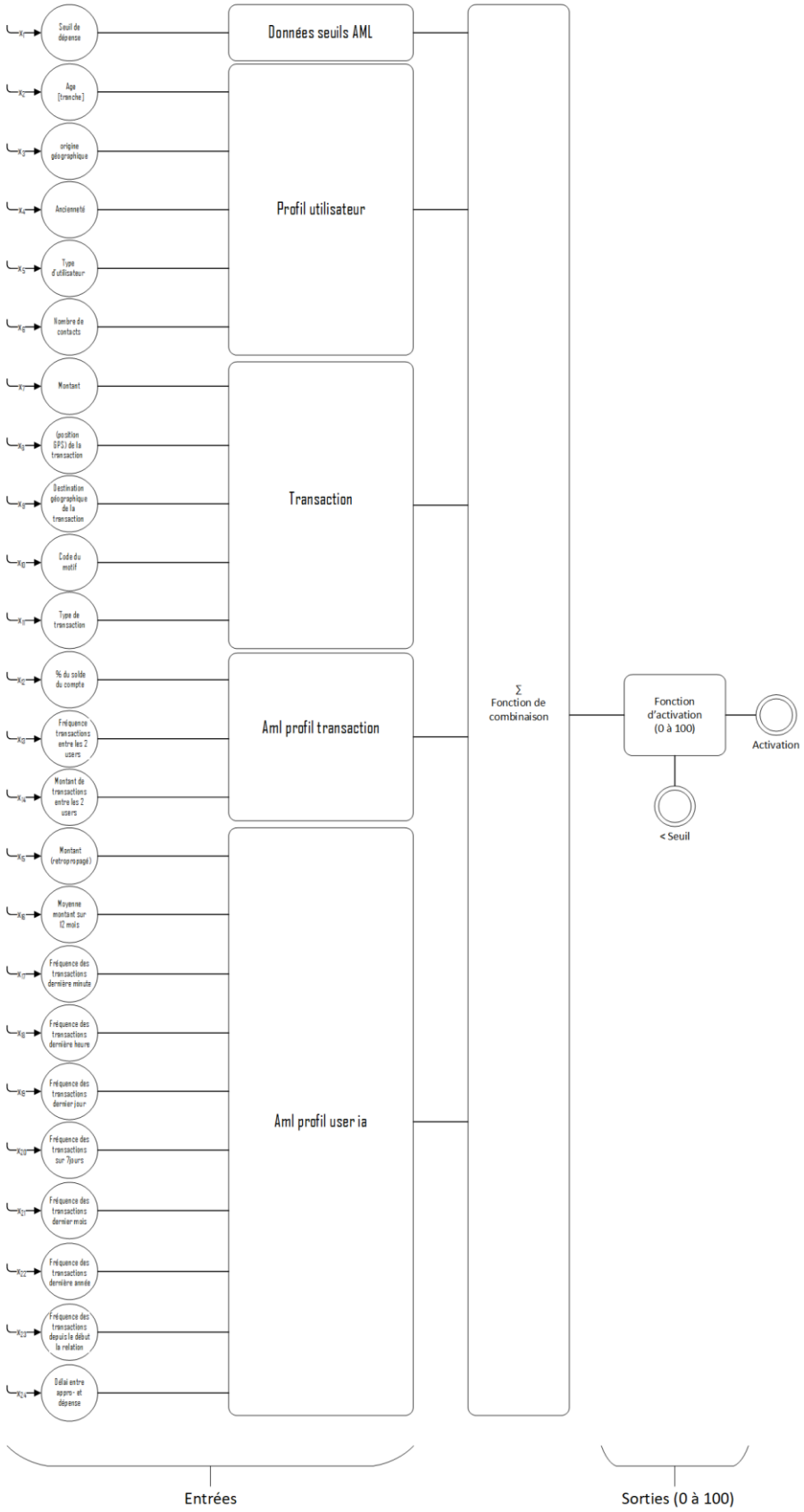


Fig 21 : Schéma du réseau de neurones

L'application industrielle de Hemdall

L'entreprise industrielle dans laquelle a été mis en production le Module est lié à la gestion des transactions financières afin d'en faire un outil d'apprentissage automatique, permettant d'estimer et suivre des modèles de scoring (en enlevant la dernière couche OUI/NON) des transactions afin de détecter les transactions atypiques.

Le Module Hemdal est divisé en plusieurs sous-parties.

Le module crée un classement des variables sur la base des 24 mesures. Grâce à ces 24 variables, le module identifie les caractéristiques qui ont un impact important sur le caractère de la transaction et son atypie.

L'ensemble des variables peut être comparé à une matrice ayant un nombre de dimensions inconnues. Chaque dimension de la matrice est en corrélation entre toutes les variables. Un poids pour chaque variable est défini par l'utilisateur dans la première forme des prédicats (forme figée). Un second poids est calculé en fonction des transactions réelle. L'écart type entre ces deux poids permet de calculer la distance (ou corrélation) entre l'apprentissage automatique issu des transactions et la forme figée.

Le module d'Estimation du Modèle permet d'estimer et comparer différents modèles. Pour évaluer ces modèles, nous choisissons parmi les mesures statiques suivantes (chacune avec un rapport détaillé complet)

- Courbes de Lift et de Gain avec une tolérance manuelle
- Valeur Informative (VI),
- Statistique de Kolmogorov - Smirnov,

Une évaluation régulière (manuelle ou semi-automatique) permet de corriger le modèle figé afin de le rentre plus strict ou moins strict en fonction du profil des utilisateurs.

Le module offre un certain nombre d'outils analytiques pour comparer deux jeux de données afin de détecter toute modification significative dans la structure des caractéristiques ou dans la population des emprunteurs. Toute distorsion significative dans la base de données peut nécessiter la ré-estimation des paramètres du modèle. Ce module produit des rapports de la stabilité des transactions afin d'évaluer en permanence l'efficacité du module

La Sélection des Seuils permet de définir les valeurs optimales pour séparer les transactions acceptables et les transactions douteuses. Le processus décisionnel est ensuite un signalement de la transaction, associée à un blocage de compte. Deux actions sont ensuite possibles : soit la transaction est acceptée manuellement avec un éventuel réajustement des variables acceptables pour ce profil, soit la transaction est notée comme atypique avec un blocage du compte de l'utilisateur et une notification aux services compétents.

6 Prévision d'activité pour les entreprises

Une grande partie des entreprises sont météo-dépendantes, c'est-à-dire que leurs résultats financiers et la satisfaction de leurs clients varient selon la météo ; les mouvements sociaux telles que les grèves, difficultés liées aux transports, arrivée et départ des touristes impactent également leurs résultats au quotidien.

6.1 Météo-Biz : Prévision d'activité pour les TPE

En fin d'année 2018, l'idée de recherche une solution de prediction d'activité des TPE a commencé à poindre. Le projet Météo-Biz a donc été formalisé au sein de l'entreprise Beepway.com. En 2020 l'apparition du COVID-19 en France et notamment dans les outremers nous conduit à la création d'une branche au projet Météo-Biz. Cette branche Méteo-Biz/COVID-19 vise à identifier des modèles de « comportements » de résistance des TPE des outres mer et de tenter de prédire la durée de vie des entreprises en se basant sur les premiers travaux réalisés depuis fin 2018.

Notre réseau de neurones possède un intérêt important s'il a une capacité d'apprentissage avec des données connues. Un des exemples forts de l'intérêt de cette méthode de ces derniers mois est AL-PHA Go qui a battu le meilleur joueur du jeu de GO 4-1. Quelques mois plus tard Alpha Zero, une autre version de l'IA, en utilisant une méthode algorithmique de réseaux de neurones par renforcement bat APHA GO 100-0. La puissance de ce type d'algorithmes dans des situations données permet de prouver leur puissance.

L'un des paramètres importants (au moins à l'initialisation des poids des neurones) concerne la prévision météo qui est un des tous premiers domaines scientifiques à faire usage de la simulation numérique. Au début des années 1900 les travaux du mathématicien, météorologue Lewis Fry Richardson (REF), posent les bases de la prévision et il décrit ce que pourrait être une forecast factory (usine à prévision) météorologique. A la fin des années 1940 la première prévision numérique du temps par une équipe pilotée par Jule Charney et John Von Neumann en 1950. Cette prévision numérique est présentée par Lynch, P. (2008) The ENIAC forecasts [27].

Depuis, des systèmes de prévision ont été développés progressivement pour toutes les sciences environnementales:

Météorologie, (27. Hello, G. (2002). Prise en compte de la dynamique associée aux dépressions des latitudes moyennes dans la détermination des conditions initiales des modèles météorologiques) [28]

Océanographie, (Monbet, V. (2009). Quelques apports à la modélisation stochastique en océanographie et météorologie) [29]

Pollution de l'air à partir de mesurages effectués dans deux sites urbains différents des relations statistiques ont été établies avec différents paramètres météorologiques à partir des valeurs quotidiennes (Masoudi, M., & Asadifard, E. (2015)) [30]

Hydraulogie fluviale (Thèse sur une Application à la prévision hydrologique sur les grands bassins fluviaux de la Saône et de la Seine) Daoud, A. B. (2010) [31]

Les systèmes de prévision météorologique sont basés sur la connaissance physique des phénomènes mis en jeu. Ces lois physiques, qui traduisent en général des principes simples comme la conservation de la masse ou de l'énergie, sont exprimées sous forme d'équations mathématiques. Beaucoup sont d'ailleurs connues depuis les travaux d'Euler sur les fonctions (1734) , la transformation de Fourier (année) présenté par Gold, B., & Rader, C. M. (1983) [32], la présentation de la perturbation de la force de Coriolis par Lackner, J. R., & Dizio, P. (1994) [33] ou l'approximation Boussinesq (1877) il y a deux siècles par par Joseph Boussinesq.

Ces équations mathématiques permettent de mettre en jeu la vitesse, la pression, la température. Toutefois le contrôle de ces trois variables se révèle trop complexe pour que l'équation soit résolue exactement. Dès lors, il convient de construire un modèle numérique afin de les résoudre de façon approchée. Autrement dit, on remplace l'équation mathématique exacte par une expression approchée calculable par un ordinateur. Comme le précise Niels Bohr Prix créateur de la théorie quantique (1913) [34] et Nobel 1922 disait: «Prévoir est très difficile, surtout lorsque cela concerne l'avenir » (p.xx). Ainsi notre modèle numérique ne nous sera que de peu d'utilité pour prédire la situation de demain s'il n'est pas bien « calé », c'est-à-dire renseigné aussi précisément que possible sur la situation d'aujourd'hui et celle d'hier.

Concrètement, il faut fixer les valeurs de toutes les variables du modèle à un instant donné, et donc utiliser pour cela les observations disponibles au même moment. Des techniques mathématiques sophistiquées ont été développées pour cela, qu'on englobe sous le terme « d'assimilation de données ». Cette approche vise à estimer des variables d'un modèle en combinant de façon optimale des valeurs a priori et des observations, tout en prenant en compte notre connaissance sur l'incertitude de ces informations. On se rapproche des modèles en logique flous présentés par exemple par Bühler, H. (1994). Réglage par logique floue [35] mais en les intégrant comme point d'entrée dans notre réseau de neurones. Cette approche a aussi été développée à par Yann Michel [2010] Météo France [36].

Le premier point que nous aborderons est la collecte des données. Cette étape est essentielle dans météo-biz. Dans cette étude les données servant à alimenter le réseau de neurones sont de plusieurs sources.

Ces "sources" alimentent un réseau de neurones.

Notre modèle en plus d'intégrer des données météo, en prévision complète sur une zone, intégre les données liées à l'activité économique. C'est à dire que nous intégrons aussi bien des données liées à l'arrivée et au départ des avions, des bateaux de croisières, les grèves (notamment transport public) et grèves par secteurs (codes APE), les mouvements sociaux..etc. L'agréagation de ces éléments constitue la base des entrées de notre modèle. La météo est un élément important mais n'est pas le seul qui influe sur la prédiction d'activité de l'entreprise.

À ce stade, on dispose donc d'un système qui, utilisant les données sur l'état actuel, est visé à prédire un état futur avec le minimum d'incertitude. L'incertitude de la prévision est évaluée généralement par des techniques de « prévisions d'ensemble », consistant à réaliser de nombreuses prévisions en perturbant légèrement à chaque fois un paramètre du système, comme son état initial. Si les prévisions demeurent très cohérentes entre elles, alors on dira que la prévision est fiable; si au contraire elles présentent des disparités fortes, on considérera que la prévision est incertaine.

La premiere version du réseau de neurones est constitué des éléments suivants :

- Activité de l'entreprise
    - Code APE
    - Tranche de Salariés
    - Dans une zone d'activité
- Position géographique
    - Région
    - Département
    - Commune
    - Sous section
    - Adresse
    - Position GPS
- Météo
    - Vent
        - Force
        - Direction
    - Visibilité (km)
    - Portée visuelle sur piste
    - Détail du temps
    - Nuages
    - Température
    - Point de rosée
    - Pression barométrique
- Evenements internes entreprise
    - Bénéficiare PGE
    - Greves
- Evenement exogenes à l'entreprise
    - Greves
        - Transport
        - Service public
        - Service Privé
    - Disponibilité électrique

- Disponibilité Eau
- Disponibilité internet

6.1.1 Le reseau de neurones de météo-biz

Il est possible de schématiser de façon simplifié le reseau de neurones avec la fig 23.

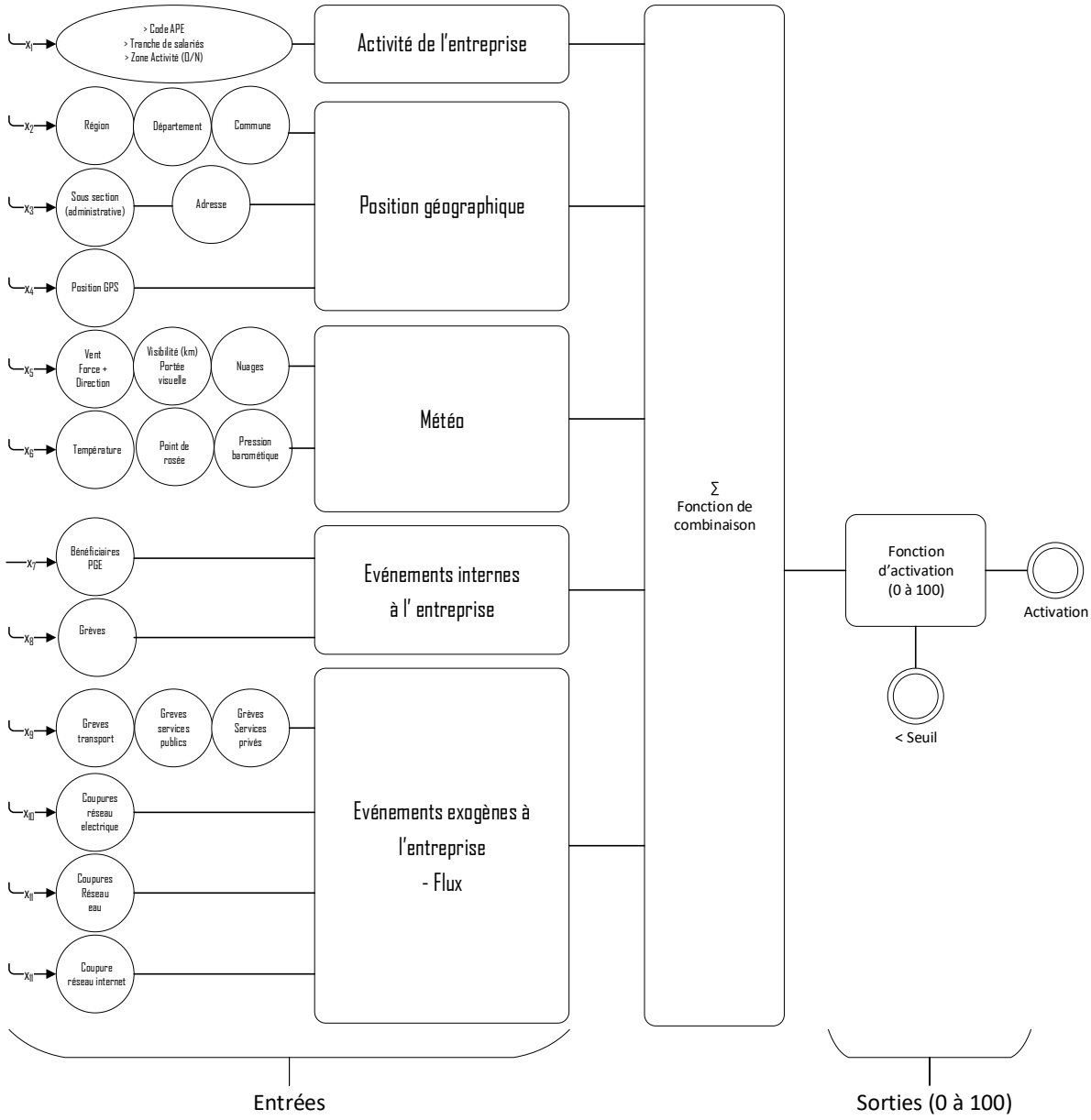


Fig 23. Modèle réseau de neurones de prévision d'activité économique

Conclusion

Mon travail avait pour objectif de trouver comment mêler science et informatique pour la résolution efficace des problèmes dans plusieurs domaines différents. Dans le domaine de la géolocalisation touristique, une approche a été mise sur pied grâce aux algorithmes k-means qui permettent de créer des groupes de comportements. Dans la recommandation, nous prenons en compte de façon très fine la notion de positionnement géographique, mais aussi la donnée temporelle. Nous apportons une solution à la question de la modélisation des déplacements touristiques en intégrant le lieu et le temps.

Dans le domaine médical, il était question de présenter des outils pour l'amélioration de la connaissance du client au sens générique, de la prédiction de ses comportements et de l'optimisation de l'offre dans un hôpital. Nous avons tenté de mettre en évidence l'importance de la fouille de données, avons identifié des outils, qui proposé aux spécialistes des données du domaine permettent d'optimiser l'offre de soins. Cette recherche a eu des limites car plusieurs données n'étant pas prises en compte. Il conviendra de proposer des outils qui mettront l'accent sur la valorisation des séjours et l'incidence financière du parcours patient en se basant notamment sur les algorithmes de groupage utilisés par l'ATIH. D'autres outils mettant l'accent sur l'optimisation des dépenses seront également étudiés (consommation médicamenteuses…).

Quant au domaine financier, nous avons développé le module Hemdall, qui permet de définir une transaction atypique ou non en tenant compte de plusieurs variables, le but étant de lutter contre les transactions financières illicites. Ce module est utilisé en 2020 dans une entreprise de paiement.

Enfin, nous avons mis sur pied le projet meteobiz qui vise à prédire l'activité économique en utilisant notamment les données météorologiques. À ce niveau, nous disposons donc d'un système qui prédit un état futur avec le minimum d'incertitudes possibles étant donné l'avancée des travaux. Météobiz fera l'objet d'une thèse avec bourse Cifre de 2020 à 2023. L'objectif est de partir des travaux actuels et d'améliorer et d'optimiser le moteur, de proposer de nouvelles orientations.

Au vu de ces quatre expériences « de terrain » je me suis attelé a proposer des solutions scientifiques en utilisant des outils informatiques. Ces travaux permettent dans le cas du

PMSI et de Hemdall, d'améliorer l'existant et dans les cas géotourisme et de meteobiz d'avoir un nouveau champ d'utilisation des algorithmes. Ces travaux m'ont permis de formaliser un certain nombre d'idées dans un cadre formel et encadré.

Ce travail de recherche m'a aussi permis d'intégrer la réflexion du chercheur au travail de l'ingénieur. La mutation des systèmes informatiques que l'on utilise ou que l'on construit demande de plus en plus une vision de l'analyse des données pour leur fonctionnement en amont ou a posteriori.

Cette thèse que je voulais professionnelle s'appuie sur les données que je manipule au quotidien dans mon environnement professionnel.

L'Université des Antilles pourrait promouvoir ce type de travaux axés sur des données endogènes ; l'objectif étant de déceler des talents en informatique et de promouvoir cette science qui attire de plus en plus de jeunes.

Cette thèse m'a permis de formaliser cette analyse tout en suivant une partie des travaux et publications dont le nombre et la complexité sont croissants.

Les travaux abordés dans le projet Météo-Biz ont permis d'entamer une relation avec la recherche académique par la proposition d'un travail de thèse CIFRE. Cette relation avec l'Institut des Mines Télécom (Mines d'Alès) a été validée par l'ANRT qui a validé la poursuite des travaux par une doctorante qui devrait démarrer lors du premier semestre 2021.

Enfin, je participe à la création de Digitribe, premier Living Lab privé des Antilles. Ce projet soutenu par le maire de la capitale vise à regrouper des chercheurs, des professionnels et des particuliers afin réfléchir et de trouver des solutions à des problématiques locales.

Pour aller au-delà de cette conclusion qui a repris les éléments importants de ce mémoire je souhaite redire que ce travail a été motivé par à la rencontre des personnes que je remercie à nouveau, mais surtout par le besoin d'inscrire mon travail quotidien dans une réflexion indispensable au développement d'une TPE dans le tissu économique et territorial de la Martinique.

Ainsi tout au long de ce travail de doctorat j'ai pris la mesure d'une démarche qui dépasse la rédaction de ce mémoire. C'est ici la première conclusion personnelle et très positive que je veux souligner. Bien sûr, j'ai acquis et renforcé mes connaissances dans le domaine

de l'informatique et particulièrement dans le domaine de la fouille de données, mais au-delà de cet aspect attendu j'ai surtout compris que toute recherche même appliquée doit avoir une dimension prospective intégrant l'intentionnalité. C'est ainsi que j'ai pris conscience que ce travail dépasse la finalité initiale de développer des algorithmes et j'ai également réaliser que la mutation des systèmes informatiques que l'on utilise ou que l'on construit, demande de plus en plus une vision de l'analyse des données en amont ou a posteriori. Même si cette finalité donne un sens au travail de recherche, ce qui est important pour moi c'est d'avoir compris que mes choix tant théoriques que pratiques ont structuré ma vision de la recherche et du terrain en même temps qu'ils donnaient un périmètre et des perspectives à mon travail. C'est donc plus une méthode, une façon d'appréhender les problèmes de terrain que j'ai apprises au cours de ces trois ans.

La seconde conclusion personnelle que je souhaite partager ici relève des relations université-entreprise. Pour beaucoup de chefs d'entreprise, ces relations sont souvent illusoires ou inutiles. Les temps, les besoins, les contraintes sont différentes. La relation profite le plus souvent aux étudiants, mais rarement à l'entreprise. Or, dans mon cas, les bénéfices sont concrets et j'y ai trouvé un enrichissement croisé des compétences et des intérêts personnels. Il me semble que le Professeur Agostinelli a été amené à tenir compte des finalités opérationnelles de l'entreprise pour assurer la direction de cette thèse. Mes collaborateurs ont également compris l'importance d'un regard « scientifique » sur les activités quotidiennes et ils sont devenus particulièrement disponibles. Le bénéfice direct repose aujourd'hui sur un environnement professionnel modifié à travers ma démarche de recherche. Pour moi, cette thèse que je voulais professionnelle s'appuie sur les données que je manipule au quotidien dans mon environnement professionnel et ce travail de recherche m'a donc permis d'intégrer la réflexion du chercheur au travail de l'ingénieur. Par expérience je sais que de nombreuses TPE en informatique existent en Martinique et aux Antilles. Si beaucoup d'entre elles participent par des propositions de stages à l'insertion professionnelle des étudiants, elles ne pensent pas que la recherche et donc l'Université des Antilles pourrait promouvoir ce type de travaux axés sur des données endogènes avec pour objectif de déceler des talents en informatique et promouvoir cette science qui attire de plus en plus de jeunes.

