# Système d'assistance aux recherches épidémiologiques et de surveillance des maladies : Modélisation Booléenne

**SABRI Mohammed, ATMANI Baghdad**
mailto:ram.sabri@gmail.com, mailto:atmani.baghdad@univ-oran.dz

Equipe de recherche « Simulation, Intégration et Fouille de données (SIF)»
Laboratoire d'Informatique d'Oran (LIO)
Département d'Informatique, Faculté des Sciences, Université d'Oran,
BP 1524, El-M'Naouer, 31000, Oran, Algérie.

**Mots clés:**
Fouille de données, Entrepôts de données, Extraction des connaissances à partir des données (ECD), Graphe d'induction, SIPINA, Extraction de règles, CASI, Santé publique, Epidémiologie, Pharmacie.
**Keywords:** Data Mining, Data Warehouse, Knowledge Discovery in Databases (KDD), Induction Graph, SIPINA, Rules Discovery, CASI, Public health, Epidemiology, Pharmacy.
**Palabras clave :**
Minería de datos, Data Warehouse, Extracción de conocimiento a partir de datos (ECD), Inducción Gráfico, SIPINA, Extracción de reglas, CASI, Salud pública, Epidemiología, Farmacia.

## Résumé

Cet article concerne notre contribution dans le domaine de la santé publique et de l'épidémiologie à travers la conception d'un Système d'Assistance à la Recherche Epidémiologique et de Surveillance des Maladies (SARESM). SARESM apporte aux différents acteurs de la santé publique une assistance à l'établissement de politiques sanitaires, notamment en matière de planification d'acquisition des produits pharmaceutiques, selon la distribution géographique de leur utilisation. Cette distribution géographique est établie par rapport à une mesure pathologique. Notre contribution dans ce domaine est de fournir des modèles de prédiction des maladies chroniques. Ces modèles sont basés sur des techniques de fouille de données, en l'occurrence une nouvelle approche de modélisation booléenne des graphes d'induction, inspirée du principe de la machine CASI (Cellular Automata for Symbolic Induction). Le but, après une modélisation booléenne des règles de prédiction épidémiologique, est double : d'une part affiner le suivi par une fouille de donnée orchestrée par CASI, et d'autre part réduire la complexité de gestion des connaissances, ainsi que le temps de réponse.

# 1 Introduction

Le développement des systèmes d'informations et des technologies des ordinateurs a permis l'automatisation des activités dans tous les domaines du monde réel, ce qui a entraîné un accroissement rapide de l'information disponible, le développement des entrepôts de grands volumes de données, et finalement, l'émergence du Data Mining. Le but de ce dernier est d'extraire des connaissances disponibles, jusque là cachées au sein des données, pour être exploitées dans différents domaines tels que le commerce, les banques, la santé publique, etc. Le domaine de la santé publique est la principale préoccupation de toute la population mondiale et s'appuie sur plusieurs disciplines, pour le bien-être de tous.

Nous étudierons, dans le cadre de ce travail, une démarche de la fouille de données dans le monde de la santé publique, où notre contribution concerne l'extraction de modèles pour la surveillance des maladies chroniques, basés sur une exploitation des données réelles de ventes en détail de médicaments dans des pharmacies privées.

L'étude a été menée de la manière suivante :

- Etude des maladies chroniques suivantes : l'asthme, l'hyper tension artérielle et le diabète.
- Elaboration du modèle de règles de prédiction épidémiologique par la fouille de données sur des données réelles des pharmacies (région ouest de l'Algérie), basée sur une nouvelle approche de modélisation booléenne des graphes d'induction inspirée du principe de la machine CASI [2] (Cellular Automata for Symbolic Induction).
- Validation des résultats des différentes expérimentations en collaboration avec le laboratoire de bio