Ma troisième conclusion personnelle porte sur les projets mis en place entre la recherche académique et mon entreprise. Le prolongement du projet Météo-Biz se concrétise aujourd'hui par une thèse CIFRE. Cette thèse avec l'Institut des Mines Télécom (Mines d'Alès) a été validée par l'ANRT et devrait démarrer lors du premier semestre 2021.

Son point de départ est qu'environ 70% des entreprises sont météo-dépendantes, c'est-à-dire que leurs résultats financiers et la satisfaction de leurs clients varient avec la météo. Cette thèse va donc contribuer à la mise au point d'algorithmes concernant la prédiction de fréquentation de lieux (en fonction de contraintes multiples) afin de faciliter la prise de décision en environnement complexe.

Enfin, je participe à la création de Digitribe, premier Living Lab privé des Antilles. Ce projet soutenu par la mairie de Fort-de-France vise à regrouper des chercheurs, des professionnels et des particuliers afin de réfléchir et trouver des solutions à des problématiques locales. Ce Living Lab sera piloté par un conseil scientifique dans lequel le président et une large majorité des membres sont des universitaires.

Bibliographie

1. Agostinelli, S. (2018). La compréhension intuitive des données : de Hal à Watson. In, C., Alcantara, F., Charest, & S., Agostinelli (eds.). BigData et visibilité en ligne (pp. 13-24). Paris : Presses des Mines
2. Kohonen, T. (1997). *Self-Organizing Maps*. Berlin: Springer.
3. Krishna, K., & Murty, M. N. (1999). *Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3), 433-439.
4. Zhang, T., Ramakrishnan, R. & Livny, M. (1996) Birch : *an efficient data clustering method for very large datables. SIGMOD 96 6/96* (103-114). Canada, Montreal. Association for Computing Machinery.
5. Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).
6. Yuan, M., Pavlidis, Y., Jain, M., & Caster, K. (2016, November). *Walmart online grocery personalization: Behavioral insights and basket recommendations. In International Conference on Conceptual Modeling* (pp. 49-64). Gifu, Japan. Springer, Cham.
7. Nedjema, Benlaiter (2012). *Extraction des regles d'association pour l'aide a la decision etude de cas: pharmacie* (Doctoral dissertation, Universite Mohamed Boudiaf M'sila: Faculte Des Mathematiques Et De L'informatique: Département d'Informatique).
8. Elisabeth, Sebastien (2014) - *Les déterminants de l'efficience des établissements de santé. XVIIIéme*. Conférence des Fédérations Hospitalières des Antilles et de la Guyane. Martinique.
9. *Journal de la Société Française de Statistique, Vol. 159 No. 3* 1-39 http://www.sfds.asso.fr/journal
10. Collard, M. (2003). *Fouille de données, Contributions Méthodologiques et Applicatives*. Doctoral dissertation, Université Nice Sophia Antipolis.
11. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11), 27-34.

12. LeCun, Y., & Bengio, Y. (1995). *Convolutional networks for images, speech, and time series*. The handbook of brain theory and neural networks (pp. 5-8), 3361(10), 1995.

13. https://www.college-de-france.fr/site/yann-lecun/course-2015-2016.htm

14. McNicholas, P. D., Murphy, T. B., & O'Regan, M. (2008). *Standardising the lift of an association rule*. Computational Statistics & Data Analysis, 52(10), 4712-4721.

15. Feno, D. R. (2007). *Mesures de qualité des règles d'association: normalisation et caractérisation des bases* (Doctoral dissertation). Université de La Réunion.

16. Guillaume, S., & Papon, P. A. (2013). *Étude comparative d'extraction de règles d'association positives et négatives et optimisations*. Université d'Auvergne. Hermann.

17. Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015). *The use of sequential pattern mining to predict next prescribed medications*. Journal of biomedical informatics, 53, 73-80.

18. Jensen, S., & SPSS, U. (2001, September). *Mining medical data for predictive and sequential patterns: PKDD 2001*. In Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases. 5th PKDD 2001: Freiburg, Germany.

19. Garofalakis, M. N., Rastogi, R., & Shim, K. (1999, September). *SPIRIT: Sequential pattern mining with regular expression constraints*. In VLDB (Vol. 99, pp. 7-10).

20. Masseglia, F. et Teisseire M. et Poncelet P. (2004) *Extraction de motifs séquentiels, Problèmes et méthodes*. Série ISI: Ingénierie des Systèmes d'information, Lavoisier, 2004, 9 (3/4), pp.183-210. INRIA Sophia Antipolis.

21. Lucchese, L., & Mitra, S. K. (2001). *Colour image segmentation: a state-of-the-art survey*. Proceedings-Indian National Science Academy Part A, 67(2), 207-222.

22. Elisabeth, E., Richard, N., and Célimene, F. "*Demonstrator of a tourist recommendation system*." International Conference on Big Data Analytics. Springer, Cham, 2013.

23. Elisabeth E. « *Fouille de données spatio-temporelle : Application à un système de modélisation des déplacements touristiques* » (pp. 5-15). Document numérique et Société - 5e conference, (Mai 2015). Maroc

24. Elisabeth E., Sébastien N. (2017). *Les déterminants de l'efficience des établissements de santé : l'apport de la fouille de données, Colloque Big Data et visibilité en ligne* - Fort de France, Martinique.

25. Nock, R., & Nielsen, F. (2013). *Information-geometric lenses for multiple foci+ contexts interfaces*. In SIGGRAPH Asia 2013 Technical Briefs (pp. 1-4).

26. Nock, Richard, et al. *"Staring at economic aggregators through information lenses."* arXiv preprint arXiv:0801.0390 (2008).

27. Lynch, P. (2008). *The ENIAC forecasts: A re-creation. Bulletin of the American Meteorological Society*, 89(1), 45-56.

28. Hello, G. (2002). *Prise en compte de la dynamique associée aux dépressions des latitudes moyennes dans la détermination des conditions initiales des modèles météorologiques* (Doctoral dissertation, Toulouse 3).

29. Monbet, V. (2009). *Quelques apports à la modélisation stochastique en océanographie et météorologie*. Université de Bretagne SudLab-STICC - UMR 3192.

30. Masoudi, M., & Asadifard, E. (2015). *Status and prediction of Nitrogen Dioxide as an air pollutant in Ahvaz City*, Iran. 2268-3798.

31. Daoud, A. B. (2010). *Améliorations et développements d'une méthode de prévision probabiliste des pluies par analogie. Application à la prévision hydrologique sur les grands bassins fluviaux de la Saône et de la Seine* (Doctoral dissertation, Université de Grenoble).

32. Gold, B., & Rader, C. M. (1983). *Digital processing of signals*. Krieger Publishing Co., Inc.., Krieger.

33. Lackner, J. R., & Dizio, P. (1994). *Rapid adaptation to Coriolis force perturbations of arm trajectory. Journal of neurophysiology*, 72(1), 299-313.

34. Bohr, N. (1913). XXXVII*. On the constitution of atoms and molecules*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 26(153), 476-502.

35. Bühler, H. (1994). *Réglage par logique floue* (No. BOOK). Presses polytechniques et universitaires romandes.

36. Yann Michel (2010]. *Techniques Mathématiques en Assimilation de Données pour la Prévision Météorologique*. http://www.umr-cnrm.fr/recyf/IMG/pdf/journee_maths_INPT_YM.pdf

37. Jean-Pierre, C. (1983). *L'homme neuronal*. Paris, Fayard.

38. Mehdi Mouhadil (23 Fév 2018) *De l'humain au deep learning*. (meritis.fr/ia/deep-learning/)

39. Roubens, M. (1980). *Analyse et agrégation des préférences: modélisation, ajustement et résumé de données relationnelles*. JORBEL-Belgian Journal of Operations Research, Statistics, and Computer Science, 20(2), 36-67.

40. Bentayeb, F., Boussaid, O., Favre, C., Ravat, F., & Teste, O. (2009). *Personnalisation dans les entrepôts de données: bilan et perspectives*. In EDA (pp. 7-22).

41. ATIH (Agence de Traitement de l'information hospitalière), *Guide méthodologique de production des informations relatives à l'activité médicale et à sa facturation en médecine, chirurgie, obstétrique et odontologie* (http://www.atih.sante.fr/index.php?id=0002300005FF)

42. Nock, R. (1998). *Apprentissage de formules logiques de taille limitée: aspects théoriques, méthodes et résultats* (Doctoral dissertation, Montpellier 2).

43. Pasquier N. et Lakhal L. (2000) Data mining: *Algorithmes d'extraction et de réduction des règles d'association dans les bases de donnes*. (Doctoral dissertation). Université Blaise Pascal - Clermont-Ferrand 2.

44. Roubens, M. (1980). *Analyse et agrégation des préférences: modélisation, ajustement et résumé de données relationnelles*. JORBEL-Belgian Journal of Operations Research, Statistics, and Computer Science, 20(2), 36-67.

45. Krishna, K., & Murty, M. N. (1999). *Genetic K-means algorithm*. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439.

46. Hello, G. (2002). *Prise en compte de la dynamique associée aux dépressions des latitudes moyennes dans la détermination des conditions initiales des modèles météorologiques* (Doctoral dissertation, Toulouse 3).

47. Antonie, M. L., & Zaïane, O. R. (2004, September). Mining positive and negative association rules: an approach for confined rules. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 27-38). Springer, Berlin, Heidelberg.

48. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. *20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

49. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, *43*(1), 59-69.

50. Kohonen, T., & Honkela, T. (2007). Kohonen network. *Scholarpedia*, *2*(1), 1568.

51. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., ... & Hsu, M. C. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on knowledge and data engineering*, *16*(11), 1424-1440.

52. Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine learning*, *42*(1-2), 31-60.

53. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. *Advances in knowledge discovery and data mining*, *12*(1), 307-328.

54. Tarby, J. C. (1993). *Gestion automatique du dialogue homme-machine à partir de spécifications conceptuelles* (Doctoral dissertation).

ANNEXES

- [publication]
  Elisabeth, E., Nock, R., & Célimene, F. (2013, December). Demonstrator of a tourist recommendation system. In International Conference on Big Data Analytics (pp. 171-175). Springer, Cham.
- [Colloque]
  Deveraux, G. Elisabeth, E. Sébastien, N. (2014) Les déterminants de l'efficience des établissements de santé : l 'apport de la fouille de données- Congrès FHM Fédération hospitalière de France.
- [publication]
  Elisabeth, E. (2015). Fouille de données spatio-temporelle : Application à un système de modélisation des déplacements touristiques Document numérique et Société - 5e conférence
- [publication]
  Elisabeth, E. Sébastien, N. (2018). Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie : l'apport de la fouille de données. BigData et visibilité en ligne (pp. 51-58). Paris : Presses des Mines.

ANNEXE 1

- [publication]
  Elisabeth, E., Nock, R., & Célimene, F. (2013, December). Demonstrator of a tourist recommendation system. In International Conference on Big Data Analytics (pp. 171-175). Springer, Cham.

ANNEXE 2

- [Colloque]
  Deveraux, G. Elisabeth, E. Sébastien, N. (2014) Les déterminants de l'efficience des établissements de santé : l 'apport de la fouille de données- Congrès FHM Fédération hospitalière de France.

ANNEXE 3

- [publication]
  Elisabeth, E. (2015). Fouille de données spatio-temporelle : Application à un système de modélisation des déplacements touristiques Document numérique et Société - 5e conférence

ANNEXE 4

- [publication]
  Elisabeth, E. Sébastien, N. (2018). Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie : l'apport de la fouille de données. BigData et visibilité en ligne (pp. 51-58). Paris : Presses des Mines.

ANNEXE 1

- [publication]
  Elisabeth, E., Nock, R., & Célimene, F. (2013, December). Demonstrator of a tourist recommendation system. In International Conference on Big Data Analytics (pp. 171-175). Springer, Cham.

Demonstrator of a tourist recommendation system

Erol Elisabeth, Richard Nock, and Fred Célimène

CEREGMIA - Centre d'Etude et de Recherche en Economie, Gestion, Modélisation et Informatique Appliquée,
Campus de Schoelcher, B.P. 7209 97275 Schoelcher Cédex, Martinique
erol@beepway.com,
Richard.Nock@martinique.univ-ag.fr,
Fred.Celimene@martinique.univ-ag.fr,
http://www.ceregmia.eu

Abstract. This paper proposes a way of using data collected from tracking gps installed in rental tourist cars. Data has been collected during more than one year. The gps positions are lined to the gps positions of the tourist sites (restaurants, beaches, museums ...). [9] These links are presented as a summary of the data. This summary is used to run specific versions of machine learning algorithms because of their geo-graphical dimension. This experiment shows how gps summaries of data can be used to extract relationships between stops of a car and touristic places.

Keywords: gps, association rules, sequential patterns, k-means,Q patterns, Geographical Center of Sequential patterns

1 Introduction

In this paper, we begin with data summaries and we use 5 types of data mining algorithms to process these summaries: association rules, sequential patterns, Q patterns, geographical center of sequential patterns and k-Means.

The aim is to provide the best recommendation for another tourist site for a tourist in his car.

The device used in the car is a tracking gps with a PND (Personal Navigation Device). With this PND it is possible to send in real time a recommendation to the tourist, and if he accepts the recommendation, the system shows him/her the best way to join this place as any gps navigation system.

2 Data Summaries

In our demonstration we have 852 dierent data summaries. These are the activities of 12 cars during more than 14 months in 2008/2009. A path is a succession of stops between the first stop of the first day in the car and the return of the car to the park. The rows are a gps position (date, time, latitude, longitude, instant speed, altitude, cap, status of the car [stopped, running]). These rows are

used to produce a table of gps stops. After resuming, the positions collected are organized in sequential rows [date, gps stop position of, duration]. We associate the stop with the gps position of the tourist place. The result is a synthetic table with the succession of the visited tourist sites. Each row is presented as follow :

[sequential number] {date/time, tourist site, duration of visit}
[1] {date/time,1 'Casino Bateliere Plazza', 1:55}
{date/time,2 'Casino Bateliere Pointe du Bout', 3:15}
[2] {date/time,1 'Habitation Clement', 0:53}
{date/time,1 'Casino Bateliere Pointe du Bout', 5:15}

3 Association rules

Association rule mining [2] is defined as : $I = i_1, i_2, ..., i_n$ as a set of $_n$ binary attributes called items. $D = t_1, t_2, ..., t_n$ a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$where $X, Y \subseteq IandX \cap Y = \oslash$. The conviction of a rule [3] is defined $conv(X \Rightarrow Y)$ With association rules algorithm, it is possible to know the relationship between 2 sites. In the screen display below we can see 33,3% of tourists who went to Casino Batelire Plazza went also to Casino de laa Pointe du Bout ; but 23,1% of tourists who went to 2 : Casino de la Pointe du Bout went also to 1 : Casino Batelire Plazza. With the proposed map we can see that the behaviors of tourists from south of the island are dierent from the one from the center.

Support X U X` :	Total AR X X`	Total AR X` X	AR X X`	AR X` X
1 2 = 3 => Support : 0.023	{1 }=>{2} = 0.333	{2 }=>{1} = 0.231	{1 }=>{2} = 0.333	{2 }=>{1} = 0.231
1 3 = 2 => Support : 0.015	{1 }=>{3} = 0.222	{3 }=>{1} = 0.069	{1 }=>{3} = 0.222	{3 }=>{1} = 0.069
1 15 = 1 => Support : 0.008	{1 }=>{15} = 0.111	{15 }=>{1} = 0.5		
1 16 = 1 => Support : 0.008	{1 }=>{16} = 0.111	{16 }=>{1} = 0.25		
1 23 = 3 => Support : 0.023	{1 }=>{23} = 0.333	{23 }=>{1} = 0.111	{1 }=>{23} = 0.333	{23 }=>{1} = 0.111

Casino Batelière Plazza

Casino de la Pointe du Bout

Fig. 1. Association rule between 2 sites

4 Sequential patterns

Extraction of sequential patterns [1] and [4] make possible the discovery of temporal relations between 2 sites.In this sequence we notice a relationship between

52 : 'ART et Nature' AND 257 : 'Hotel le Panoramique'. We may suppose that before coming back to the 'hotel Le Panoramique' the tourist went to 'Art et Nature'. We propose two types of representations. The geographical map of pattern allows to have a visual representation of behaviors of tourist stops. It is possible to create an oriented diagram that shows the inter site links.

(52) ART et NATURE	=>	(257) Hotel Le Panoramique **	9	0.0124826629681
(1) Casino Batelière Plazza	=>	(23) BIBLIOTHEQUE SCHOELCHER	6	0.0083217753120 7
(1) Casino Batelière Plazza	=>	(3) La Galleria	2	0.0027739251040 2
(1) Casino Batelière Plazza	=>	(80) HERTZ	4	0.0055478502080 4
(1) Casino Batelière Plazza	=>	(2) Casino de la Pointe du Bout	2	0.0027739251040 2
(188) TOTAL	=>	(257) Hotel Le Panoramique **	2	0.0027739251040 2
(2) Casino de la Pointe du Bout	=>	(80) HERTZ	4	0.0055478502080 4
(2) Casino de la Pointe du Bout	=>	(134) Anse Noire	2	0.0027739251040 2
(2) Casino de la Pointe du Bout	=>	(1) Casino Batelière Plazza	2	0.0027739251040 2
(2) Casino de la Pointe du Bout	=>	(267) Hotel Pagerie ***	2	0.0027739251040 2
(2) Casino de la Pointe du Bout	=>	(137) Anse Mitan	2	0.0027739251040 2
(3) La Galleria	=>	(80) HERTZ	7	0.0097087378640 8
(3) La Galleria	=>	(23) BIBLIOTHEQUE SCHOELCHER	7	0.0097087378640 8
(3) La Galleria	=>	(280) Hotel Valmeniere***	5	0.0069348127600 6

Fig. 2. sequential pattern between sites

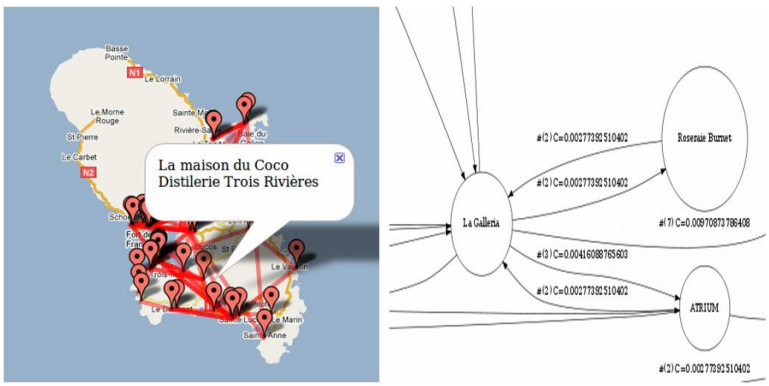


Fig. 3. Map of sequential pattern Diagram inter site links

4.1 Q patterns

The Q patterns are like patterns but the item sets do not have a fixed dimension [5] and [8]. For example item sets are in the database $conv(A, B \Rightarrow C)$ and in the same database we can also have $conv(A, B, D, F \Rightarrow T)$

23 : Bibliothque Schoecher , 217 : La Kasa Saveurs, 298 : Karibea Rsidence La Goelette, 12 : Habitation Clment, 39 : Habitation Depaz, 40 : Distillerie Neisson, 130 : Grande Anse d Arlet

$(23, 217, 298, 12, 298, 39, 40, 298 \Rightarrow 130)$ We can have a representation where each pattern has a specific color on an oriented graph. This sequential pattern shows a succession of stops with the same topic : rum distillery. We can have a representation where each pattern has a specific color on an oriented graph. Each Bubble is a tourist site.

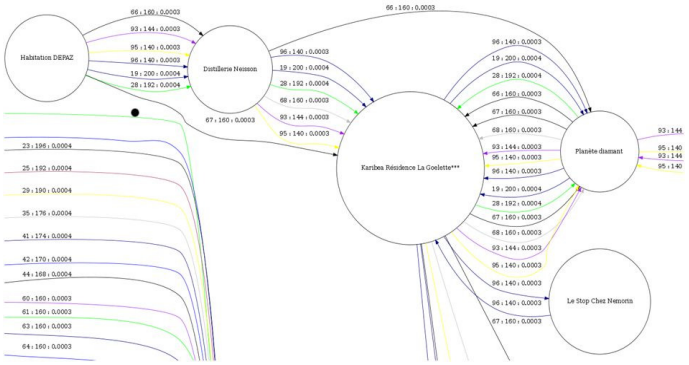


Fig. 4. Part of Q pattern

4.2 Geographical Center of Sequential patterns

We also compute a geographical solution as a recommendation when we already have found the sequential pattern or the cluster in a specific k-means. The objective is to find -in real time- the best tourist site next to a car when we already classify it in a k-means cluster or a sequential pattern [6] and [7]. In this example if the car is IN the cluster (12, 107) or in a sequential pattern where there are item sets (12 AND 107) and if the car is next to the centroid of this data, we can propose a new activity.

4.3 k-Means

We can have a k-means representation using 2 clusters. The cluster A (in red) is around the Center of the island and the south, the second one B (in yellow) is in the south.

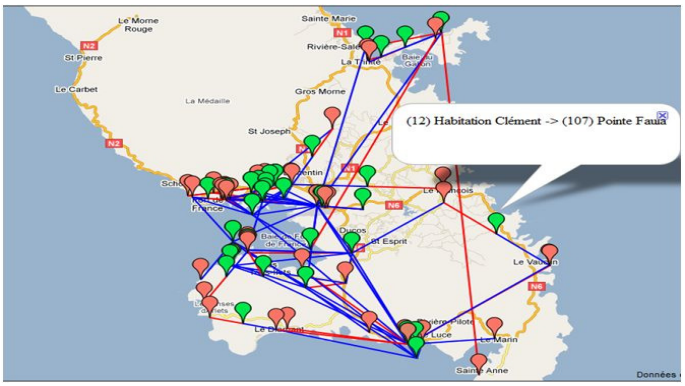


Fig. 5. Map of centroid of sequential patterns

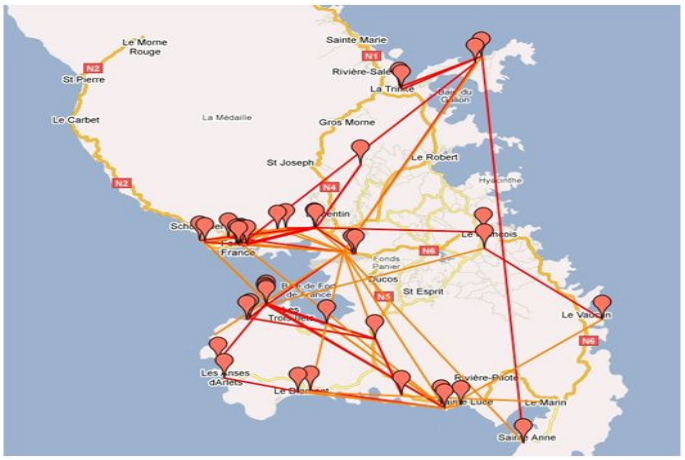


Fig. 6. K means (3 clusters)

5 Conclusion

In this paper, we have applied a set of data mining algorithms using data collected from tracking gps installed in rental tourist cars. Tourist organizations and agencies could look into these applications to find the best way to extract knowledge from their own database systems. GPS tracking companies can also find ideas to improve the uses of their collected data.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In Philip S. Yu and Arbee L. P. Chen : Proceedings of the Eleventh International Conference on Data Engineering, March 6-10 (1995)
2. Heikki Mannila and Hannu Toivonen : Multiple uses of frequent sets and condensed representations (extended abstract) (1996)
3. Stephane Lallich et Olivier Teytaud : Evaluation et validation de l'interet des regles d'association (2000)
4. Pasquier Nicolas et Lakhal Lotfi :Data mining : Algorithmes d'extraction et de rduction des regles d'association dans les bases de donnes (2000)
5. Mohammed Javeed Zaki. : Spade : An efficient algorithm for mining frequent sequences. :Machine Learning, 42(1/2) : 3160 (2001)
6. F. Masseglia et M. Teisseire et P. Poncelet , Extraction de motifs sequentiels, Problemes et methodes (2004)
7. J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. : Mining sequential patterns by pattern-growth : The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering (2004)
8. Nistor Grozavu, Youns Bennani : Classification collaborative non supervisee LPN UMP CNRS, 249-264 CAP 2010
9. Nicolas Bchet, Marie-Aude Aufaure, Yves Lechevallier Construction et peuplement de structures hirarchiques de concepts dans le domaine du e-tourisme : INRIA - 475-506, IFIA 2011

ANNEXE 2

- [Colloque]
  Deveraux, G. Elisabeth, E. Sébastien, N. (2014) Les déterminants de l'efficience des établissements de santé : l 'apport de la fouille de données-Congrès FHM Fédération hospitalière de France.

Les déterminants de l'efficience des établissements de santé : l 'apport de la fouille de données

Charles DEVEREAUX, Docteur en médecine
Erol ELISABETH, Ingénieur en Informatique, Doctorant en Informatique
Nathalie SEBASTIEN, Directeur adjoint au CHUM

XVIII^éme^ Conférence des Fédérations Hospitalières des Antilles et de la Guyane
Martinique, 23, 24, 25 Avril 2014

Sommaire

- Le PMSI, clé de voute de l'analyse médico-économique au sein de l'hôpital
- Pour des usages plus poussés du PMSI : l'apport de la fouille de données
- Application à la prise en charge du patient âgé
- Orientations et pistes de recherche

Le PMSI, clé de voute de l'analyse médico-économique au sein de l'hôpital (1/2)

- La mise en place du PMSI (Programme de médicalisation des Systèmes d'information) s'inscrit dans une **démarche de description de l'activité médicale des établissements publics et privés**
  - **à usage interne** (établissements de santé) et
  - **à usage externe** (Assurance-maladie, Services de l'Etat) en vue d'une optimisation de l'organisation de l'offre de soins
- Une **base de données médico-administrative** constituée de l'ensemble des données des établissements de santé publics et privés, remontées mensuellement, selon un **format normé et standardisé** (RSS) **retraçant le séjour hospitalier du patient** (motif de prise en charge, unité médicale de prise en charge, durée du séjour, etc.)

Le PMSI, clé de voute de l'analyse médico-économique au sein de l'hôpital (2/2)

- Une base de données médico-administrative, parmi les plus grandes au monde **insuffisamment utilisées à ce jour au regard de la puissance d'analyse qu'elle autorise**
  - Une **connaissance possible de la cible, *patient par patient (dans le respect de l'anonymat)***, à l'échelle d'un établissement, d'un territoire de santé…
  - Possibilité d'**analyses diachroniques et synchroniques**
- Un **volume de données important** qui pose la question de l'**exploitation optimale des données**
  - Volume annuel de données :
    - 1,2 milliards de feuille de soins
    - 500 millions d'actes médicaux
    - Environ 20 millions de séjours hospitaliers (MCO)

Pour des usages plus poussés du PMSI : l'apport de la fouille de données (1/3)

- La **fouille de données**, une des composantes du CRM analytique (Customer Relationship Management ou Gestion de la Relation Client), se définit comme le **processus d'extraction à partir de données de connaissances intéressantes, non triviales, implicites, inconnues et potentiellement utiles** (Fayad et al., 1996)
- Une discipline qui diffère de la statistique…
  - Une ***démarche sans « a priori »***
  - Pour faire ***émerger, à partir des données, des inférences dont il s'agit d'évaluer la qualité***
  - A partir de modèles, réalisation d'analyses de prédiction (***« apprentissage non supervisé »***)

Pour des usages plus poussés du PMSI : l'apport de la fouille de données (2/3)

- La **problématique**, au regard du volume de données traitées…
  - **produire des recommandations de qualité tout en minimisant l'effort** (*les temps de calculs*) **requis,** dans un apprentissage non supervisé,
  - **minimiser les temps humains nécessaires et les biais introduits par l'utilisation de connaissances *a priori***
- **L'objectif central des travaux de recherche présentés…**
  - **Détermination de modèles, établis à partir de « clusters »** (regroupement d'items similaires dans des classe de données représentatives), **au service de l'*amélioration de la connaissance du patient*** au sens générique**, de la *prédiction de ses comportements* et de l'*optimisation de l'offre de soins proposée***
  - **Création de savoir à destination des spécialistes du domaine de données** (hospitaliers, tutelles, chercheurs…)

Pour des usages plus poussés du PMSI : l'apport de la fouille de données (3/3)

- **Méthodologie de Recherche**
  - **Recherche bibliographique à partir de 1994**
  - **Base de données PMSI MCO pour la période 2007-2012** (soit plus de 132 millions d'enregistrements), ***après avis favorable de la CNIL et de l'ATIH***
  - **Développement d'algorithmes d'analyse et de représentation des données avec un focus sur « l'utilisabilité » des environnements de fouilles de données par les utilisateurs spécialistes du domaine des données**
  - **Deux pré-réquis :**
    - Utilisation des connaissances du domaine dans l'ensemble du processus de fouille
    - Amélioration de la « compréhensibilité » et de la confiance de l'utilisateur dans le modèle chemin faisant

Prise en charge du patient âgé : Apport de la fouille de données

- **Focus sur la pris en charge du patient âgé sur les Antilles et la Guyane**
  - **Analyse concurrentielle** reposant sur le profil d'activité des établissements et leur positionnement sur le marché (zones de recrutement, intensité concurrentielle, parts de marché, …)
  - **Analyse du parcours patient** dans une logique de médecine de parcours
    - À l'échelle du territoire
    - A l'échelle de l'établissement

Prise en charge du patient âgé : le cas de la GUADELOUPE

- **Complication de prise en charge**

CEREGMIA - PRED - SS (Chaire PMSI) | Déconnexion (SEBASTIEN Nathalie)

Nombre moyen de DAS associés à un DP

2012

Guadeloupe

Tous les établissements

Age nimimum : 75

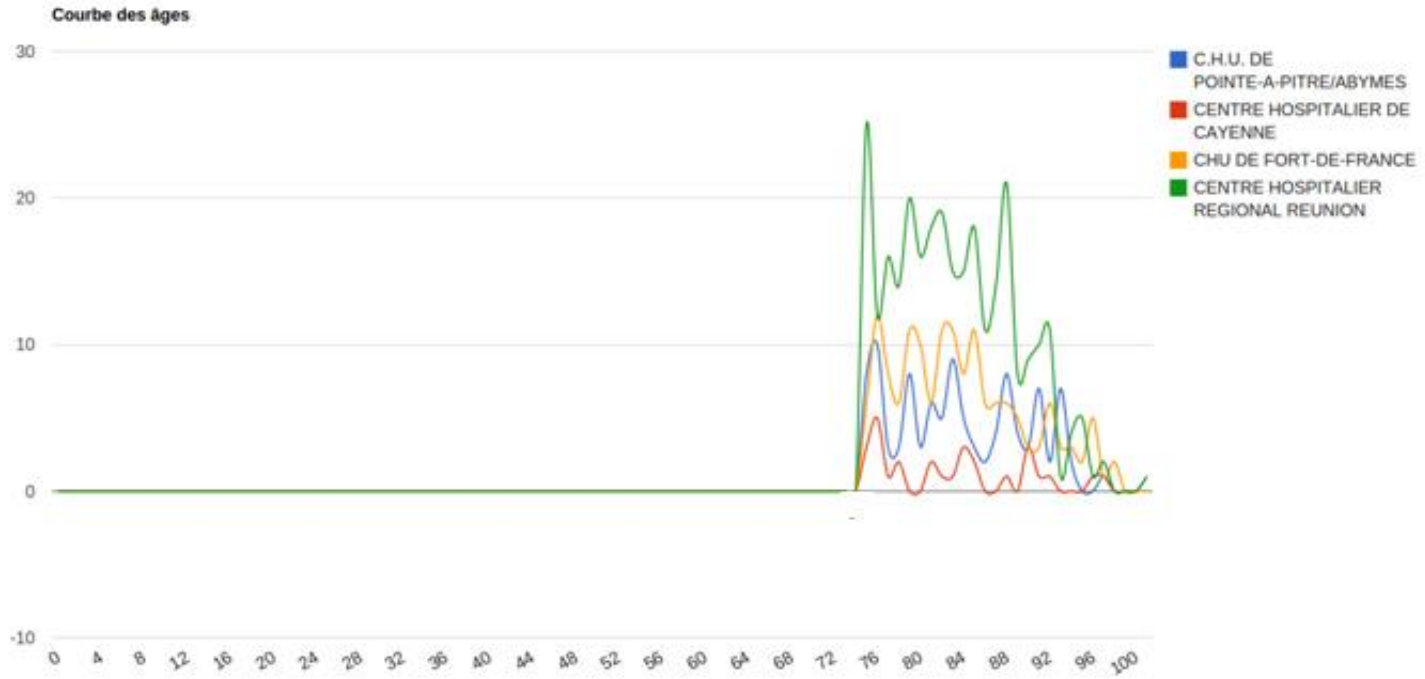
Age maximum : 130

01 Affections du système nerveux D-0108 Maladies dégénératives du système nerveux

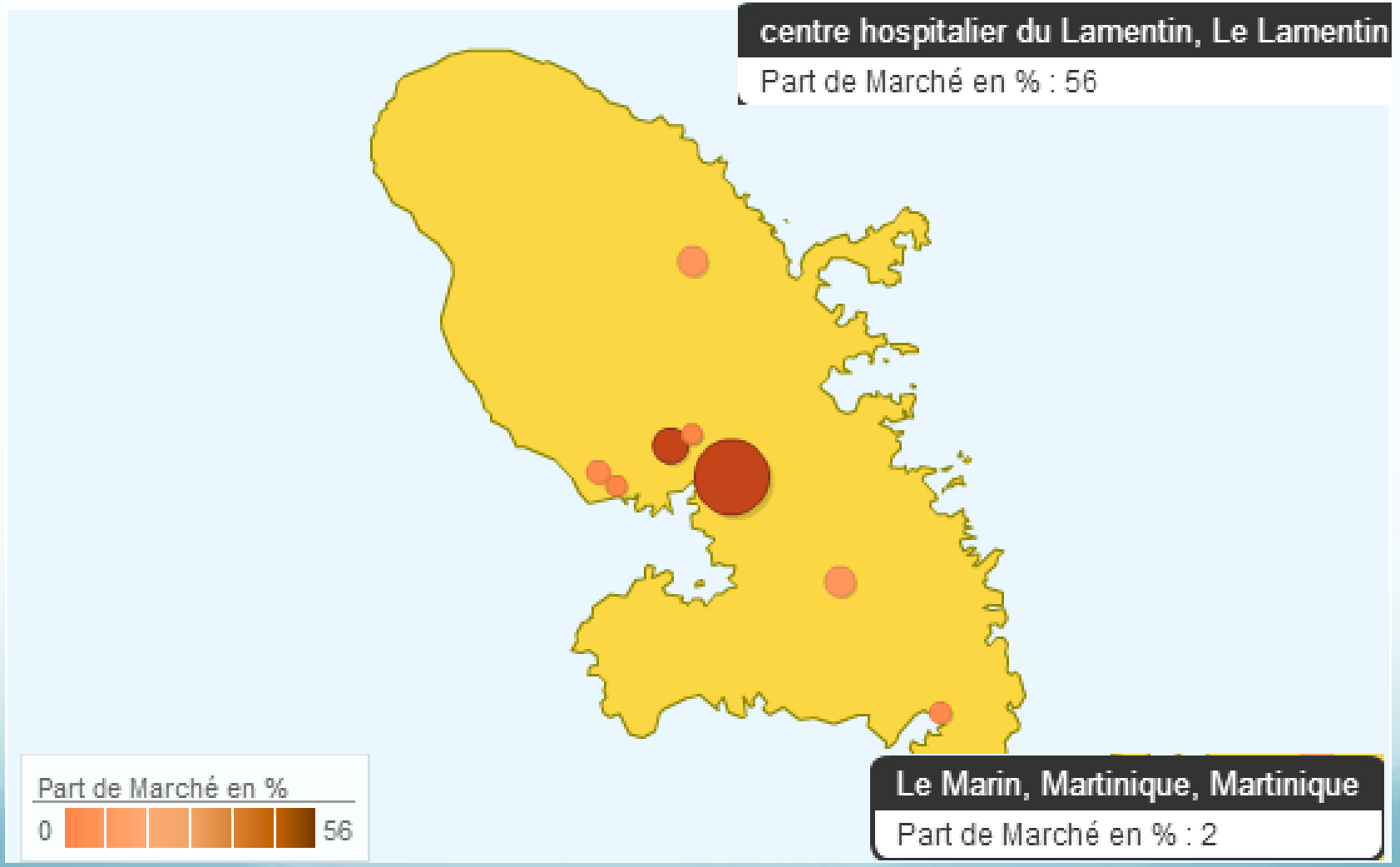
DP	POLYCLINIQUE DE LA GUADELOUPE	CENTRE MEDICO-SOCIAL	LES NOUVELLES EAUX VIVES	POLYCLINIQUE SAINT-CHRISTOPHE	HÔPITAL LOCAL IRÉNÉE DE BRUYN	CENTRE HOSPITALIER DE LA BASSE TERRE	CENTRE HOSPITALIER LOUIS CONSTANT FLEMING	CH BEAUPERTHUY	CENTRE HOSPITALIER SAINTE-MARIE	C.H.U. DE POINTE-A-PITRE/ABYMES	HÔPITAL. LOCAL DE CAPESTERRE-BELLE-EAU	CENTRE HOSPITALIER M.SELBONNE	CLINIQUE DE CHOISY	CLINIQUE LES NOUVELLES EAUX-MARINES	CLINIQUE LES EAUX CLAIRES	A.U.D.R.A.
G238 => Autres maladies dégénératives précisées des noyaux gris centraux	0	0	0	0	0	0	0	0	0	9.67	0	0	0	0	0	0
G811 => Hémiplégie spastique	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Moyenne	4.18	7.84	0	3.22	0	4.6	6	0	2	6.31	3.56	3	4.67	5.8	2	0

Prise en charge du patient âgé : ANALYSE COMPARATIVE ENTRE PLUSIEURS ÉTABLISSEMENTS

2012			
Guadeloupe	Guyane	Martinique	La Réunion
C.H.U. DE POINTE-A-PITRE/ABY	CENTRE HOSPITALIER DE CAYE	CHU DE FORT-DI	CENTRE HOSPIT/
Age min : 75	Age min : 75	Age min : 75	Age min : 75
Age max : 112	Age max : 106	Age max : 106	Age max : 104
I500 (104) (2.34%) Insuffisance c	I500 (28) (2.44%) Insuffisance ca	I500 (151) (1.95%	I500 (286) (2.58%



Prise en charge du patient âgé : le cas de la MARTINIQUE



CEREGMIA - PRED - SS (Chaire PMSI)

Déconnexion (SEBASTIEN Nathalie)

Régionale des parts de marché des Etablissements / spécialité (Groupe DP)

2012

Martinique

Tous les établissements

Age nimimum : 75

Age maximum : 130

01 Affections du système nerveux D-0108 Maladies dégénératives du système nerveux

DP	C-H-LOUIS DOMERGUE	HOPITAL DU MARIN	HOPITAL DE SAINT-ESPRIT	HOPITAL LOCAL DU FRANCOIS	CENTRE HOSPITALIER DU LAMENTIN	CHU DE FORT-DE-FRANCE	CLINIQUE SAINT PAUL	S. A. CLINIQUE SAINTE MARIE	A.T.I.R.	STEER SARL	E.T.E.E.R.	Cumul
G309 => Maladie d'Alzheimer, sans précision	3/1734 (17.65 %) / (0.17 %) cumul (0.17 %)	0/410 (0 %) / (0 %) cumul (0 %)	0/305 (0 %) / (0 %) cumul (0 %)	0/112 (0 %) / (0 %) cumul (0 %)	14/8861 (82.35 %) / (0.16 %) cumul (0.16 %)	0/11736 (0 %) / (0 %) cumul (0 %)	0/2097 (0 %) / (0 %) cumul (0 %)	0/1126 (0 %) / (0 %) cumul (0 %)	0/246 (0 %) / (0 %) cumul (0 %)	0/225 (0 %) / (0 %) cumul (0 %)	0/236 (0 %) / (0 %) cumul (0 %)	17
G20 => Maladie de Parkinson	3/1734 (18.75 %) / (0.17 %)	0/410 (0 %) / (0 %) cumul	4/305 (25 %) / (1.31 %)	0/112 (0 %) / (0 %) cumul (0	7/8861 (43.75 %) / (0.08 %)	1/11736 (6.25 %) / (0.01 %)	0/2097 (0 %) / (0 %) cumul	1/1126 (6.25 %) / (0.09 %)	0/246 (0 %) / (0 %)	0/225 (0 %) / (0 %)	0/236 (0 %) / (0 %) cumul	16

Prise en charge du patient âgé : le cas de la GUYANE

CEREGMIA - PRED - SS (Chaire PMSI) | Déconnexion (SEBASTIEN Nathalie)

Régionale des parts de marché des Etablissements / spécialité (Groupe DP)

2012

Guyane

Tous les établissements

Age nimimum : 75

Age maximum : 130

01 Affections du système nerveux D-0108 Maladies dégénératives du système nerveux

DP	CENTRE MED.-CHIRURG. DE KOUROU	CENTRE HOSPITALIER DE CAYENNE	CLINIQUE VERONIQUE	CENTRE HOSPITALIER FRANK JOLY	Cumul
G218 => Autres syndromes parkinsoniens secondaires	0/315 (0 %) / (0 %) cumul (0 %)	1/1749 (100 %) / (0.06 %) cumul (0.06 %)	0/965 (0 %) / (0 %) cumul (0 %)	0/169 (0 %) / (0 %) cumul (0 %)	1
G231 => Ophtalmoplégie supranucléaire progressive [maladie de Steele-Richardson-Olszewski]	0/315 (0 %) / (0 %) cumul (0 %)	1/1749 (100 %) / (0.06 %) cumul (0.11 %)	0/965 (0 %) / (0 %) cumul (0 %)	0/169 (0 %) / (0 %) cumul (0 %)	1
G258 => Autres syndromes précisés extrapyramidaux et troubles de la motricité	0/315 (0 %) / (0 %) cumul (0 %)	0/1749 (0 %) / (0 %) cumul (0.11 %)	0/965 (0 %) / (0 %) cumul (0 %)	1/169 (100 %) / (0.59 %) cumul (0.59 %)	1
G8100 => Hémiplégie flasque récente, persistant au-delà de 24 heures	0/315 (0 %) / (0 %) cumul (0 %)	0/1749 (0 %) / (0 %) cumul (0.11 %)	0/965 (0 %) / (0 %) cumul (0 %)	1/169 (100 %) / (0.59 %) cumul (1.18 %)	1
Cumul DP	0 (0 %)	2 (50 %)	0 (0 %)	2 (50 %)	4

Parcours Patient : le cas de la GUADELOUPE

Parcours patients entre établissements (2007 à 2012)

Guadeloupe

Établissement : C.H.U. DE POINTE-A-PITRE/ABYMES

Age nimimum : 75

Age maximum : 130

Z491 Dialyse extra-corporelle : 81

Établissement de départ	Num patient	Mois sortie	Année
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	752E6YKAEHR3KCFUE	Avril	2007
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	752E6YKAEHR3KCFUE	Février	2009
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	752E6YKAEHR3KCFUE	Avril	2011
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	752E6YKAEHR3KCFUE	Novembre	2012
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	752E6YKAEHR3KCFUE	Décembre	2012
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	7SE41QJSB6X1PTQCE	Mai	2007
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	7SE41QJSB6X1PTQCE	Février	2009
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	7SE41QJSB6X1PTQCE	Janvier	2012
C.H.U. DE POINTE-A-PITRE/ABYMES (970100228)	MUBGAUXJTPE8PK0YE	Novembre	2007
CENTRE HOSPITALIER DE BOULOGNE (620103440)	MUBGAUXJTPE8PK0YE	Novembre	2011
(060792926)	MUBGAUXJTPE8PK0YE	Mai	2012
(060792926)	MUBGAUXJTPE8PK0YE	Juin	2012
CENTRE HOSPITALIER DE BOULOGNE (620103440)	MUBGAUXJTPE8PK0YE	Août	2012
CENTRE HOSPITALIER DE BOULOGNE (620103440)	MUBGAUXJTPE8PK0YE	Septembre	2012
CENTRE HOSPITALIER DE BOULOGNE (620103440)	MUBGAUXJTPE8PK0YE	Octobre	2012
CENTRE HOSPITALIER DE BOULOGNE (620103440)	MUBGAUXJTPE8PK0YE	Novembre	2012
CENTRE HOSPITALIER DE BOULOGNE (620103440)	MUBGAUXJTPE8PK0YE	Décembre	2012

CEREGMIA - PRED - SS (Chaire PMSI)

Déconnexion (SEBASTIEN Nathalie)

Plugin CDF Mathematica (pour Demo 1, Demo 2, Demo 3)
Télécharger : Demo 1
Télécharger : Demo 2
Télécharger : Demo 3

Analyse DP/DAS(s) | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013

Gestion des diagnostiques d'entrées : Categorie majeur diag

Gestion des diagnostiques d'entrées : Groupe DP

Durées de séjours

Durées de séjours Hors Régions

Parcours patients entre établissements (2007 à 2012)

Courbe des âges

Analyse Régionale des parts de marché des Etablissements / spécialité (ou DP)

Analyse Régionale des parts de marché des Etablissements / spécialité (Groupe DP)

Analyse Régionale des parts de marché des Etablissements / spécialité (Catégories majeures)

Nombre moyen de DAS associés à un DP

Parcours patients par Région

Carte des pathologies

Analyse taux de réhospitalisation a -1mois sur un même groupe de DP

Taux de recours à l'hospitalisation régionale / par spécialité / en fonction de la population

Sévérité

Orientations et pistes de recherche (1/2)

- **Une utilisation optimale de l'information médico-économique : « tout savoir pour enfin pouvoir »…**
  - **Utilisation de l'information médico-économique,** seule ou combinées à d'autres bases de données financières, socio-économiques, etc., **à des fins stratégiques**
    - Meilleure appréhension des **déterminants des établissements en terme de description de l'activité médicale et de leur positionnement stratégique**
    - Redéfinition du **positionnement stratégique des établissements aux regards des contraintes médicales et socio-économiques du territoire et réallocation des ressources dédiées (type matrice BCG)**
    - **Réingénierie des processus de prise en charge des patients grâce à des *benchmarks d'activité entre services* et définition de l'*organisation polaire sur une base de regroupement de services de soins***

Orientations et pistes de recherche (2/2)

- **Une utilisation optimale de l'information médico-économique : « tout savoir pour enfin pouvoir »…**
  - **Développement d'outils d'aide à la décision en matière d'organisation et de régulation de l'offre de soins au niveau du territoire de santé**
  - **Spécialisation des établissements voire des territoires de santé au regard des potentialités**
- **Pour un rapprochement dans la création de savoir au bénéfice de la collectivité entre les hospitaliers et les tutelles d'une part et la recherche d'autre part** (Rapport sur la gouvernance et l'utilisation des données de santé, IGAS, septembre 2013)
- Elargissement des travaux aux données des **établissements de Psychiatrie, SSR et HAD** notamment dans une logique de recherche sur la médecine de parcours

ANNEXE 3

- [publication]
  Elisabeth, E. (2015). Fouille de données spatio-temporelle : Application à un système de modélisation des déplacements touristiques Document numérique et Société - 5e conférence

Fouille de données spatio-temporelle
Application à un système de modélisation des déplacements touristiques

ELISABETH Erol,
Doctorant
Université des Antilles et de la Guyane
Centre d'Etude et de Recherche en Economie, Gestion, Modélisation
et Informatique Appliquée (CEREGMIA)

Email : erol@beepway.com
Adresse postale : c/o CEREGMIA BP 7209 – 97275 SCHOELCHER Cedex (Martinique)
Tél. : +596 596 72 74 00

Le poids économique du tourisme mondial, le place à la fois comme première industrie avec des parts de marché qui ne cessent de croître mais également comme une économie de services en profonde et perpétuelle mutation. C'est l'offre qui crée la demande caractérisée par de fortes disparités de comportements suivant la diversité des attentes des clientèles et des revenus. Or la compétitivité des destinations touristiques passe par une compréhension desdits comportements.

Nous proposons dans cet article des outils algorithmiques permettant la modélisation non-supervisée des parcours touristiques dans une optique de recommandations personnalisées, qui s'appuie sur une modélisation des comportements.
Nous nous sommes particulièrement intéressés à l'optimisation du système de modélisation des comportements (résumé de données). Ce projet comporte en effet un volet applicatif important qui vise à fournir une méthode pour la mise place d'un système de recommandations basées sur la collecte de données de positionnement GPS enregistrées en temps réel. Il s'agit de créer, dans le cadre d'un apprentissage des déplacements, des classes représentants les caractères de ces déplacements (positions géographique, durées, ordonnancement).
La principale problématique dans le domaine de la conception de systèmes de recommandations est de produire des recommandations de qualité tout en minimisant, si possible, l'effort du traitement informatique requis de la part des producteurs et des consommateurs. Ce problème est particulièrement prégnant lorsque le touriste utilise des dispositifs mobiles embarqués.

Les résultats des travaux montrent la pertinence de cette approche.

Mots-clefs
Tourisme, parcours touristiques, gps, règles d'association, motifs séquentiels, q-motifs, kmeans

Introduction

Beaucoup de sites Web d'e-commerce rendent de nombreux services. Dans ce contexte de foisonnement d'informations, une recherche de produits pourrait renvoyer un très grand ensemble d'accès. Sans l'appui d'un système, filtrant les produits non pertinents, comparant des solutions de rechange, choisir la meilleure option peut être difficile voire impossible pour l'utilisateur connecté par un dispositif mobile.

Ces systèmes sont souvent des systèmes de recommandations. L'objectif d'un système de recommandations est d'aider les utilisateurs à faire leurs choix dans un domaine où ils disposent de peu d'informations pour trier et évaluer les alternatives possibles.

Des systèmes de recommandations manuels élaborés aident aussi à la découverte de nouveaux produits ou services, dans un cadre strictement supervisé. C'est la caractéristique principale des approches traditionnelles des systèmes de recommandation : les systèmes rassemblent des préférences d'utilisateur en interrogeant explicitement l'utilisateur. Le système exploite les préférences acquises pour activer l'algorithme spécifique de recommandation. Bien que ces préférences tendent à être fiables, l'approche a plusieurs inconvénients. D'abord, les utilisateurs doivent avoir assez de connaissances sur le domaine pour rendre la préférence explicite selon le modèle de produit (par exemple, les attributs du produit). En second lieu, les préférences incertaines ou inachevées peuvent devenir claires pendant que les utilisateurs agissent avec le système et comprennent mieux ce qu'ils veulent et quels produits sont disponibles, ce qui signifie qu'ils ne peuvent pas être demandés avant que le système ne fournisse quelques recommandations. Troisièmement, peu d'utilisateurs sont disposés à indiquer leurs préférences jusqu'à ce qu'ils reçoivent un certain "bénéfice" du système.

Le système est essentiellement algorithmique et nécessite une puissance de calculs importante pour définir un profil de goûts uniques. Ces algorithmes peuvent être utilisés dans un cadre sortant de l'internet sur des objets connectés mais hors du champ des sites internets.
L'objet connecté est un module de suivi gps installé dans des véhicules de location destiné aux touristes. Ce mode de transport destiné aux particuliers notamment au départ des aéroports, est largement utilisé par les touristes.

Spécificités de notre étude

Les systèmes de recommandation sont aujourd'hui orientés vers le e-marketing et l'amélioration de la relation client (e-CRM). Mais on doit imaginer les extensions des modèles et un élargissement des utilisations dans la recherche d'information, l'analyse des usages, la personnalisation, appliqué notamment au tourisme.

La plupart des approches fondées sur la fouille de données sont principalement des approches statistiques où l'ordre d'occurrence d'événements dans l'historique n'est pas pris en compte lors du calcul de recommandation. Or dans la modélisation des parcours et des durées cet ordre est une composante très importante, et une autre limitation dans notre cadre d'étude.

Les systèmes existants sont principalement orientés dans le domaine d'aide à "l'achat" sur le web. L'acheteur potentiel devient une structure de données qui décrit les centres d'intérêts dans l'espace des objets commerciaux à recommander. Notre approche se caractérise par la prise en compte dans la recommandation de la notion de positionnement géographique mais également de donnée temporelle.

Problématiques étudiées
Les problèmes sur lesquels il faudra se pencher pour la mise en œuvre de notre approche concernent d'une part la définition des méthodes et des techniques de mesure de similarités entre comportements et d'autre la définition des techniques de recommandations personnalisées.

Ce projet comporte un volet applicatif important : fournir une méthode pour la mise place un système de recommandation basée sur la collecte de données de positionnement GPS de véhicules loués pas les touristes. Il s'agit dans le cadre d'un apprentissage des déplacements de créer des classes représentants les caractères de ces déplacements (positions géographique, durées, ordonnancement).

Qu'est-ce que la fouille de données ? - Définition générale

Le data mining est un processus d'extraction de connaissances valides et exploitables à partir de grands volumes de données. Il a vocation à être utilisé dans un environnement professionnel et se distingue de l'analyse de données et de la statistique par les points suivants :

- Contrairement à la méthode statistique, le data mining ne nécessite jamais que l'on établisse une hypothèse de départ qu'il s'agira de vérifier. Ce sont des données elles-mêmes que sont déduites les corrélations intéressantes, le logiciel n'étant là que pour les découvrir (le data mining se situe à la croisée des statistiques, de l'intelligence artificielle, des bases de données).
- Les connaissances extraites par le data mining ont vocation à être intégrées dans le schéma organisationnel. Le data mining impose donc d'être capable d'utiliser de manière opérationnelle les résultats des analyses effectuées, souvent dans des délais très courts. Le processus d'analyse doit permettre à l'organisation une réactivité (très) importante.
- Les données traitées sont issues des systèmes de stockage en place dans l'organisation et sont ainsi hétérogènes, multiples, plus ou moins structurées. Leur raison d'être n'est donc a priori pas l'analyse (sauf dans le cas d'un entrepôt de données). Cela impose de disposer de systèmes performants de préparation ou de manipulation de données.
- Le data mining se propose de transformer en information, ou en connaissance, de grands volumes de données qui peuvent être stockés de manière diverse, dans des bases de données

relationnelles, dans un (ou plusieurs) entrepôt de données (datawarehouse), mais qui peuvent aussi être récupérées de sources riches « bien renseignées plus ou moins structurées comme internet, ou encore en temps réel (sollicitation d'un centre d'appel, retrait d'argent dans un distributeur à billets...). Lorsque la source n'est pas directement un entrepôt de données, il s'agit très souvent de construire une base de données ou un datamart dédié à l'analyse et aux analystes. Cela suppose d'avoir à sa disposition une palette d'outils de gestion de données (data management). On peut également structurer les données de l'entrepôt sous forme d'un hypercube OLAP (anglais online analytical processing abr) traitement analytique en ligne, même si cela est assez rare en matière de data mining. Parmi les utilisations du data mining on peut citer certains exemples (analyser les comportements des consommateurs dans la grande distribution, prédire le taux de réponse à un mailing, dans le secteur bancaire prédire la perte d'un client, détecter des comportements anormaux ou atypiques etc.).

Principe et méthodes

Le data mining est différent de la statistique qui fixe une hypothèse et que les données permettent ou non de confirmer. Dans le data mining on adopte pour une démarche sans a priori et cherche à faire émerger, à partir des données, des inférences que l'expérimentateur dont il faudra évaluer la qualité.

Le data mining se sert d'algorithmes issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) pour construire des modèles à partir des données, c'est-à-dire trouver des schémas « intéressants » (des patterns ou motifs en français) selon des critères fixés au départ, et extraire de ces données la connaissance utile. Cette connaissance peut être les motifs fréquents ou non fréquents. Il peut être intéressent d'obtenir les règles les plus fréquentes, mais aussi dans certains cas les règles les moins fréquentes en fonction de ce que l'on cherche.

Le mode non-supervisé

Les algorithmes non-supervisés permettent de travailler sur un ensemble de données dans lequel aucune des données ou des variables à disposition n'a d'importance particulière par rapport aux autres, c'est-à-dire un ensemble de données dans lequel aucune variable n'est considérée individuellement comme la cible, l'objectif de l'analyse.

On les utilise par exemple pour dégager d'un ensemble d'individus des groupes homogènes (typologie), pour construire des normes de comportements et donc des déviations par rapport à ces normes (détection de fraudes nouvelles ou inconnues à la carte bancaire, à l'assurance maladie...), pour réaliser de la compression d'informations (compression d'image)... Voici une liste non exhaustive des techniques disponibles :

Techniques à base de Réseau de neurones : carte de Kohonen (SOM/TOM) (carte auto adaptative).

Techniques utilisées classiquement dans le monde des statistiques : classification ascendante hiérarchique, k-means et les nuées dynamiques (Recherche des plus proches voisins), les classifications mixtes (Birch...), les classifications relationnelles... Une des techniques de classification non supervisée (clustering) les plus utilisées. kmeans :

Etant donné un entier K, K-means partitionne les données en K groupes, ou "clusters", ou "classes" ne se chevauchant pas. Ce résultat est obtenu en positionnant K "prototypes", ou "centroïdes" dans les régions de l'espace les plus peuplées. Chaque observation est alors affectée au prototype le plus proche (règle dite "de la Distance Minimale"). Chaque classe contient donc les observations qui sont plus proches d'un certain prototype que de tout autre prototype (image inférieure de l'illustration ci dessous).

Les techniques dites de recherche d'associations (elles sont à l'origine utilisées pour faire de l'analyse dite de panier d'achats ou de séquences, c'est-à-dire pour essayer de savoir parmi un ensemble d'achats effectués par un très grand nombre de clients et de produits possibles, quels sont les produits qui sont achetés simultanément (pour un supermarché par exemple ; elles sont également appliquées à des problèmes d'analyse de parcours de navigation de site web). Ces techniques peuvent donc être utilisées de manière supervisées) : algorithmes a priori, GRI, Carma, méthode ARD... Analyses de liens

Résumé de données de positions GPS

La croissance du volume des données à traiter ouvre la voie à de nouveaux champs d'étude. Dans certains cas, il est possible d'atténuer l'influence de la quantité de données sur les temps de réponse des traitements et algorithmes. Les traitements concernés peuvent se satisfaire d'approximations, généralement de moindre qualité que des résultats non approximatifs, mais obtenues en des temps plus acceptables.

C'est le cas de l'échantillonnage au sein d'une population ou des problèmes NPcomplets dont une solution approchée peut être obtenue par des algorithmes polynomiaux.

Le résumé de données permet de réduire un ensemble de données volumineux à un ensemble de taille plus réduite, épuré de ce que l'on considérera comme de l'information non pertinente ou non signifiante, comme du bruit. Elles sont ainsi très souvent, mais pas systématiquement, utilisées en amont des techniques supervisées ou non supervisées. Elles sont notamment très complémentaires des techniques non supervisées.

Le systeme de géolocalisation gps installé dans les véhicules envoi des données en continu.

Le Global Positioning System (GPS) est un système de positionnement mondial. Ce système imaginé par le physicien D. Fanelli et mis en place à l'origine par le Département de la Défense des États-Unis.

Les données transmises par les satellites peuvent être librement reçus et exploités. Un récepteur peut connaître sa position sur la surface de la Terre, avec une précision aujourd'hui de l'odre de 10 metres en moyenne. Dans notre étude 12 véhicules ont été équipé. Ces véhicules ont envoyés des données pendant plusieurs mois.

Le systeme GPS comprend au moins 24 satellites activés en orbite à 20 200 km d'altitude. Ces satellites émettent en permanence sur deux fréquences L1 (1 575,42MHz) et L2 (1 227,60 MHz) un signal complexe, constitué de données numériques et d'un ensemble de codes pseudo-aléatoires, daté précisément grâce à leur horloge atomique. Les données numériques, transmises à 50 bit/s, incluent en particulier des éphémérides permettant le calcul de la position des satellites, ainsi que des informations sur leurs horloges internes. Les codes sont un code C/A (acronyme de coarse acquisition, acquisition grossière) à 1,023 Mbit/s et de période 1 ms, et un code P (pour précision) à 10,23 Mbit/s avec une période de 280 jours. Le premier est librement accessible, le second est réservé aux utilisateurs autorisés ; il est le plus souvent chiffré. Les récepteurs commercialisés dans le domaine civil utilisent le code C/A. Quelques rares utilisateurs civils spécialisés, comme les organismes de géodésie, ont accès au code P.

Un récepteur GPS qui capte les signaux d'au moins quatre satellites équipés de plusieurs horloges atomiques peut, en calculant les temps de propagation de ces signaux entre les satellites et lui, connaître sa distance par rapport à ceux-ci et, par trilatération, situer précisément en trois dimensions n'importe quel point placé en visibilité des satellites GPS2, avec une précision de 15 à 100 mètres pour le système standard. Le GPS est ainsi utilisé pour localiser des véhicules roulants, des navires, des avions, des missiles et même des satellites évoluant en orbite basse. La précision de la position horizontale est de l'ordre de 10 mètres, la position verticale est fausse et varie énormément. Le GPS étant un système développé pour les militaires américains, une disponibilité sélective a été prévue : certaines informations, en particulier celles concernant l'horloge des satellites, peuvent être volontairement dégradées et priver les récepteurs qui ne disposent pas des codes correspondants dela précision maximale.

Pendant de nombreuses années, les civils n'avaient ainsi accès qu'à une faible précision (environ 100 m). Le 1er mai 2000, le président des Etats-Unis a mis fin à cette dégradation volontaire.

Certains systèmes GPS conçus pour des usages très particuliers peuvent fournir une localisation à quelques millimètres près. Il utilise alors un système différentiel. Le GPS différentiel (DGPS), corrige

ainsi la position obtenue par GPS conventionnel par les données envoyées par une station terrestre de référence localisée très précisément. La station terrestre connaissant son "imprécision" elle l'applique à la position reçue.

Dans certains cas, seuls trois satellites peuvent suffire. La localisation en altitude (axe des Z) n'est pas d'emblée correcte alors que la longitude et la latitude (axe des X et des Y) sont encore bonnes. On peut donc se contenter de trois satellites lorsque l'on évolue au-dessus d'une surface « plane » (océan, mer). Ce type d'exception est surtout utile au positionnement d'engins volants (tels les avions) qui ne peuvent pas se reposer sur le seul GPS, trop imprécis pour leur donner leur altitude. Il existe un modèle de géoïde mondial nommé « Earth Gravity Model 1996 » ou EGM964 associé au WGS 84 qui permet, à partir des coordonnées WGS 84, de déterminer des altitudes rapportées au niveau moyen des mers avec une précision d'environ 1 mètre.

Des récepteurs GPS évolués incluent ce modèle pour fournir des altitudes plus conformes à la réalité.

Les règles d'association sont devenues un concept majeur en fouille de données pour représenter les relations quasi-implicatives entre des variables booléennes (dénommées items). Depuis les premiers travaux d'Agrawal et al. (1993), de nombreux algorithmes ont été proposés pour découvrir efficacement de telles connaissances dans de grandes bases de données. Tous engendrent d'énormes quantités de règles, dues à l'explosion combinatoire du nombre de conjonctions d'items traitées.

Les présents travaux visent à trouver des solutions au problème du volume de données, de l'explosion combinatoire notamment pour les règles d’associations et les motifs séquentiels.
Il est nécessaire de faire un premier résumé de données. Ce résumé de donné consiste à définir ce qu'est un arrêt et à le coder en terme informatique.

La collecte de données de comportements

Le boitier GPS de géolocalisation installé dans le véhicule a une mémoire qui lui permet de ne pas perdre les données collectées en cas de réseau GSM / GPRS indisponible. En fonction du matériel embarqué la forme des données reçues est sensiblement la même :

- numéro automatique d'insertion de ligne (numauto = ex 5)
- numéro de boitier gps embarqué (numbeepway = ex 344691000067372)
- l'ip publique du boitier gps (ip = ex 193.251.163.1)
- la date et l'heure de la position collectée (date heure = ex 2009-12-22 00 :00 :3)
- la latitude de la position collectée (latitude = ex 1436.0512)
- l'orientation NS de la position collectée (ns = ex N)
- la longitude de la position collectée (longitude = ex 06056.1889)
- l'orientation EW de la position collectée (ew = ex W)
- la hauteur de la position collectée (hauteur = ex 138)
- la vitesse de la position collectée (vitesse = ex 16.01)
- des informations complémentaires de la position collectée (info =; ;20.00) (cumul en km depuis le dernier démarrage)

Une trame peut donc avoir la forme suivante :

5 344691000067372 193.251.163.1 2009-12-22 00 :00 :37 1436.0512 N 06056.1889 W 138 16.01 ; ;20.00

Un trajet étant la succession des arrêts entre le premier arrêt à partir de la première location jusqu'au retour au parking du loueur. Un premier traitement a été réalisé afin de générer pour les arrêts collectés (position longitude, latitude) une table de succession des arrêts pour chaque trajet.

Les jeux de données

Les jeux de données TRAJETS sont constitués de succession de d'arrêts caractérisés par un début : la prise du véhicule dans le parking du loueur et une fin par le retour de ce véhicule dans le parc.
La création du jeu TRAJET prend la forme suivante en langage SQL.

```
CREATE TABLE IF NOT EXISTS ' jeux_arret '          (
'num' i n t ( 1 1 ) NOT NULL AUTO_INCREMENT,
'nom_jeux ' var char (100) NOT NULL,
' l igne_jeux ' i n t ( 2 0) NOT NULL,
'num_beepway ' var char (250) NOT NULL,
'date ' datet ime NOT NULL,
' duree_ar ret ' var char ( 2 0) NOT NULL,
' num_point ' i n t ( 1 1 ) NOT NULL,
' data ' text NOT NULL,
PRIMARY KEY ( 'num' )                              )
```

Il est possible après de manipuler ces données très simplement avec les langages de programmations traditionnels. Nous pouvons fournir un exemple de code PHP destiné à lister les jeux de données :

```
1 <?php
2 // Ouverture de la connexion la base
3 // Recherche tous l e s beepways qui appar t i ennent aux f l o t t e s de $tab_f lot tes_4 $ s q l =
"SELECT DISTINCT ( 'nom_jeux ' ) FROM ' jeux_ar ret ' " ;
5 $ r e s u l t = mysql_db_query ( $base , $sql , $ l i n k ) ;
6
7 // . . . s i on a des r s u l t a t s
8 while ( $row = mysql_fetch_array ( $ r e s u l t ) ) {
9 ?php echo $row [ 0 ] ;
10 }
11 // . . . fermeture de la connexion la base de données
12 ?>
```

Dans cette étude le jeu de données est constitué d'un ensemble d'arrêts. Une structure simple et ordonnée du stockage permet d'avoir les bases pour des traitements complexes et multiples (règles d'association, motifs séquentiels, q-motifs…kmeans...). Ce papier explique comment obtenir les analyser ces données avec les règles d'association et les motifs séquentiels.

Le volet applicatif

Dans cette situation, la fouille de données se propose de donner les outils et/ou techniques nécessaires pour l'extraction de ces connaissances. Deux classes de motifs se sont alors avérées très utiles et simultanément utilisées dans la pratique, à savoir : les itemsets fréquents les règles d'association Un itemset est une conjonction d'items relatifs au contexte d'extraction alors qu'une règle d'association est une expression causale avec parties prémisse et conséquence et probabiliste ayant une fréquence ou support et une force ou confiance des « co-occurrences » entre les items de la prémisse et ceux de la conclusion.

Toutefois, l'ensemble de tous les itemsets fréquents et de toutes les règles valides (par rapport aux mesures de support et de confiance) extrait à partir des contextes réels est généralement de taille importante, dont une bonne partie est redondante.
Pour pallier à cette situation, un nombre important de travaux propose d'extraire seulement un sous-ensemble représentatif, appelé représentation concise.

Dans ce volet, nous allons détailler les notions relatives à ces deux classes de motifs et donner un aperçu des principales approches permettant de réduire la taille des ensembles extraits. Le cadre de cet exposé concerne l'extraction des connaissances à partir des données (ECD).

Le but d'un processus d'ECD est d'extraire des connaissances qui sont non triviales, potentiellement utiles et significatives. Plus précisément, nous investiguons deux classes de motifs qui se sont avérées très utiles dans l'ECD.

Les itemsets fréquents, Les règles d'association. L'extraction de connaissances dans les bases de données, également appelé data mining, désigne le processus non trivial permettant d'extraire des informations et des connaissances utiles qui sont enfouies dans les bases de données, les entrepôts de données (data warehouses) ou autres sources de données.
Nous traitons des problèmes de la génération efficace des règles d'association et de la pertinence et de l'utilité des règles d'association extraites. Une règle d'association est une implication conditionnelle entre ensembles d'attributs binaires appelés items. Dans l'ensemble des travaux existants, l'extraction de règles d'association est décomposée en deux sous-problèmes qui sont la recherche des ensembles fréquents d'items et la génération des règles d'association à partir de ces ensembles. Le premier sous-problème, dont la complexité est exponentielle dans la taille de la relation et qui nécessite de parcourir à plusieurs reprises celle-ci, constitue la phase la plus coûteuse en termes de temps d'exécution et d'espace mémoire.

Nous proposons une nouvelle sémantique pour le problème de l'extraction des règles d'association. Nous démontrons que les ensembles fermés fréquents d'items constituent un ensemble générateur non redondant pour les ensembles fréquents d'items et les règles d'association.

Les résultats expérimentaux démontrent que ces algorithmes permettent de réduire les temps d'extraction et l'espace mémoire nécessaire dans le cas de jeux de données constitués de données denses ou corrélées.

Nous adaptons pour cela les bases pour les règles d'implication définies en analyse de données et nous définissons de nouvelles bases constituées des règles non redondantes d'antécédents minimaux et de conséquences maximales à partir des ensembles fermés fréquents. Nous proposons également des algorithmes efficaces de génération de ces bases.

Les règles d'associations

Les règles d'associations utilisent des méthodes de génération combinatoires et engendrent un nombre élevé de règles qui sont difficilement exploitables. Plusieurs approches de réduction de ce nombre ont été proposées comme l'usage de mesures de qualité, le filtrage syntaxique par contraintes, la compression par les bases représentatives ou génériques. Cependant, ces approches n'intègrent pas l'expert dans le déroulement du processus limitant ainsi l'aspect interactif du processus. En effet, l'expert ne sait pas toujours initialement quelle connaissance il souhaite obtenir. Nous analysons l'activité cognitive de l'expert dans différents processus de recherche de règles d'association et nous montrons que dans ces approches, l'expert n'intervient pas durant les tâches du processus. Pour accroître cette interactivité avec l'expert, il est nécessaire que celui-ci soit au coeur du processus afin de répondre à l'un des objectifs de l'ECD.

L'utilisation d'une interface graphique adaptée s'avère donc nécessaire pour que l'expert puisse interagir de manière optimale avec le processus. L'efficacité de cet algorithme a été montrée sur un problème réel de marketing faisant intervenir des experts du monde touristique.

En outre, les résultats de la fouille de données avec une représentation visuelle présente un intérêt non négligeable puisque l'esprit humain peut traiter une plus grande quantité d'informations. Comme des quantités très importantes de règles sont générées, la fouille de données « visuelles » s'avère être une étape incontournable pour améliorer encore notre approche.
Parmi ces représentations, nous nous focalisons sur les représentations de type matrice 2D présentant la particularité de générer des occlusions. Une occlusion est un chevauchement d'objets dans un environnement 2D rendant certains de ces objets pas ou peu visibles. Après avoir défini formellement le problème d'occlusions, nous montrons qu'il s'agit d'un problème d'optimisation qui est de trouver le

meilleur ordre possible des itemsets sur les deux axes pour limiter les occlusions. Nous proposons une heuristique permettant de réduire significativement les occlusions générées. Les résultats que nous avons obtenus sont présentés et discutés.

L'intérêt des règles d'associations est qu'elles sont faciles à interpréter. La méthode réalise de l'apprentissage non supervisé ; elle est basée sur des calculs élémentaires, elle est très coûteuse en temps elle marche pour des découvertes de faits fréquents, elle peut produire des règles triviales et inutiles.

Stéphane LALLICH et Olivier TEYTAUD - Évaluation et validation de l'intérêt des règles d'association : Comme le soulignent (Hajek et Rauch 1999), l'une des premières méthodes de recherche des règles d'association est la méthode GUHA initiée par (Hajek, Havel et Chytil 1966), où apparaissent déjà les notions de support et de confiance. L'intérêt pour les règles d'association a été renouvelé par les travaux de (Agrawal, Imielinski et Swami 1993), (Agrawal et Srikant 1994), puis (Srikant et Agrawal 1995) ayant trait à l'extraction de règles d'association à partir des grandes bases de données qui enregistrent le contenu des transactions commerciales. Nous nous basons sur ces travaux afin de rendre les résultats simples et compréhensibles tout en intégrant la notion géographique sans ajouter de complexité aux résultats.

Les règles d'association ont été, initialement, utilisées en analyse de données puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données organisées selon des schémas relationnels de grandes tailles (Simon and Napoli, 1999). Elles ont été, par la suite, appliquées à la fouille de textes [(Feldman and Dagan, 1995),(Toussaint et al., 2000)]. Soit une base de données transactionnelle où chaque transaction est une liste d'items (achats par un client lors d'une visite). Beaucoup de mesures de qualité existent dans la littérature. La mesure de la qualité est dans notre approche liée à la finesse du choix du « support » et de la « confiance ».

La recherche de règles d'association intéressantes est un thème privilégié de l'extraction des connaissances à partir des données. Les algorithmes du type Apriori fondés sur le support et la confiance des règles ont apporté une solution élégante au problème de l'extraction de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes.
Il faut disposer d'autres mesures venant compléter le support et la confiance.
Dans la pratique chaque enregistrement est une transaction alors que les différents champs correspondent aux données susceptibles de composer la transaction. n est le nombre de transactions
p le nombre de données

Dans la mesure où l'on s'intéresse à la présence-absence de chaque article dans les différentes transactions, on associe à chaque article l'acte d'achat correspondant, appelé item, qui est une variable booléenne.

La formalisation du problème d'extraction de règles associatives a été introduit par Agrawal et al. [10]
Soit $I = \{i1, i2, \dots, in\}$ l'ensemble des items pouvant faire partie d'une base de données transactionnelle de D.

Soit *T* une transaction ; *T* est donc un sous-ensemble de $I : T \subseteq I$

Une règle d'association est de la forme $X \Rightarrow Y$ ou :

$X \subset I, Y \subset I \quad et \; X \cap Y = \emptyset$

Sur l'ensemble des transactions, on obtient une matrice booléenne de dimensions *n* et *p*. A un ensemble d'articles, on associe la conjonction des actes d'achat correspondant, ou itemset, qui est aussi une variable booléenne.

À partir de la matrice booléenne qui indique les articles présents dans chaque transaction, on veut extraire des règles si un touriste client va sur le site A et sur le site B. Il est probable qu'il aille aussi sur le site C.

Une règle d'association est ainsi une expression *r* du type $A \rightarrow B$, ou l'antécédent A et le conséquent B sont des itemsets qui n'ont pas d'items communs.

Dans la mesure où le nombre de règles d'association possibles croît exponentiellement avec le nombre d'items, il est capital de pouvoir se limiter à l'extraction des règles les plus intéressantes. Il faut pour cela être capable de définir celles-ci et de les identifier, puis il faut les valider.

Les algorithmes doivent être affinés en les liants aux critères de support et de confiance. Par support et confiance, on désigne précisément les algorithmes d'extraction qui recherchent de façon exhaustive les règles d'association dont le support et la confiance dépassent des seuils fixés au préalable par l'utilisateur, notés *min supp* et *min conf* .

Les algorithmes d'extraction liés à l'approche support-confiance parcourent les itemsets afin de rechercher les itemsets fréquents, donc ceux dont le support dépasse *min supp*, pour en déduire les règles d'association dont la confiance dépasse *min conf*.
Dans notre étude cette approche est essentielle car étant donné les volumes de données et leur nature géographique, une représentation qui ne prendrait pas en compte des seuils serait simplement illisible.

Cette méthode est appropriée car deux types de recherches sont possibles ; d'une part les itemsets fréquents, d'autre part les itemsets non fréquent.

L'algorithme fondateur « Apriori, (Agrawal et Srikant 1994) » procède en deux temps : premièrement on recherche les itemsets fréquents, ceux dont le support dépasse minsupp, en balayant le treillis des itemsets dans sa largeur et en calculant les fréquences par comptage dans la base, ce qui impose une passe sur la base; ensuite pour chaque itemset fréquent X, dont la confiance dépasse le seuil *min conf* .

Les Règles d'association Non Redondantes pourront être utiles dans le prolongement ce cette études, elles sont introduites par, Zaki [11]. Il introduit une nouvelle base générique de règles d'association, appelée base de règles d'association non Redondantes. Zaki propose d'utiliser l'axiome de transitivité de Luxenburger [12] ainsi que l'axiome d'augmentation d'Armstrong [13] ce qui permet de dériver l'ensemble de toutes les règles redondantes.

L'utilisation de ces axiomes ne garantit pas la génération de toutes les règles valides. Ainsi, le mécanisme d'inférence utilisé n'est pas complet, mais dans le cadre de notre étude il pourrait permettre d'obtenir un résultat plus rapidement. Il conviendrait de l'évaluer et de la comparer au résultats obtenus avec « Apriori ».

Au point de départ de l'algorithme utilisé, il faut fixer un seuil de support minimal pour que seules les règles d'association avec un support plus grand ou égal à ce seuil soient générées. Il génère tous les sous-ensembles de k items potentiellement fréquents à partir des sous ensembles des (k-1) items fréquents ; il élague tous les sous-ensembles de k items qui ne peuvent être fréquents. L'algorithme fait un k-ième passage dans la base de données pour calculer le support des sous-ensembles de k items générés et retenus.

Les règles déduites des itemsets fréquents ont nécessairement une confiance supérieure au seuil de support, dans la mesure où *Supp (A ! B) < Conf (A ! B)*. L'efficacité de Apriori diminue en présence de données denses ou fortement corrélées. Toute la difficulté de l'extraction des fréquents consiste à identifier la bordure entre itemsets fréquents et itemsets non-fréquents dans le treillis des itemsets [3].

La recherche peut se faire en largeur ou en profondeur. Dans chaque cas, on peut procéder par comptage direct de la fréquence de chaque itemset dans la base, ou procéder par intersection des deux itemsets qui constituent l'itemset candidat.

// Recherche des « beepways » appartenant aux flottes de $tab_flottes_recommandation

$ s q l = "SELECT DISTINCT ('nom_jeux ') FROM ' jeux_arret ' " ;

D'un point de vue pratique, l'extraction de toutes les règles valides se fait en deux étapes tout d'abord détermination des *itemsets* fréquents c'est à dire le support puis on pratique une dérivation des règles d'association valides c'est-à-dire de confiance.

La première étape est généralement la plus coûteuse car elle nécessite des accès au contexte d'extraction alors que la deuxième ne nécessite aucun.

Implémentation (Règle) R : Site touristique A -> Site touristique B

Dans ces jeux de données il est possible de connaitre la relation entre deux sites si elle existe. Cette existence est pondérée par un support et une confiance dans le calcul des règles d'associations.

Etape 1 : On fixe les paramètres et le jeu de donné résultant du résumé de données (liste des arrêts géo référencés)

Etape 2 : On calcule la matrice (Lignes = parcours ; Sites = Identifiant ID du site touristique)

Exemple : dans le parcours 8 le touriste a été sur le site
ID :12 (Habitation Clément) Rhumerie
Puis sur le site
ID : 16 (Les Rails de la Canne à Sucre)

Lignes\Sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1															
2		1														
3		1														
4			1													
5	1															
6		1														
7												1				
8												1				1
9																
10																
11			1			1										

Etape 3 : On calcule les supports et la confiance minimale afin d'éliminer les règles peu fréquentes.

Support X U X` :	Total AR X X`	Total AR X` X	AR X X`	AR X` X
1 2 = 3 => Support : 0.023	{1 }=>{2} = 0.333	{2 }=>{1} = 0.231	{1 }=>{2} = 0.333	{2 }=>{1} = 0.231
1 3 = 2 => Support : 0.015	{1 }=>{3} = 0.222	{3 }=>{1} = 0.069	{1 }=>{3} = 0.222	
1 15 = 1 => Support : 0.008	{1 }=>{15} = 0.111	{15 }=>{1} = 0.5		
1 16 = 1 => Support : 0.008	{1 }=>{16} = 0.111	{16 }=>{1} = 0.25		
1 23 = 3 => Support : 0.023	{1 }=>{23} = 0.333	{23 }=>{1} = 0.111	{1 }=>{23} = 0.333	
1 24 = 1 => Support : 0.008	{1 }=>{24} = 0.111	{24 }=>{1} = 0.143		
1 27 = 1 => Support : 0.008	{1 }=>{27} = 0.111	{27 }=>{1} = 0.25		
1 44 = 1 => Support : 0.008	{1 }=>{44} = 0.111	{44 }=>{1} = 0.111		
1 52 = 2 => Support : 0.015	{1 }=>{52} = 0.222	{52 }=>{1} = 0.182	{1 }=>{52} = 0.222	{52 }=>{1} = 0.182
1 57 = 1 => Support : 0.008	{1 }=>{57} = 0.111	{57 }=>{1} = 0.111		
1 80 = 7 => Support : 0.053	{1 }=>{80} = 0.778	{80 }=>{1} = 0.056	{1 }=>{80} = 0.778	
1 124 = 1 => Support : 0.008	{1 }=>{124} = 0.111	{124 }=>{1} = 0.167		
1 129 = 1 => Support : 0.008	{1 }=>{129} = 0.111	{129 }=>{1} = 0.25		
1 134 = 2 => Support : 0.015	{1 }=>{134} = 0.222	{134 }=>{1} = 0.286	{1 }=>{134} = 0.222	{134 }=>{1} = 0.286
1 135 = 1 => Support : 0.008	{1 }=>{135} = 0.111	{135 }=>{1} = 0.167		
1 137 = 1 => Support : 0.008	{1 }=>{137} = 0.111	{137 }=>{1} = 0.083		
1 140 = 1 => Support : 0.008	{1 }=>{140} = 0.111	{140 }=>{1} = 0.333		
1 179 = 1 => Support : 0.008	{1 }=>{179} = 0.111	{179 }=>{1} = 0.5		
1 187 = 2 => Support : 0.015	{1 }=>{187} = 0.222	{187 }=>{1} = 0.133	{1 }=>{187} = 0.222	
1 188 = 2 => Support : 0.015	{1 }=>{188} = 0.222	{188 }=>{1} = 0.4	{1 }=>{188} = 0.222	{188 }=>{1} = 0.4
1 197 = 1 => Support : 0.008	{1 }=>{197} = 0.111	{197 }=>{1} = 0.5		
1 200 = 1 => Support : 0.008	{1 }=>{200} = 0.111	{200 }=>{1} = 0.037		
1 209 = 1 => Support : 0.008	{1 }=>{209} = 0.111	{209 }=>{1} = 0.167		
1 217 = 2 => Support : 0.015	{1 }=>{217} = 0.222	{217 }=>{1} = 0.286	{1 }=>{217} = 0.222	{217 }=>{1} = 0.286
1 218 = 2 => Support : 0.015	{1 }=>{218} = 0.222	{218 }=>{1} = 0.222	{1 }=>{218} = 0.222	{218 }=>{1} = 0.222

Illustration #1

Note : La figure (copie écran) fait apparaitre toutes les données

Les Colonnes AR XX' & AR X'X permettent de « voir » les valeurs des règles. Certaines colonnes sont vides car le support ou la confiance est inférieur au seuil donné en paramètre.

Les motifs séquentiels A-> B

Introduits dans Agrawal (1995), la recherche de motifs séquentiels consiste ainsi à extraire des enchaînements d'ensembles d'items, couramment associés sur une période de temps bien définie. Le motif séquentiel met en évidence des associations inter-transactions, contrairement à celle des règles d'association qui extrait des combinaisons intra-transactions.
Les travaux récemment menés ont montré que toutes les approches qui visent à extraire l'ensemble des motifs séquentiels deviennent cependant inefficaces dès que le support minimal spécifié par l'utilisateur est trop bas ou lorsque les données sont fortement corrélées. En effet, dans ce cas, et plus encore que pour les itemsets, les recherches sont pénalisées par un espace de recherche trop important.
Pour essayer de gérer au mieux ces problèmes de et l'introduction de complexités spatiale et temporelle, deux grandes tendances se distinguent à l'heure actuelle. Dans le premier cas, les propositions comme PrefixSPAN Pei et al. (2004)[7] ou SPADE Zaki (2001) [5] se basent sur de nouvelles structures de données et une génération de candidats efficace. Les approches de la seconde tendance considèrent l'extraction d'une représentation condensée Mannila et Toivonen (1996) [2].
Depuis une bonne quinzaine d'années de nombreux travaux en extraction d'information et en fouille de textes appliquées au domaine biomédical ont vu le jour. Deux tâches sont particulièrement explorées correspondant aux deux requêtes mentionnées précédemment : la première est la reconnaissance d'entités nommées de type biologique (noms de gènes, protéines, fonctions biologiques, etc.) et la deuxième concerne l'identification et le typage de relations entre entités biologiques précédemment reconnues.

Ces approches ont fait leurs preuves pour des applications traitant les transactions financières, les suivis de navigations sur le web, les données musicales, la sécurité informatique etc.

A l'inverse de la problématique d'extraction des itemsets, les travaux sur l'extraction des motifs séquentiels, de part leur complexité, sont rares et de nombreux axes de recherche restent encore à étudier F. Masseglia et M. Teisseire et P. Poncelet (2004) [6].

L'équipe Tatoo développe un thème de recherche sur l'extraction des motifs séquentiels et ses applications à des gros volumes de données comme les données médicales (garantir le respect de la vie privée pour les données), les données multidimensionnelles (fouille de cube), les données du web (web sémantique). Le problème de l'extraction de motifs séquentiels peut sembler proche de celui de l'extraction de règles d'association. Ce rapprochement s'avère cependant très fragile en raison d'un élément clé qui est propre à l'extraction de motifs séquentiels : la temporalité. Cette notion permet à la fois de distinguer à l'intérieur des enregistrements un ordre d'apparition mais aussi de regrouper certains éléments.
Si les règles d'association s'appliquent à des données de type itemsets (et permettent l'extraction de règles intra-transaction), la recherche de motifs séquentiels s'applique à des données de type séquences d'itemsets (et permet donc l'extraction de règles inter-transactions). « *Introduits dans [AGR 95b] et largement étudiés dans [MAS 02], les motifs séquentiels peuvent être vus comme une extension de la notion de règles d'association intégrant diverses contraintes temporelles . La recherche de tels motifs consiste ainsi à extraire des enchaînements d'ensembles d'items, couramment associés sur une période de temps bien spécifiée. En fait, cette recherche met évidence des associations inter-transactions, contrairement à celle des règles d'association qui extrait des combinaisons intra-transactions. Par exemple, des motifs séquentiels peuvent montrer que 60 pourcent des gens qui achètent une télévision, achètent un magnétoscope dans les deux ans qui suivent. » F. Masseglia et M. Teisseire et P. Poncelet (2004) [6]*.

Ce problème, posé à l'origine dans des contextes de marketing, est important dans des domaines divers (détection de fraudes), la finance, ou encore la médecine (identification des symptômes précédant les maladies, données de facturations médicales RSS/RSF,).

La prise en compte importante de la temporalité dans les enregistrements à étudier permet une plus grande précision dans les résultats, mais implique aussi un plus grand nombre de calculs et de contraintes.

Définition
Soit D une base de données de transactions de clients où chaque transaction T est composée de :
Un identifiant du véhicule noté *Cid*
Le temps utilisé, notée temps
Un ensemble d'items 'itemset' intervenants dans la transaction i_t
Une transaction constitue, pour un touriste C, l'ensemble des items arrêts effectués par C à lors d'un même parcours (de la récupération du véhicule à sa restitution).

Une transaction s'écrit sous la forme d'un ensemble : id-client, id-date, itemset.
Dans notre exemple la fin de séquence est caractérisée par le retour du véhicule au parking du loueur. La rupture de séquence intervient donc à ce niveau.

le parcours réalisé est donc : 52,257,1,23,1,3,1,80 (Parking de remise du véhicule)

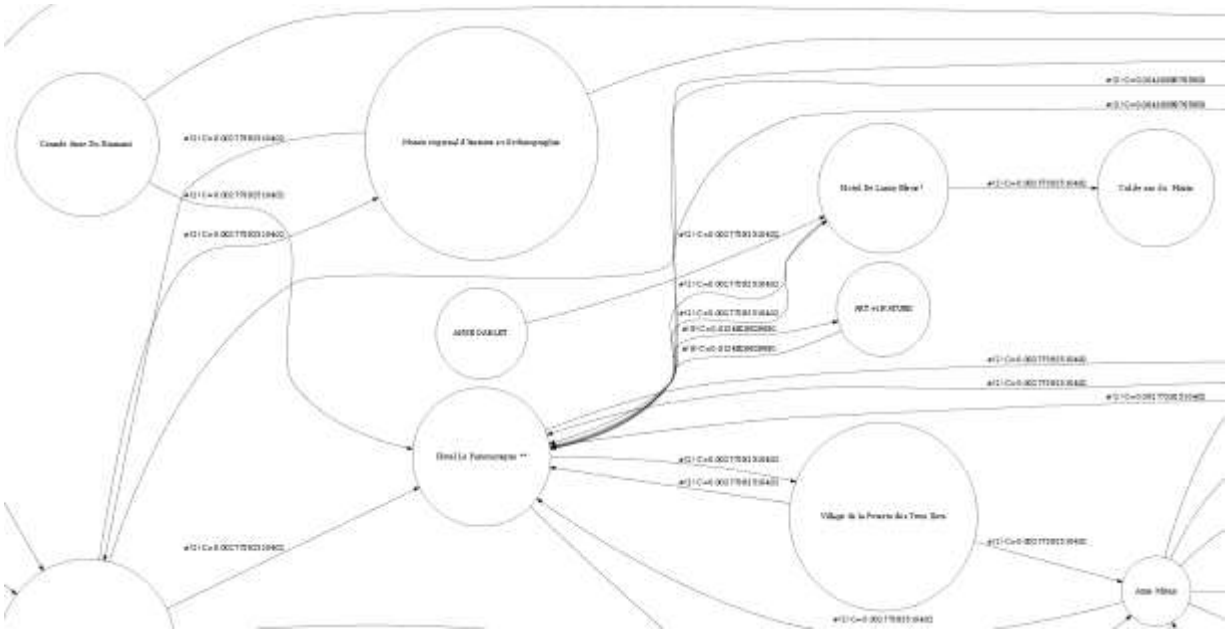


Illustration #2
Dans cette séquence on note la relation entre 52 : "ART et Nature" AND 257 : "Hotel le Panoramique".
Nous pouvons proposer deux types de représentations. La carte géographique des Motifs
conv(A → B) permet une représentation visuelle des comportements« arrêts touristiques ». Il est alors possible de créer un graph orienté de relations intersites.

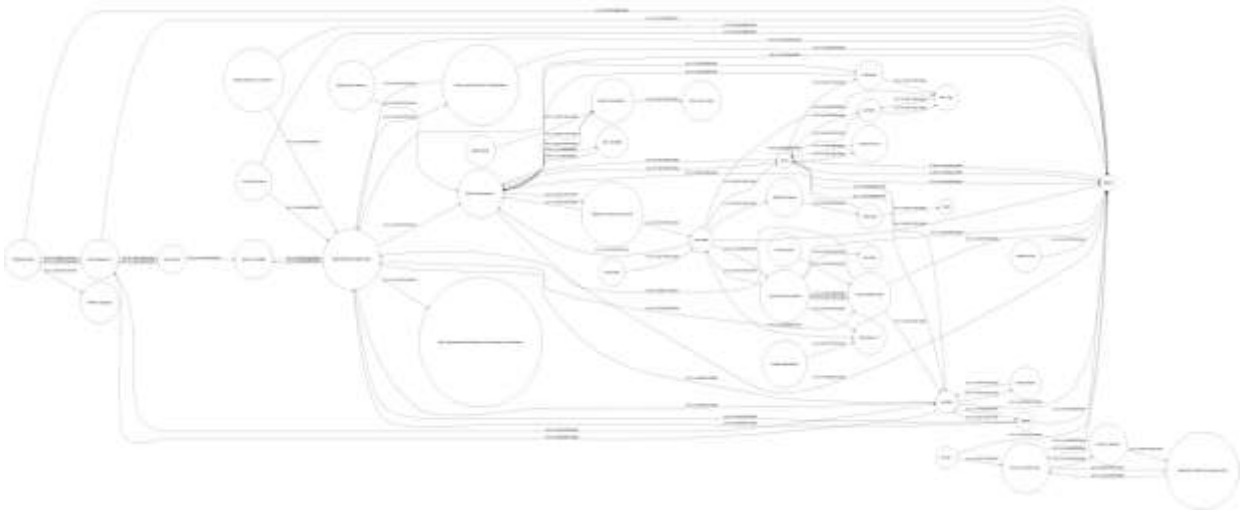


Illustration #3
Nous pouvons explorer l'ensemble des motifs séquentiels permettant de retracer les parcours des touristes de toute l'ile. Ce graph orienté est la base de travail pour les recommandations.

Conclusion

Notre sujet d'étude concerne le développent d'outils permettant la modélisation non-supervisé des parcours touristiques. L'inclusion d'appareil mobiles, aujourd'hui appelés « appareils connectés », dans notre étude permet une utilisation et une extension des algorithmes informatiques hors du champ traditionnel de l'internet.

Dans le domaine de la conception de systèmes de recommandations l'effort du traitement informatique requis est minimisé par l'introduction des résumés de données. Notre approche du résumé prend la forme de courtes séquences informatiques qui « résument » le parcours du touriste.

Ce volet applicatif a permis de mettre en lumière plusieurs approches de traitement des données. Il est l'élément de base qui permettra de proposer des recommandations aux touristes en temps réel en fonction des associations des sites « précédemment » visités, avec les sites dans lesquels son parcours s'inscrit dans le cadre du motif séquentiel.

Une approche complémentaire utilisant les règles d'association non Redondantes devra être explorée afin d'affiner et d'optimiser les résultats.

Cette approche possède aussi plusieurs extensions, notamment la « clustérisations » grâce aux algorithmes k-means qui permet de créer des groupes de comportements. Ces approches couplées à une vision en temps réel selon la position et la direction du véhicule de location permet de proposer une approche géographique de l'analyse des comportements et par extension de la recommandation aux touristes.

Dans la recommandation nous prenons en compte de façon très fine la notion de positionnement géographique, mais aussi la donnée temporelle. Nous apportons une solution à la question de la modélisation des déplacements touristiques en intégrant le lieu et le temps.

Remerciements

Je tiens à remercier chaleureusement les membres du laboratoire d'accueil de ma thèse UAG (CEREGMIA) pour leur aide et plus particulièrement le directeur du laboratoire, le Professeur Fred CELIMENE et mon directeur de Thèse, le Professeur Richard NOCK.

Références

1. AGRAWAL R. and SRIKANT R. (1995) Mining sequential patterns. In YU Philip S. and CHEN Arbee L. P.: Proceedings of the Eleventh International Conference on Data Engineering, March 6-10.
2. MANNILA Heikki and TOIVONEN Hannu (1996) Multiple uses of frequent sets and condensed representations (extended abstract) .
3. LALLICH Stephane et TEYTAUD Olivier (2000) Evaluation et validation de l'intérêt des règles d'association.
4. NICOLAS Pasquier et LOTFI Lakhal (2000) Data mining : Algorithmes d'extraction et de réduction des règles d'association dans les bases de donnes.
5. ZAKI Mohammed Javeed (2001) Spade : An efficient algorithm for mining frequent sequences. :Machine Learning, 42(1/2) : 3160.
6. MASSEGLIA Florent et TEISSEIRE M. et PONCELET Pascal (2004) Extraction de motifs séquentiels, Problèmes et méthodes.
7. PEI J., HAN J., MORTAZAVI-ASL B., WANG J., PINTO H., CHEN Q., DAYAL U., and HSU M.-C. (2004) Mining sequential patterns by pattern-growth : The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering.
8. GROZAVU Nistor, BENNANI Youns (2010) Classification collaborative non supervisé LPN UMP CNRS, 249-264 CAP.
9. BUCHET Nicolas, AUFAURE Marie-Aude, LECHEVALLIER Yves (2011) Construction et peuplement de structures hiérarchiques de concepts dans le domaine du e-tourisme : INRIA - 475-506, IFIA.
10 R. Agrawal, T. Imielinski et A. Swami. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD Intl. Conference on Management of Data, Washington, USA, pages 207–216, June 1993.
11 M. J. Zaki. Mining non-redundant association rules. Data Mining and Knowledge Discovery : An International Journal(DMKDJ'04). Pages 223–248, November 2004
12 M. Luxenburger. Implication partielles dans un contexte. Mathématiques et Sciences Humaines, 29(113) :35–55, 1991.
13 W.W. Armstrong. Dependency structures of database relationships. In IFIP Congress, pages 580–

ANNEXE 4

- [publication]
  Elisabeth, E. Sébastien, N. (2018). Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie : l'apport de la fouille de données. BigData et visibilité en ligne (pp. 51-58). Paris : Presses des Mines.

6 Novembre 2017

« Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie : l'apport de la fouille de données »

Erol ELISABETH, Nathalie SEBASTIEN

Université des Antilles et de la Guyane. Faculté de Droit et d'Economie de la Martinique. Campus de Schœlcher – Martinique FWI
BP 7209 – 97275 Schœlcher Cedex – Tél. : 0596.72.74.00 – Fax : 0596.72.74.03

Programmes de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique et odontologie : l'apport de la fouille de données

Erol Elisabeth[1] Nathalie Sébastien[2]

Résumé

Depuis la loi du 31 juillet 1991 portant réforme hospitalière puis l'arrêté du 20 septembre 1994 et la circulaire du 10 mai 1995, les établissements de santé publics et privés ont l'obligation de procéder à l'évaluation et à l'analyse de leur activité médicale mais également à la transmission de celle-ci aux services de l'Etat et de l'Assurance maladie. La mutation de leur environnement les a conduits à mettre en place le Programme de médicalisation des systèmes d'information (PMSI). Il permet de disposer d'informations quantifiées et standardisées, tenant compte notamment des pathologies et des modes de prise en charge, utilisées tant pour la gestion des établissements que pour le financement des établissements de santé et l'organisation de l'offre de soins.
Compte tenu du volume de données importants et en croissance exponentielle produits par les établissements de santé, la fouille de données permet, grâce aux modèles exploratoires développés, une utilisation optimale de celles-ci et une compréhension de l'activité hospitalière.
Nous présentons au travers d'illustrations l'apport de la fouille de données dans l'analyse des données du PMSI et la création de savoir.

Mots clés : PMSI, fouille de données, big data, motifs séquentiels, diagnostics principaux, diagnostics associés, atih

JEL Classification : I10

Abstract : Since the law of 31 July 1991 on hospital reform and subsequently the decree of 20 September 1994 and the circular of 10 May 1995, public and private health establishments have an obligation to evaluate and analyze their but also to the transmission of it to the State and Health Insurance services. The mutation of their environment led them to set up the Program of medicalization of the information systems (PMSI). It makes it possible to have quantified and standardized information, taking into account, in particular, pathologies and modes of care, used both for the management of establishments and for the financing of healthcare establishments and the organization of the provision of care.

Given the large and exponentially growing data produced by healthcare facilities, data mining makes optimal use of these data and an understanding of hospital activity possible thanks to the exploratory models developed.
We present by means of illustrations the contribution of the data mining in the analysis of the data of the PMSI and the creation of knowledge.

[1] CEREGMIA, Université des Antilles, Pôle Universitaire Régional de Martinique, email : erol.elisabeth@gmail.com

[2] CEREGMIA, Université des Antilles, Pôle Universitaire Régional de Martinique, email : ns.sebastien@gmail.com

1 Introduction

La fouille de données, qui est une des composantes du CRM (Customer Relationship Management ou Gestion de la Relation Client) analytique, largement déployé dans les entreprises (Lefebure, 2004), est le processus d'extraction à partir de données de connaissances intéressantes, non triviales, implicites, inconnues et potentiellement utiles (Fayad et al., 1996). Elle s'appuie sur des algorithmes issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) pour construire des modèles à partir des données stockées dans des entrepôts de données. Autrement dit, il s'agit de trouver des schémas pertinents (des « patterns » ou motifs en français) ou des ensembles similaires, selon des critères fixés au départ, et d'extraire de ces données les connaissances utiles. Cette connaissance permet d'aboutir à la découverte des motifs fréquents ou non fréquents. L'identification des règles les plus fréquentes, mais aussi dans certains cas des règles les moins fréquentes, a tout son sens selon la nature de la recherche.
La fouille de données diffère toutefois de la statistique qui fixe une hypothèse que les données permettent ou non de confirmer. Ainsi, dans la fouille de données il s'agit d'opter pour une démarche sans a priori et de chercher à faire émerger, à partir des données, des inférences dont il faudra évaluer la qualité. Cette approche permet donc au travers de la découverte des connaissances et de modèles de faire des analyses de prédictions.
Compte tenu du volume de données traitées, la principale problématique dans le domaine de la conception de systèmes liés à l'apprentissage est de produire des recommandations de qualité tout en minimisant, si possible, l'effort (les temps de calculs) requis.
Pour faire face à l'explosion du volume des données, un nouveau domaine technologique a vu le jour : le Big Data. Inventées par les géants du web, ces solutions sont dessinées pour offrir un accès en temps réel à des bases de données géantes.
L'objectif de cette communication est précisément de déterminer des modèles, établis à partir de « clusters » (regroupement d'items similaires dans des classes de données représentatives), au service de l'amélioration de la connaissance du client au sens générique, de la prédiction de ses comportements et de l'optimisation de l'offre proposée.
Ces modèles ayant vocation à être utilisés par des utilisateurs spécialistes du domaine de données, chercheurs en économie de la santé et sciences de gestion ou professionnels du secteur étudié, ces travaux de recherche mettent l'accent sur l'« utilisabilité » des environnements de fouille de données. Pour ce faire, cette approche conduite par une équipe pluridisciplinaire (informatique, santé publique et sciences de gestion) s'appuie sur deux pré-requis : l'utilisation des connaissances du domaine dans l'ensemble du processus de fouille, et l'amélioration de la compréhensibilité et de la confiance de l'utilisateur dans le modèle grâce une association de celui-ci dans sa création

Nous retiendrons comme champs pour notre réflexion un environnement « symptomatique » de ces questions : l'hôpital. Deux raisons essentielles justifient ce choix. En premier lieu, les établissements de santé publics et privés disposent de volumes importants d'informations quantifiées et standardisées sur leur activité médicale au travers du Programme de Médicalisation des Systèmes d'information (PMSI), qui s'inscrit dans la réforme hospitalière avec comme ambition l'optimisation de l'organisation de l'offre de soins et la réduction des inégalités de ressources entre les établissements de santé (Ordonnance n°96-346 du ordonnance du 24 avril 1996). Compte tenu du volume de données enregistrées, l'usage interne (établissements de santé) et externe (assurances maladie, services de l'Etat) de ces informations soulèvent des problématiques d'exploitation optimale. En second lieu,

l'hôpital est confronté depuis 2002 à des mutations importantes avec comme objectif une amélioration de la performance de la gestion hospitalière tout en maîtrisant les dépenses d'hospitalisation, qui sont en augmentation continue (Ordonnance n°2003-850 du 04 septembre 2003, Ordonnance n°2005-406 du 2 mai 2005, Loi du 21 juillet 2009 dite Loi Hôpital Patients Santé Territoires (HPST)). Ces réformes se sont accompagnées d'un nouveau mode d'allocation des ressources avec l'application d'une tarification à l'activité (T2A) pour les établissements dits MCO, qui affecte plus spécifiquement les établissements auparavant sous dotation globale, dont les activités seront désormais fonction de l'activité réalisée. Elle remet en cause les positions historiques entre le secteur privé et le secteur public « subventionné » (établissements publics de santé et établissements privés participant aux missions de service public) et obligent à un changement de paradigme, où la performance des établissements est centrale et soulève des questions de connaissance de son « client », le patient et de l'offre de soins proposée. Dans ce contexte, le PMSI prend toute sa place en tant qu'outil de gestion.

Pour conduire ces travaux de recherche les auteurs ont utilisé les bases de données PMSI MCO (Médecine, chirurgie, obstétrique et odontologie) gérées par l'ATIH, qui regroupent l'ensemble des RSS (résumés de sorties standardisées) des établissements de santé publics et privés de France pour les années 2007 à 2012 relatifs à des prises en charge de patients (activités médicales avec ou sans hébergement, cancérologie) et comportent plus de 132 millions d'enregistrements. Cette utilisation était conditionnée par un avis favorable de la CNIL et de l'ATIH[3].

Le plan de cette communication est le suivant : le domaine d'activité auquel s'appliquent ces travaux de recherche ainsi que les enjeux liés aux analyses des données produites, à usage interne et externe, sont présentés à la section 2 ; les outils développés pour les utilisateurs spécialistes du domaine des données sont décrits en section 3 ; la section 4 livre des éléments de conclusion.

2 Les enjeux de l'hôpital

L'hospitalisation française a connu depuis plusieurs années des changements majeurs, avec comme objectifs centraux la description de l'activité médicale et plus récemment la performance du système de soins, qui repose très largement sur l'hôpital. Les établissements de santé sont inscrits, par ces réformes, dans une dynamique d'amélioration de la performance de la gestion hospitalière qui doit concilier la maîtrise des dépenses, qui depuis 2002 augmentent de façon continue à plus de 2,5% par an en volume, la qualité des soins et la garantie de l'accès des soins à tous. Les principaux leviers de ces réformes sont l'évaluation de la qualité des soins et des pratiques, notamment grâce à la certification des établissements, la rénovation des règles de planification hospitalière avec l'instauration de la contractualisation (Ordonnance n°2003-850 du 4 septembre 2003), le déploiement d'une

[3] Agence Technique de l'Information sur l'Hospitalisation, instituée par le décret n°2000-1282 du 26 décembre 2000, repris dans le Code de la Santé publique aux articles R. 6113-33 et suivants, dont les missions portent essentiellement sur la collecte des données et de la gestion des référentiels, le calcul des tarifs et des coûts de prestation et la contribution au suivi et à l'analyse financière et médico-économique de l'activité des établissements de santé et l'évolution des classifications.

nouvelle organisation au sein de l'hôpital public axée sur le développement du pilotage économique avec la création des pôles et la réforme des instances institutionnelles (Ordonnance n°2005-406 du 2 mai 2005). Ces principes sont réaffirmés à chaque campagne tarifaire annuelle.

Ces réformes se sont accompagnées de l'instauration de la T2A, qui vise à une répartition plus équitable des moyens financiers entre établissements assurant des missions de service public (établissements publics et établissements privés assurant des missions de service public) et à une responsabilisation des acteurs, les recettes dépendant majoritairement de l'activité produite par les établissements. Il est à noter que celle-ci n'est pas neutre sur le secteur privé, dans la mesure où elle marque également un rapprochement des modes d'allocation des ressources entre secteur public et privé.

L'instauration de la T2A, qui s'appuie sur le PMSI, contraint les établissements de santé à changer de paradigme. En effet, la T2A vise à améliorer la transparence : elle assure en effet une plus grande transparence dans le financement des soins hospitaliers en liant le financement à la production des soins. Par ailleurs, elle est voulue comme un mécanisme « équitable » dans la mesure où on paie le même prix pour un même service pour tous les fournisseurs de soins. Cette équité dépend toutefois de la fiabilité de la classification de l'activité en groupes tarifaires : il est impératif que cette classification soit suffisamment fine, et les groupes suffisamment homogènes, pour que les établissements qui attirent systématiquement les patients les plus lourds ne soient pas pénalisés. Il faut également bien prendre en compte les facteurs exogènes liés au contexte local et que les établissements ne contrôlent pas, car ils peuvent influencer fortement les coûts. La T2A vise également à améliorer l'efficience, à la fois de chaque établissement individuellement et de l'ensemble du marché : elle introduit en effet une forme de compétition stimulant l'efficience dans un contexte où ces pressions compétitives étaient inexistantes jusqu'alors. Ceci suppose toutefois que les prix reflètent correctement les coûts des producteurs les plus efficients.

Il est donc désormais essentiel pour les établissements d'être en mesure de décrire leur activité médicale, mais également d'en comprendre les leviers et les incidences financières. Ceci passe par une connaissance du type de patients pris en charge (pathologies, durées du séjour, âge, lieu de résidence) avec une mise en perspective sur temps long.

En effet, à l'instar d'autres entreprises de services elles sont en concurrence avec d'autres acteurs, offreurs de soins publics ou privés, à un patient qui est un acteur à part entière du processus de soins. Le patient est le consommateur du service proposé (le soin), mais il va influer tout au long de la prise en charge sur le service produit. De fait, il existe un facteur d'incertitude du fait du caractère « intrinsèquement hétérogène » de la cible (Bancel-Charensol et Jougluex, 1997). Il apparaît donc difficile pour cette activité de service « obtenir de clients (…) des normes comportementales prévisibles et pratiquement impossible. Ces « participants à la production faiblement encadrés, introduisent des incertitudes fortes sur les processus et leurs résultats et sur leur qualité. Chacun revendique le droit de se comporter comme une exception » (Gadrey, 1996). Cette difficulté est accrue lorsque les producteurs de soins appréhendent les patients comme de des « exceptions ».

L'environnement contraint des établissements ne leur permet plus d'accepter comme une fatalité ces situations. Ils ont donc l'obligation de réduire cette incertitude par une meilleure connaissance de la « cible », le patient.

3 Les outils

Les modèles présentés répondent à deux des thématiques majeures en fouille de données, à savoir d'une part l'extraction d'informations et de connaissances de données informatisées et d'autres part la visualisation d'informations et les approches visuelles et interactives pour la représentation et l'extraction d'informations et de connaissances.

3.1. Structure de la base de données PMSI MCO

Tous les établissements de santé publics et privés (établissements dits MCO qui exercent des activités de médecine, chirurgie, obstétrique et odontologie mais aussi les activités ambulatoires et la cancérologie, établissements de soins de suite et de réadaptation (SSR), établissements exerçant des activités d'hospitalisation à domicile (HAD), établissements exerçant des activités de santé mentale). Cependant les modalités de recueil de l'information, de traitement de l'information médicalisée et les règles de tarification diffèrent selon le type d'établissements.

La base de données PMSI MCO est constituée de l'ensemble des données des établissements de santé publics et privés, remontées mensuellement selon un format normé et standardisé, le *Résumé de sortie standardisé* (RSS). Les RSS sont établis à l'occasion de tout séjour hospitalier dans un établissement public ou privé et retrace le séjour du patient (indication de (ou des) l'unité(s) médicale(s) de prises en charge, description de la prise en charge au sein de l'unité médicale au travers du *Résumé d'unité médicale* (RUM), etc.). Le RUM contient un nombre limité de rubriques, d'ordre administratif et médical, définies par un arrêté du 22 février 2008 modifié[4]. Ces informations sont codées selon des nomenclatures imposées (Figure 1).
Les informations ainsi recueillies font l'objet d'un traitement automatique aboutissant au regroupement des RSS en classes représentatives et cohérentes (homogènes) d'un point de vue médical et économique, appelées Groupes homogènes de malades (GHM).
Ce traitement algorithmique de **classification** permet de regrouper les RSS (*Groupage*) **en fonction du critère médical de prise en charge** (i.e. appareil fonctionnel ou motif notoire d'hospitalisation), appelé *Diagnostic principal*. Ce premier niveau de classement est appelé *Catégorie majeure de diagnostic* (CMD). Cette classification en GHM peut toutefois être altérée si un évènement médical majeur (acte médical), appelé *Acte classant* est intervenu au cours de la prise en charge du patient.

[4] Arrêté du 28 février 2008 modifié relatif au recueil et au traitement des données d'activité médicale et des données de facturation correspondantes, produites par les établissements de santé publics ou privés ayant une activité en médecine, chirurgie, obstétrique et odontologie, et à la transmission d'informations issues de ce traitement dans les conditions définies à l'article L. 6113-8 du code de la santé publique

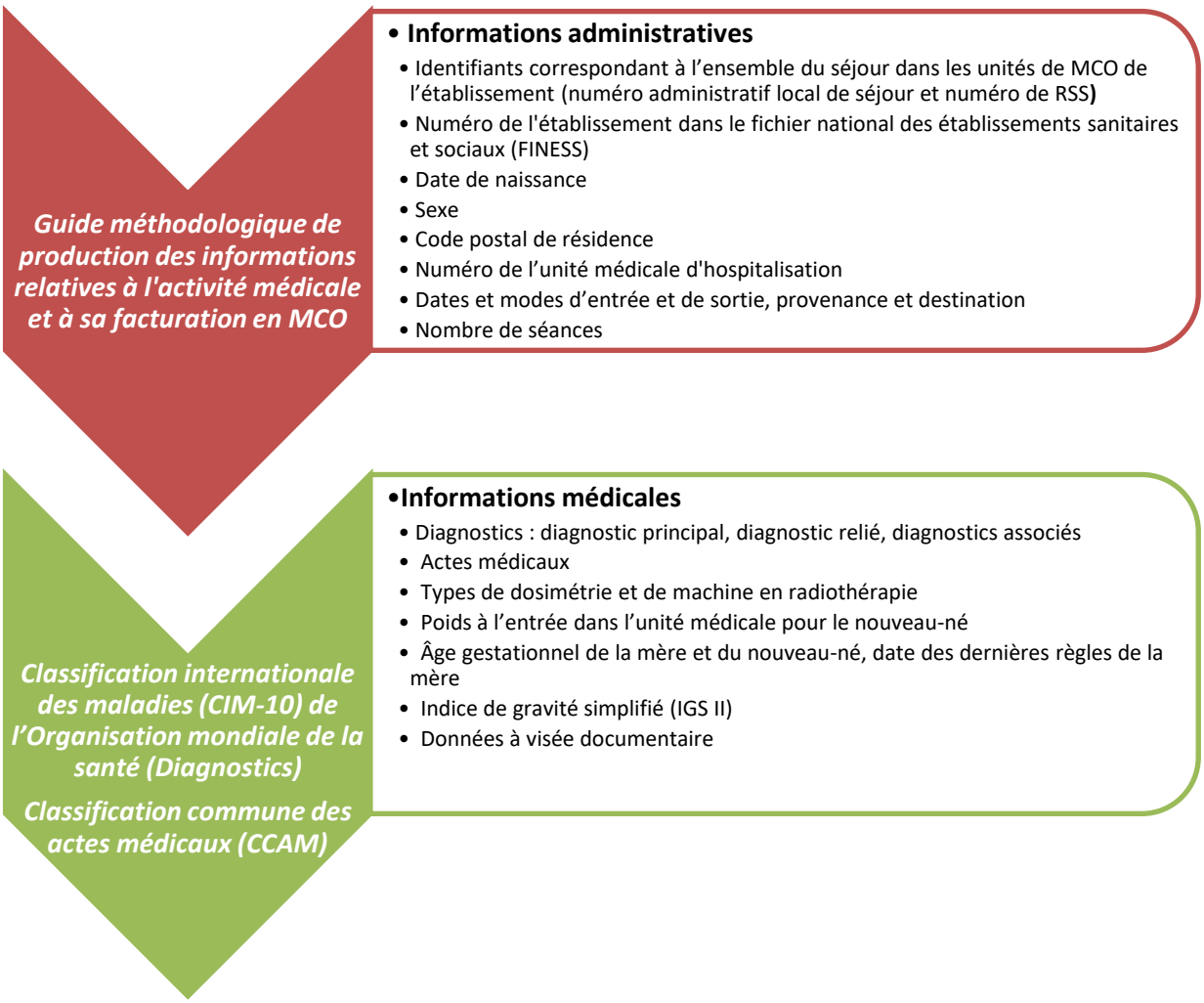


Figure 1 – Rubriques du RUM et nomenclatures de codification

Un deuxième niveau de classification en GHM est lié aux *Complications ou morbidités associés* (CMA), qui permettent de mettre en exergue la sévérité des cas pris en charge, et les incidences sur la durée de la prise en charge (*durée de séjour*). Ces CMA sont également appelés Diagnostic associés.

L'âge peut être également un facteur déterminant dans la classification en GHM dans la mesure où l'âge peut être soit de nature à accroitre le niveau de sévérité soit de nature à influer sur la survenance de la pathologie (exemple : limite d'âge de 18 ans pour les affections touchant fréquemment les enfants…).

Dans cette communication un focus particulier est fait sur les diagnostics principaux, les diagnostics associés et sur les durées moyennes de séjour.

3.2. Visualisation des données en 3D interactives

Pour cette visualisation des données en 3D interactive, nos travaux s'appuient sur le couplage de deux outils, Mathematica et Wolfram SystemModeler.

Mathematica est un puissant moteur de calcul symbolique et numérique. Il est doté de capacités graphiques et de visualisation exceptionnelles, un langage robuste de programmation et de développement et un environnement intuitif pour l'édition scientifique et technique. C'est le logiciel de calcul de référence dans le monde de l'enseignement et de la

recherche.
Wolfram SystemModeler offre une nouvelle approche à la modélisation et la simulation de systèmes complexes. Le couplage avec Mathematica permet de bénéficier d'une suite logicielle fonctionnelle pour notre étude.
Le premier module développé permet pour une année de visualiser et de comparer pour deux établissements les durées de séjours (Figure 2). Ce module paramétrable permet de sélectionner un AGE, une durée de séjours MAXIMUM et affiche pour DEUX établissements les durées de séjour moyennes : les diagnostics principaux sont cumulés

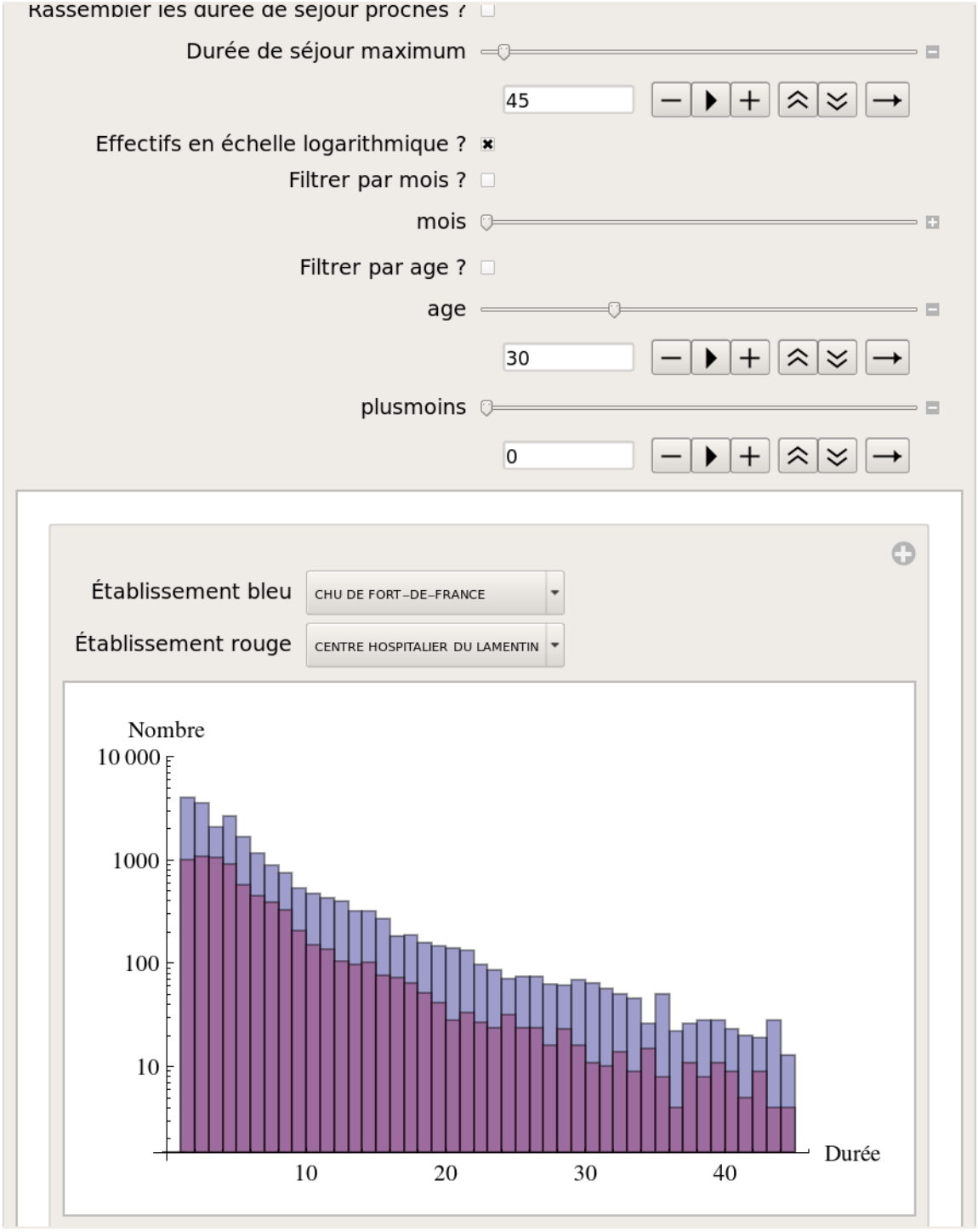


Figure 2 – Comparaison des durées de séjours entre le Centre Hospitalier du Lamentin et le CHU de Fort-de-France en 2007

Le deuxième module développé permet de visualiser avec une granularité mensuelle les durées de séjours pour un établissement et pour une année donnée (Figure 3). Il est possible de ne prendre en compte que des durées maximales.

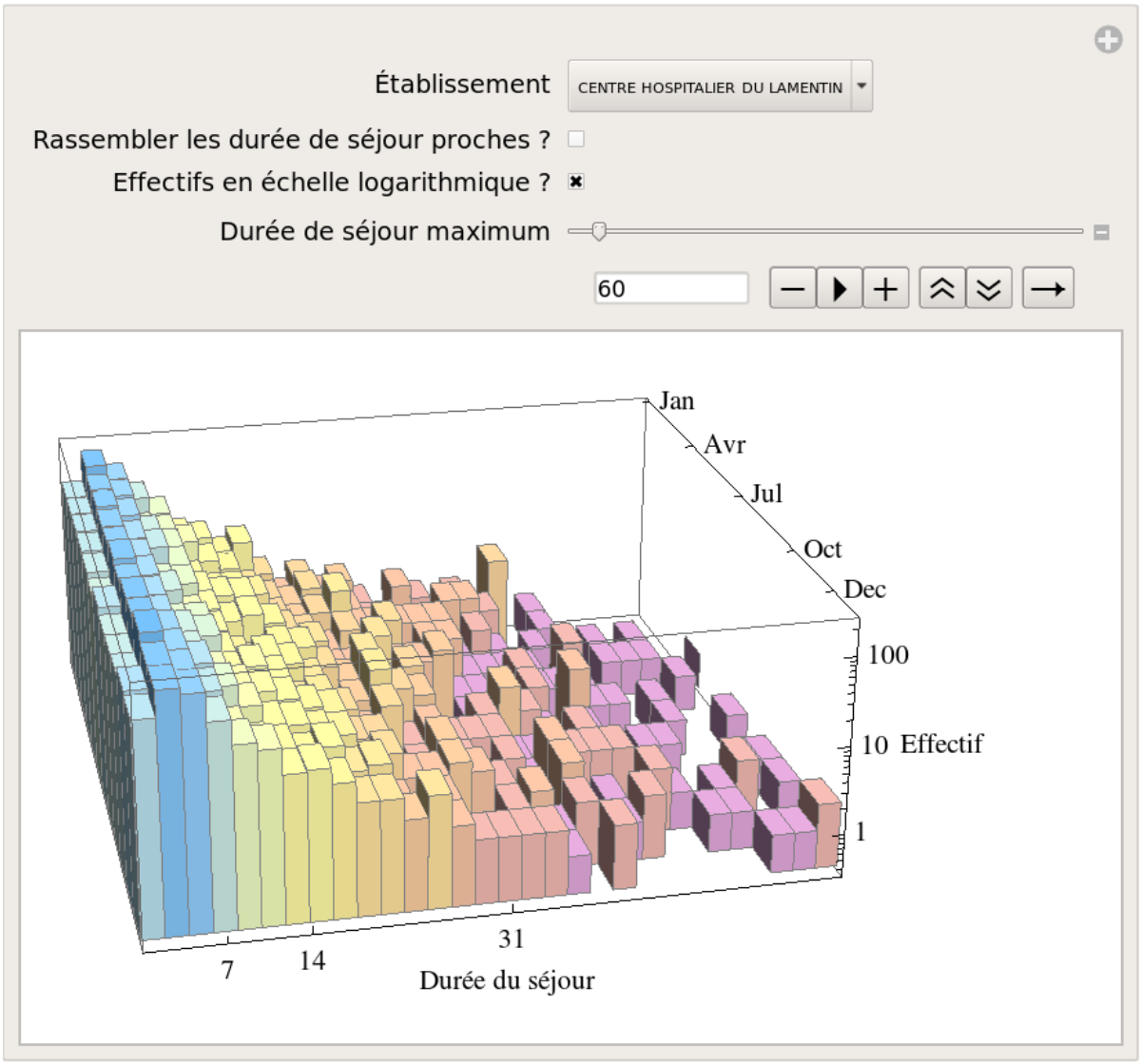


Figure 3 – Liste des durées de séjour du Centre hospitalier du Lamentin en 2007

Le troisième module permet de comparer côte à côte les durées de séjour de deux établissements (Figure 4). Les comparaisons visuelles sont facilitées. Il est possible d'appliquer un filtre pour les durées de séjours maximales, pour un mois donné et pour un âge donné. Le module fonctionne pour une année donnée. Les échelles sont différentes, l'utilisateur doit donc être vigilant lors de l'analyse des résultats. Le graphique est interactif et un passage de la souris sur un histogramme affiche l'effectif.

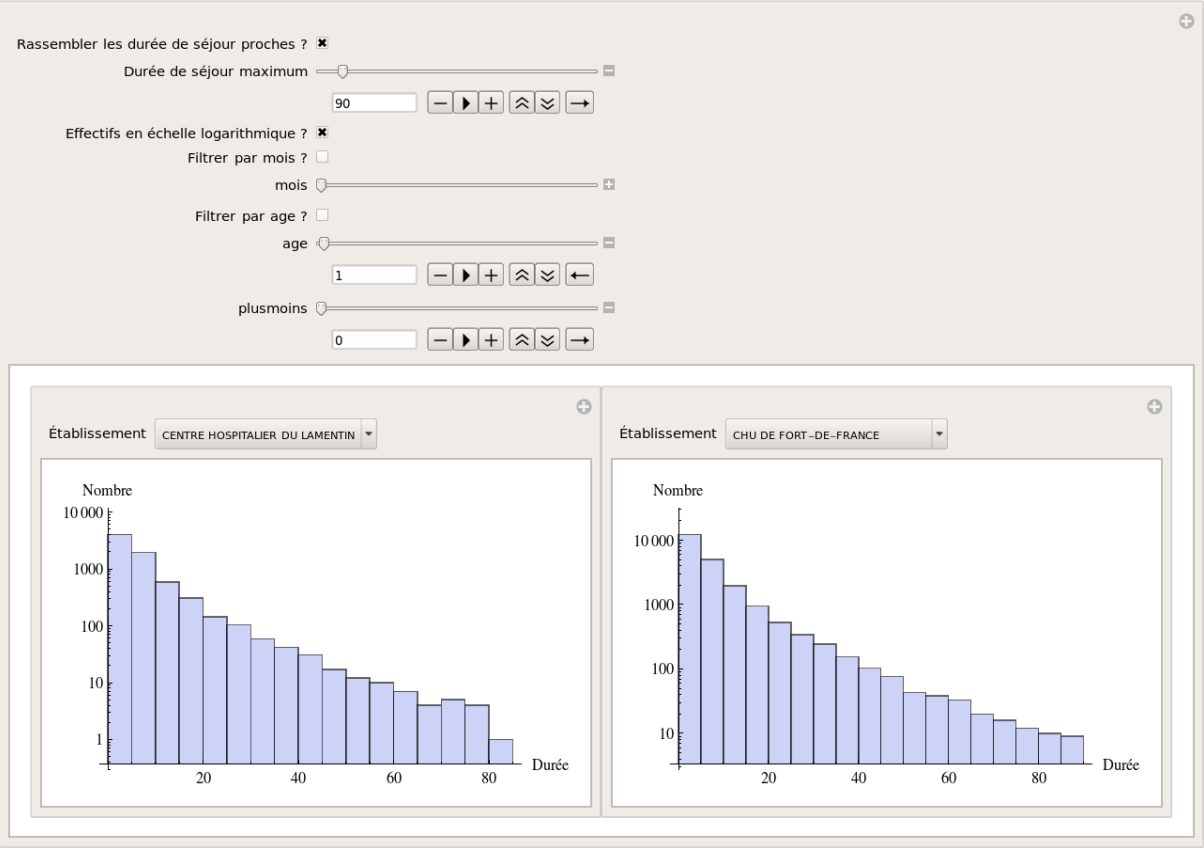


Figure 4 – Comparaison des durées de séjour entre le Centre hospitalier du Lamentin et le CHU de Fort-de-France en 2007

3.3. Analyse des Diagnostics principaux et des diagnostics associés

L'outil suivant permet après avoir sélectionné une année puis un établissement de lister, d'afficher sous forme graphique et d'exporter les diagnostics principaux et les diagnostics associés en fonction de leur effectif (Figure 5a et s.).

Dans un souci d'optimisation de l'utilisabilité de l'outil, celui-ci offre la possibilité à l'utilisateur, s'il le souhaite, d'exporter toutes ces données au format Excel (personnalisation des analyses et des présentations...).

Analyse DP/DAS(s) 2008

CENTRE HOSPITALIER DU LAMENTIN - Martinique

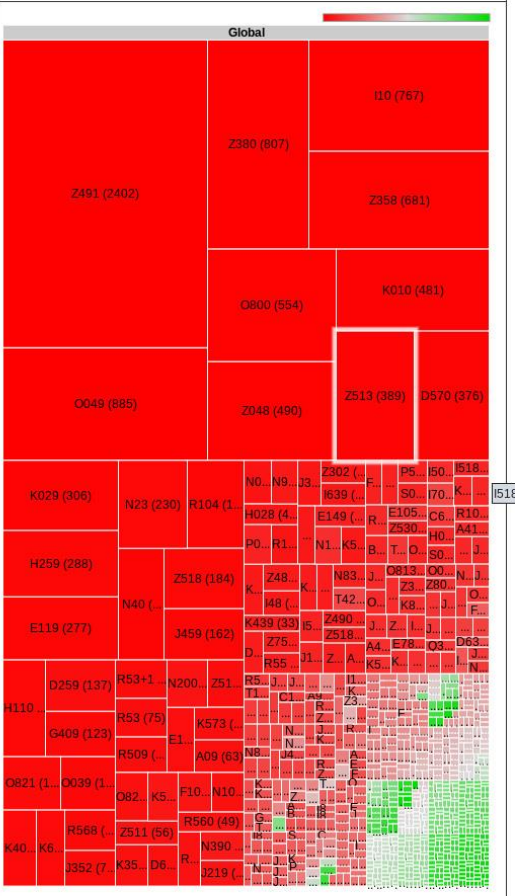
Diagnostic principal	Diagnostic principal (Nombre d'occurences)
Z491 Dialyse extra-corporelle O049 Avortement médical complet ou sans précision, sans complication Z380 Enfant unique, né à l'hôpital I10 Hypertension essentielle (primitive) Z358 Surveillance d'autres grossesses à haut risque O800 Accouchement (unique) spontané par présentation du sommet Z048 Examen et mise en observation pour d'autres raisons précisées K010 Dents incluses Z513 Transfusion sanguine (sans mention de diagnostic) K029 Carie dentaire, sans précision D570 Anémie à hématies falciformes [anémie drépanocytaire] avec crises N40 Hyperplasie de la prostate E119 Diabète sucré non insulino-dépendant, sans complication R104 Douleurs abdominales, autres et non précisées H259 Cataracte sénile, sans précision J459 Asthme, sans précision H110 Ptérygion Z518 Autres formes précisées de soins médicaux N23 Colique néphrétique, sans précision G409 Épilepsie, sans précision O039 Avortement spontané complet ou sans précision, sans complication O821 Accouchement (unique) par césarienne d'urgence K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène) K625 Hémorragie de l'anus et du rectum R568 Convulsions, autres et non précisées J352 Hypertrophie des végétations adénoïdes R53+1 Malaise R53 Malaise et fatigue R509 Fièvre, sans précision Z515 Soins palliatifs D259 Léiomyome de l'utérus, sans précision E109 Diabète sucré insulino-dépendant, sans complication K573 Diverticulose du côlon, (sans perforation ni abcès) A09 Autres gastroentérites et colites d'origine infectieuse ou non précisée O820 Accouchement (unique) par césarienne programmée D649 Anémie, sans précision K359 Appendicite aiguë, sans précision N10 Néphrite tubulo-interstitielle aiguë R101 Douleur localisée à la partie supérieure de l'abdomen Z511 Séance de chimiothérapie pour tumeur R560 Convulsions fébriles J219 Bronchiolite (aiguë), sans précision N049 Syndrome néphrotique, sans précision N200 Calcul du rein F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance R11 Nausées et vomissements J329 Sinusite (chronique), sans précision K566 Occlusions intestinales, autres et sans précision Z302 Stérilisation H028 Autres affections précisées des paupières I639 Infarctus cérébral, sans précision D069 Carcinome in situ du col de l'utérus, sans précision K590 Constipation K429 Hernie ombilicale, (sans occlusion ni gangrène) I48 Fibrillation et flutter auriculaires N390 Infection des voies urinaires, siège non précisé N948 Autres affections précisées des organes génitaux de la femme et du cycle menstruel Z488 Autres soins de contrôle chirurgicaux précisés K439 Autres hernies de la paroi abdominale antérieure, (sans occlusion ni gangrène) D509 Anémie par carence en fer, sans précision Z7588 Autres difficultés liées aux installations médicales et autres soins de santé, non classées ailleurs K859 Pancréatite aiguë, sans précision N47 Hypertrophie du prépuce, phimosis et paraphimosis R55 Syncope et collapsus (sauf choc)	

Figure 5a – Diagnostics principaux du Centre Hospitalier du Lamentin en 2008

Analyse DP/DAS(s) 2008

CHU DE FORT-DE-FRANCE - Martinique

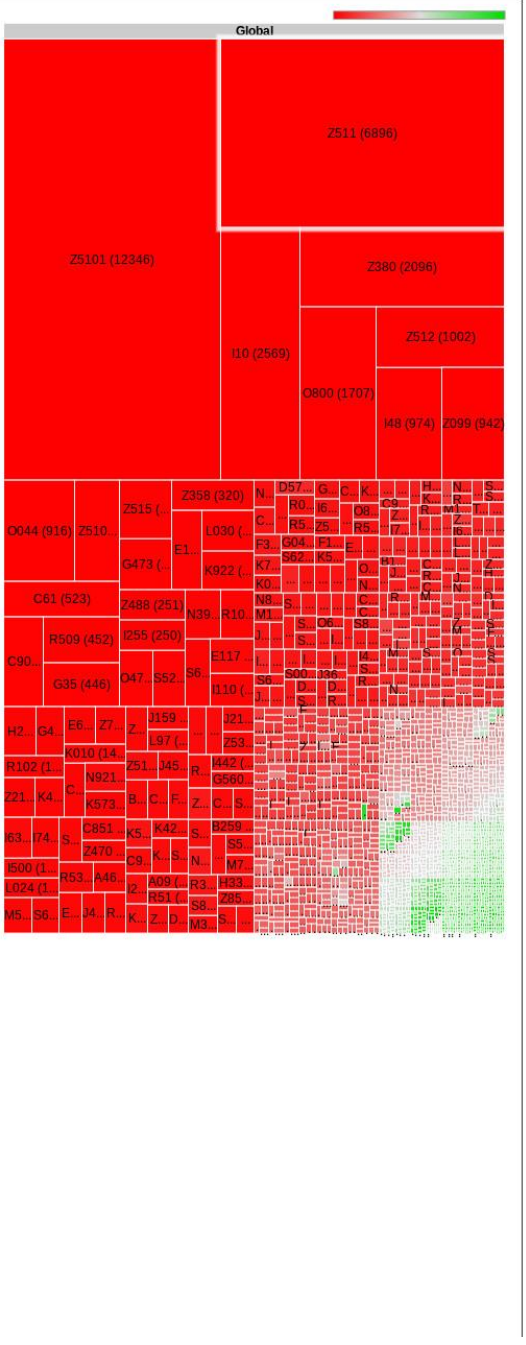
Diagnostic principal	Diagnostic principal (Nombre d'occurences)
Z5101 Séance d'irradiation Z511 Séance de chimiothérapie pour tumeur I10 Hypertension essentielle (primitive) Z380 Enfant unique, né à l'hôpital O800 Accouchement (unique) spontané par présentation du sommet I48 Fibrillation et flutter auriculaires Z512 Autres formes de chimiothérapie Z099 Examen de contrôle après traitement pour une affection, sans précision O044 Avortement médical incomplet, sans complication Z5100 Séance de préparation à une irradiation C61 Tumeur maligne de la prostate R509 Fièvre, sans précision C900 Myélome multiple G473 Apnée du sommeil Z358 Surveillance d'autres grossesses à haut risque G35 Sclérose en plaques Z515 Soins palliatifs L030 Phlegmon des doigts et des orteils K922 Hémorragie gastro-intestinale, sans précision E107 Diabète sucré insulino-dépendant, avec complications multiples Z488 Autres soins de contrôle chirurgicaux précisés O470 Faux travail avant 37 semaines entières de gestation [Menace d'accouchement prématuré] S5250 Fracture fermée de l'extrémité inférieure du radius R104 Douleurs abdominales, autres et non précisées S610 Plaie ouverte de(s) doigt(s) (sans lésion de l'ongle) I255 Myocardiopathie ischémique I110 Cardiopathie hypertensive, avec insuffisance cardiaque (congestive) N390 Infection des voies urinaires, siège non précisé H259 Cataracte sénile, sans précision G409 Épilepsie, sans précision R102 Douleur pelvienne et périnéale Z21 Infection asymptomatique par le virus de l'immunodéficience humaine [VIH] K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène) E660 Obésité due à un excès calorique Z751 Sujet attendant d'être admis ailleurs, dans un établissement adéquat C504 Tumeur maligne du quadrant supéro-externe du sein K573 Diverticulose du côlon, (sans perforation ni abcès) L024 Abcès cutané, furoncle et anthrax d'un membre M511 Atteintes d'un disque lombaire et d'autres disques intervertébraux avec radiculopathie S663 Lésion traumatique du muscle et du tendon extenseurs d'un autre doigt au niveau du poignet et de la main I743 Embolie et thrombose des artères des membres inférieurs S611 Plaie ouverte de(s) doigt(s) avec lésion de l'ongle Z470 Soins de contrôle impliquant l'enlèvement d'une plaque et autre prothèse interne (orthopédique) de fixation A46 Érysipèle E105 Diabète sucré insulino-dépendant, avec complications vasculaires périphériques I500 Insuffisance cardiaque congestive E117 Diabète sucré non insulino-dépendant, avec complications multiples J450 Asthme à prédominance allergique R100 Syndrome abdominal aigu Z018 Autres examens spéciaux précisés C851 Lymphomes à cellules B, sans précision J159 Pneumopathie bactérienne, sans précision R53+1 Malaise L97 Ulcère du membre inférieur, non classé ailleurs K010 Dents incluses J459 Asthme, sans précision B182 Hépatite virale chronique C C509 Tumeur maligne du sein, sans précision F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance I633 Infarctus cérébral dû à une thrombose des artères cérébrales D500 Anémie par carence en fer secondaire à une perte de sang (chronique) K802 Calcul de la vésicule biliaire (sans cholécystite)	

Figure 5b – Diagnostics principaux du CHU de Fort-de-France en 2008

Il est possible d'afficher et d'exporter les diagnostics associés en cliquant sur les diagnostics principaux. Ainsi dans notre exemple il est possible de quantifier les naissances pour l'année 2008 (Figure 6).

Figure 6 – Diagnostics associés pour le Centre Hospitalier du Lamentin en 2008 pour le diagnostic principal « Accouchements »

Cette même analyse est faite pour le CHU de Fort-de-France pour l'année 2008 (Figure 7). L'outil permet donc des comparaisons entre établissements sur des prises en charge similaires.

Analyse DP/DAS(s) 2008

CHU DE FORT-DE-FRANCE - Martinique

Diagnostic principal	Diagnostic associé (Nombre d'occurences)
Z5101 Séance d'irradiation	Z370 Naissance unique, enfant vivant
Z511 Séance de chimiothérapie pour tumeur	(1512 occurence(s))
I10 Hypertension essentielle (primitive)	
Z380 Enfant unique, né à l'hôpital	Pas de diagnostique associé
O800 Accouchement (unique) spontané par présentation du sommet	(72 occurence(s))
I48 Fibrillation et flutter auriculaires	
Z512 Autres formes de chimiothérapie	O471 Faux travail à ou après la 37ème semaine entière de gestation
Z099 Examen de contrôle après traitement pour une affection, sans précision	Z370 Naissance unique, enfant vivant
O044 Avortement médical incomplet, sans complication	(39 occurence(s))
Z5100 Séance de préparation à une irradiation	
C61 Tumeur maligne de la prostate	Z358 Surveillance d'autres grossesses à haut risque
R509 Fièvre, sans précision	(37 occurence(s))
C900 Myélome multiple	
G473 Apnée du sommeil	Z340 Surveillance d'une première grossesse normale
Z358 Surveillance d'autres grossesses à haut risque	Z370 Naissance unique, enfant vivant
G35 Sclérose en plaques	(26 occurence(s))
Z515 Soins palliatifs	
L030 Phlegmon des doigts et des orteils	Z371 Naissance unique, enfant mort-né
K922 Hémorragie gastro-intestinale, sans précision	O358 Soins maternels pour d'autres anomalies et lésions foetales (présumées)
E107 Diabète sucré insulino-dépendant, avec complications multiples	(8 occurence(s))
Z488 Autres soins de contrôle chirurgicaux précisés	
O470 Faux travail avant 37 semaines entières de gestation [Menace d'accouchement prématuré]	
S5250 Fracture fermée de l'extrémité inférieure du radius	Z340 Surveillance d'une première grossesse normale
R104 Douleurs abdominales, autres et non précisées	O48 Grossesse prolongée
S610 Plaie ouverte de(s) doigt(s) (sans lésion de l'ongle)	Z370 Naissance unique, enfant vivant
I255 Myocardiopathie ischémique	(6 occurence(s))
I110 Cardiopathie hypertensive, avec insuffisance cardiaque (congestive)	
N390 Infection des voies urinaires, siège non précisé	Z370 Naissance unique, enfant vivant
H259 Cataracte sénile, sans précision	O757 Accouchement par voie vaginale après une césarienne
G409 Épilepsie, sans précision	O48 Grossesse prolongée
R102 Douleur pelvienne et périnéale	(4 occurence(s))
Z21 Infection asymptomatique par le virus de l'immunodéficience humaine [VIH]	
K409 Hernie inguinale, (unilatérale ou) sans précision(, sans occlusion ni gangrène)	Z358 Surveillance d'autres grossesses à haut risque
E660 Obésité due à un excès calorique	O48 Grossesse prolongée
Z751 Sujet attendant d'être admis ailleurs, dans un établissement adéquat	O700 Déchirure obstétricale du périnée, du premier degré
C504 Tumeur maligne du quadrant supéro-externe du sein	(2 occurence(s))
K573 Diverticulose du côlon, (sans perforation ni abcès)	
L024 Abcès cutané, furoncle et anthrax d'un membre	O364 Soins maternels pour mort intra-utérine du foetus
M511 Atteintes d'un disque lombaire et d'autres disques intervertébraux avec radiculopathie	N751 Abcès de la glande de Bartholin
S663 Lésion traumatique du muscle et du tendon extenseurs d'un autre doigt au niveau du poignet et de la main	Z371 Naissance unique, enfant mort-né
	(1 occurence(s))
I743 Embolie et thrombose des artères des membres inférieurs	
S611 Plaie ouverte de(s) doigt(s) avec lésion de l'ongle	
Z470 Soins de contrôle impliquant l'enlèvement d'une plaque et autre prothèse interne (orthopédique) de fixation	
A46 Érysipèle	
E105 Diabète sucré insulino-dépendant, avec complications vasculaires périphériques	
I500 Insuffisance cardiaque congestive	
E117 Diabète sucré non insulino-dépendant, avec complications multiples	
J450 Asthme à prédominance allergique	
R100 Syndrome abdominal aigu	
Z018 Autres examens spéciaux précisés	
C851 Lymphomes à cellules B, sans précision	
J159 Pneumopathie bactérienne, sans précision	
R53+1 Malaise	
L97 Ulcère du membre inférieur, non classé ailleurs	
K010 Dents incluses	
J459 Asthme, sans précision	
B182 Hépatite virale chronique C	
C509 Tumeur maligne du sein, sans précision	
F102 Troubles mentaux et du comportement liés à l'utilisation d'alcool : syndrome de dépendance	
I633 Infarctus cérébral dû à une thrombose des artères cérébrales	
D500 Anémie par carence en fer secondaire à une perte de sang (chronique)	
K802 Calcul de la vésicule biliaire (sans cholécystite)	
Z538 Acte non effectué pour d'autres raisons	

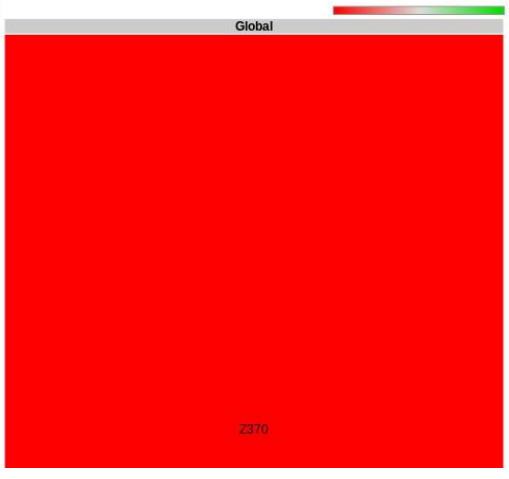


Figure 7 – Diagnostics associés pour le CHU de Fort-de-France en 2008 pour le diagnostic principal « Accouchements »

3.4. Durées de séjours des patients

Ce module permet d'extraire, à partir de résumés de données, les durées de séjours différenciés pour tous les diagnostics ou pour un diagnostic principal spécifique (Figure 8).

CEREGMIA - PREG Santé (Chaire PMSI) : Durées de séjours

2007

Guadeloupe

C.H.U. DE POINTE-A-PITRE/ABYMES

Tous les DP

Exécutée en 0.000869 secondes.

Age	Durée moyenne Janvier	Durée moyenne Février	Durée moyenne Mars	Durée moyenne Avril	Durée moyenne Mai	Durée moyenne Juin	Durée moyenne Juillet	Durée moyenne Août	Durée moyenne Septembre	Durée moyenne Octobre	Durée moyenne Novembre	Durée moyenne Décembre
0	H:11.6349 F:9.2254	H:7.7581 F:10.7826	H:17.6000 F:8.4386	H:7.2545 F:19.9250	H:9.0167 F:11.1951	H:18.1364 F:9.6735	H:10.5000 F:9.4615	H:9.9467 F:12.7451	H:11.8947 F:14.0862	H:8.1176 F:13.5000	H:5.4203 F:7.7833	H:12.6129 F:10.1000
1	H:1.8333 F:2.7826	H:2.3077 F:1.7619	H:2.3600 F:2.0500	H:1.4783 F:1.9375	H:20.0500 F:1.9444	H:2.2593 F:4.8261	H:3.7143 F:1.8235	H:1.1250 F:1.6500	H:2.0952 F:2.0588	H:2.6452 F:2.2800	H:2.0333 F:2.2000	H:2.5500 F:1.6087
2	H:2.0769 F:2.4667	H:1.7143 F:6.3750	H:1.8800 F:2.6111	H:3.3333 F:5.6500	H:3.9048 F:2.0667	H:1.2632 F:3.4118	H:2.6000 F:3.7895	H:2.0500 F:1.9375	H:1.4211 F:2.6364	H:2.5926 F:0.8000	H:2.3636 F:1.5625	H:1.4706 F:2.9412
3	H:1.2500 F:1.4000	H:2.5556 F:2.4444	H:1.4615 F:2.7333	H:2.7500 F:0.7500	H:0.8750 F:0.8750	H:2.1667 F:3.8889	H:0.5000 F:2.9286	H:1.2857 F:3.0833	H:1.3750 F:1.7143	H:1.3077 F:2.2308	H:2.3333 F:8.6667	H:1.7143 F:1.7222
4	H:2.0625 F:4.2000	H:1.8667 F:3.5000	H:1.6667 F:1.3333	H:4.2857 F:2.8333	H:15.7273 F:1.5556	H:1.9091 F:2.0000	H:1.8889 F:1.3333	H:1.6364 F:1.2857	H:2.7778 F:2.5556	H:3.6667 F:2.9000	H:3.5714 F:1.2500	H:1.2308 F:1.8750
5	H:1.3571 F:2.2500	H:3.1111 F:1.2308	H:1.2000 F:0.7500	H:1.4444 F:1.7778	H:3.7778 F:1.5556	H:2.0000 F:1.3333	H:1.1875 F:4.1429	H:6.6000 F:1.9375	H:1.6667 F:1.4444	H:1.0000 F:1.1429	H:3.9000 F:3.8333	H:0.7143 F:6.5556
6	H:2.3333 F:3.2000	H:0.5000 F:9.4000	H:3.4545 F:2.6667	H:2.0833 F:0.7500	H:2.8000 F:1.1000	H:2.5714 F:1.3333	H:5.5000 F:2.1667	H:2.1333 F:0.8000	H:1.6000 F:2.0000	H:1.6923 F:2.2000	H:1.9000 F:6.1111	H:2.6667 F:1.5000
7	H:2.2000 F:2.6000	H:4.9000 F:2.0833	H:1.6250 F:1.8571	H:2.1667 F:11.1429	H:1.1667 F:1.7000	H:1.5000 F:1.7500	H:2.0000 F:1.6000	H:1.4000 F:2.5714	H:3.2857 F:1.8571	H:2.7333 F:3.2667	H:2.7143 F:3.4286	H:2.2857 F:4.4444
8	H:3.4000 F:2.0000	H:2.1429 F:1.5000	H:2.4286 F:2.1250	H:3.1818 F:1.0000	H:7.8333 F:13.5714	H:1.0000 F:2.0000	H:2.0000 F:1.2000	H:2.7778 F:3.0000	H:3.3077 F:2.7500	H:2.2727 F:2.7143	H:1.8889 F:7.0000	H:2.9091 F:2.8571
9	H:9.3333 F:2.5833	H:2.5714 F:3.0000	H:2.7000 F:0.7143	H:7.6667 F:1.2222	H:2.8750 F:2.0000	H:1.9000 F:7.8889	H:1.3636 F:3.2500	H:3.0000 F:3.5714	H:3.0000 F:1.7500	H:2.9000 F:3.0000	H:5.8571 F:1.8750	H:0.5000 F:2.6000
10	H:3.0833 F:3.7778	H:0.8333 F:3.0000	H:3.3333 F:3.8571	H:1.3750 F:1.5000	H:1.6154 F:2.5000	H:4.4444 F:1.2000	H:1.5000 F:1.2000	H:1.2000 F:1.2000	H:2.2500 F:6.0000	H:4.6667 F:2.5000	H:2.5714 F:4.2000	H:7.5000 F:3.2000
11	H:3.7000 F:1.4286	H:1.6667 F:4.3750	H:2.0000 F:5.3000	H:5.4286 F:1.0000	H:2.5000 F:1.8000	H:4.5000 F:5.8333	H:2.0000 F:5.2222	H:5.3636 F:3.0000	H:1.6667 F:2.1667	H:3.7273 F:2.1818	H:4.7500 F:2.0000	H:0.5000 F:14.6667
12	H:5.2727 F:1.8571	H:2.1111 F:2.5000	H:2.5714 F:1.0000	H:1.4000 F:2.5556	H:1.3000 F:3.3333	H:1.5000 F:3.0000	H:2.2000 F:5.7143	H:2.4615 F:2.6000	H:0.8889 F:0.6000	H:4.6429 F:1.5000	H:2.5455 F:2.8333	H:3.0000 F:2.0000
13	H:2.3636 F:1.5000	H:7.0000 F:2.2500	H:1.7143 F:1.0000	H:2.0000 F:3.3333	H:7.6364 F:2.3333	H:1.9091 F:2.3750	H:3.3333 F:1.6667	H:2.2000 F:5.0000	H:3.6667 F:2.6667	H:5.6667 F:3.0000	H:4.4000 F:4.0000	H:1.8889 F:5.0000
14	H:2.3125 F:1.8750	H:1.8750 F:3.2000	H:1.7778 F:7.5000	H:0.8333 F:2.5000	H:2.6923 F:3.2727	H:4.3333 F:2.5000	H:3.8000 F:2.2000	H:2.7778 F:3.1250	H:1.1250 F:2.1429	H:2.1111 F:2.7143	H:2.4000 F:2.1111	H:3.8571 F:4.4545
15	H:2.0000 F:2.1111	H:1.7500 F:3.6250	H:3.1429 F:3.6667	H:0.8750 F:1.5714	H:1.5714 F:3.1667	H:1.6000 F:3.1429	H:6.1429 F:1.3750	H:1.3750 F:3.5556	H:3.5385 F:2.5714	H:6.8750 F:1.1111	H:1.5000 F:2.6000	H:3.0833 F:2.5714
16	H:5.4000 F:4.7273	H:1.0000 F:5.8000	H:6.5000 F:1.8750	H:2.2857 F:2.7500	H:3.4444 F:3.0000	H:5.6000 F:2.0000	H:4.7778 F:2.4444	H:2.0000 F:4.0000	H:1.5000 F:2.5833	H:1.0000 F:2.8182	H:4.4286 F:3.9167	H:1.5000 F:2.7500
17	H:2.8889 F:4.8750	H:3.3333 F:4.6000	H:2.6000 F:4.4167	H:1.6000 F:1.9000	H:4.2000 F:1.7778	H:2.0000 F:3.2727	H:3.5714 F:3.1765	H:3.8889 F:5.0000	H:2.7778 F:2.1667	H:6.4167 F:1.9000	H:5.0909 F:2.7273	H:1.8571 F:3.7647
18	H:2.3333 F:4.5600	H:16.2000 F:5.3684	H:1.3333 F:4.2500	H:2.4286 F:2.8000	H:5.3333 F:3.0909	H:2.8571 F:4.2727	H:1.5000 F:3.8750	H:2.0556 F:3.0000	H:3.6667 F:2.7143	H:9.5556 F:4.4375	H:1.0000 F:3.0000	H:4.8571 F:6.3846
19	H:3.9231 F:3.6000	H:3.6667 F:3.0769	H:3.0000 F:3.0000	H:5.6250 F:5.2143	H:5.5455 F:2.3333	H:0.7500 F:3.4118	H:4.2667 F:7.2105	H:3.3333 F:2.9231	H:3.9286 F:4.0000	H:4.6000 F:4.4286	H:2.3333 F:4.1538	H:1.5000 F:4.2632
20	H:3.5714	H:6.1818	H:4.2857	H:12.2500	H:2.3333	H:8.9000	H:1.6667	H:5.7500	H:2.5000	H:1.8750	H:6.5000	H:3.5714

Figure 8 – Durées de séjours Hommes et Femmes tous diagnostics principaux confondus pour le CHU de Pointe-à-Pitre en 2007

3.5. Durées de séjours des patients hors Région

Ce module permet d'extraire, à partir de résumés de données, les durées de séjours

Ce module permet d'afficher, pour une année et pour un établissement, les diagnostics principaux des patients pris en charge non originaire de la région d'implantation de l'établissement (Figure 9).

Au regard de l'organisation de l'offre de soins, ce module permet de conduire des analyses sur le « taux de fuite » pour les patients résidents français. L'indicateur « Taux de fuite » correspond au rapport entre le nombre de séjours de la zone géographique sélectionnée pris en charge en dehors de cette zone sur le nombre total de séjours issus de la zone sélectionnée (une zone géographique qui ne contient aucun établissement hospitalier présente un taux de fuite de 100 %). En outre, ce module est une des bases à un module destiné à quantifier le tourisme médical. Il est aussi important pour générer la cartographie diagnostics principaux/Origine géographique des patients.

2007

Guadeloupe

C.H.U. DE POINTE-A-PITRE/ABYMES

Tous les DP

Exécutée en 0.000975 secondes.

Age	Durée moyenne Janvier	Durée moyenne Février	Durée moyenne Mars	Durée moyenne Avril	Durée moyenne Mai	Durée moyenne Juin	Durée moyenne Juillet	Durée moyenne Août	Durée moyenne Septembre	Durée moyenne Octobre	Durée moyenne Novembre	Durée moyenne Décembre
0	H:8.1892 F:7.3590	H:6.1068 F:7.5244	H:11.6383 F:6.1800	H:5.9286 F:12.4189	H:7.3953 F:7.6667	H:12.3148 F:6.7976	H:7.8019 F:6.7579	H:7.7478 F:9.4250	H:8.8977 F:10.4667	H:6.4375 F:9.2697	H:4.5565 F:6.4457	H:9.4300 F:7.3645
1	H:1.8000 F:2.3103	H:2.0000 F:1.4231	H:2.2692 F:1.9524	H:1.4167 F:1.7778	H:19.0952 F:1.9474	H:1.9063 F:4.0714	H:3.6207 F:1.7778	H:1.0000 F:1.5714	H:2.0000 F:1.9474	H:2.5143 F:2.0000	H:1.8684 F:2.1176	H:2.5455 F:1.6087
2	H:1.8235 F:2.4667	H:1.5000 F:5.1000	H:1.5484 F:2.0435	H:2.9583 F:5.1364	H:3.9048 F:1.6842	H:1.1304 F:2.7619	H:2.4375 F:3.4286	H:1.7826 F:1.6000	H:1.3810 F:2.6364	H:2.3750 F:0.8750	H:2.2917 F:1.5294	H:1.4762 F:2.8333
3	H:1.2500 F:1.4000	H:2.0909 F:2.4444	H:1.1875 F:2.7333	H:2.3333 F:0.7500	H:0.7778 F:0.7778	H:1.9286 F:3.6000	H:0.3333 F:2.5625	H:1.2000 F:2.5333	H:1.3529 F:1.6000	H:1.2353 F:2.0000	H:2.3684 F:7.3636	H:1.6957 F:1.7368
4	H:2.0625 F:3.0000	H:1.7647 F:3.1111	H:1.3636 F:1.0000	H:3.3333 F:2.4286	H:15.7273 F:1.4000	H:1.9091 F:2.0000	H:1.8889 F:1.0909	H:1.5000 F:1.2857	H:2.3636 F:2.7000	H:3.4000 F:2.7273	H:3.4000 F:1.3000	H:1.2667 F:1.8750
5	H:1.3333 F:2.2500	H:2.8000 F:1.2308	H:1.2500 F:0.7333	H:1.4444 F:1.4545	H:3.6000 F:1.4000	H:2.0000 F:0.8000	H:1.1667 F:2.9000	H:4.7143 F:1.9375	H:1.5000 F:1.4444	H:1.0000 F:1.1429	H:3.9000 F:3.8333	H:0.7143 F:6.5556
6	H:2.1333 F:3.2000	H:0.3750 F:9.4000	H:2.9231 F:2.6667	H:1.9231 F:0.6000	H:2.4706 F:1.1000	H:2.4000 F:1.3333	H:4.4000 F:1.8000	H:2.0625 F:0.6667	H:1.6364 F:2.0000	H:1.6923 F:2.2000	H:1.9000 F:6.1111	H:2.6667 F:1.5000

Figure 9 – Durées de séjours Hommes et Femmes tous diagnostics principaux confondus pour le CHU de Pointe-à-Pitre en 2007

4 Conclusion

Nous nous sommes attachés dans cette communication à présenter des outils établis à partir de clusters au service de l'amélioration de la connaissance du client au sens générique, de la prédiction de ses comportements et de l'optimisation de l'offre proposée, dans un environnement spécifique qu'est l'hôpital. Nous avons tenté de mettre en évidence en quoi la fouille de données peut être un levier puissant dans la connaissance du patient, au bénéfice duquel l'hôpital créé un service (prise en charge, au sens global). Nous avons identifié des outils, qui proposé aux spécialistes des données du domaine (chercheurs, établissements de santé, ARS[5], DHOS, etc.), permettent d'optimiser l'offre de soins.

Cette recherche n'a cependant pas permis d'intégrer à ce stade l'articulation entre connaissance de l'activité et performance des établissements au sens rapport qualité-coût des prestations délivrées par les établissements. Il conviendra de proposer des outils qui prendront en compte cette deuxième composante du PMSI, en mettant l'accent sur la valorisation des séjours et l'incidence financière du parcours patients en se basant notamment sur les algorithmes de groupage utilisés par l'ATIH. D'autres outils mettant l'accent sur l'optimisation des dépenses seront également étudiés (consommations médicamenteuses…).

[5] Agences Régionales de Santé, instituées par la Loi du 21 juillet 2009 dite « Loi HPST » (art. 118). Elles sont le pilier de la réforme de santé et ont comme mission d'assurer un pilotage unifié de la santé en région, de mieux répondre aux besoins et d'accroître l'efficacité du système de santé.

Bibliographie

AGRAWAL R., SRIKANT R., « Mining Sequential Patterns », Proceedings of the 11th International Conference on Data Engineering (ICDE'95)

ATIH (Agence de Traitement de l'information hospitalière), *Guide méthodologique de production des informations relatives à l'activité médicale et à sa facturation en médecine, chirurgie, obstétrique et odontologie (http://www.atih.sante.fr/index.php?id=0002300005FF)*

Direction de l'hospitalisation et de l'organisation des soins (DHOS) [2006], « Présentation de la réforme de la gouvernance hospitalière », Juillet

Loi n°91-748 du 31 juillet 1991 portant réforme hospitalière, *Journal Officiel de la République française*, 2 août 1991

Loi n°2009-879 du 21 juillet 2009 dite Loi Hôpital Patients Santé Territoires (HPST), *Journal Officiel de la République française*, 21 juillet 2009

Ordonnance n°96-346 du 24 avril 1996 portant réforme de l'hospitalisation publique et privée, Inter Bloc n°3/96, tome XV

Ordonnance n°2005-406 du 2 mai 2005, *Journal Officiel de la République française*, 3 mai 2005

BANCEL-CHARENSOL L., JOUGLEUX M., (1997), « Un modèle d'analyse des systèmes de production dans les services », *Revue Française de gestion*, Mars-Avril-Mai, pp. 71-81

GADREY, J. (1996), « Services : la productivité en question », *Sociologie économique*, Desclée de Brouwer

SAMPIERI-TEISSIER N., SAUVIAT, I. (2002), « Les évolutions du positionnement des acteurs du système hospitalier : le cas de la situation du patient-usager-client », *Revue Electronique du RECEMAP (http://www.unice.fr/recemap/contenurevue/Articles.html)*

Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares http://ic2012.crc.jussieu.fr/papiers/IC2012_51.pdf

http://chazard.org/emmanuel_/contenuperso/articles/memoire_emmanuelchazard_version_2006-08-02.pdf (P36)

http://en.wikipedia.org/wiki/Sparkline

http://code.google.com/intl/fr/apis/chart/interactive/docs/gallery/imagesparkline.html

http://code.google.com/intl/fr/apis/chart/interactive/docs/gallery/treemap.html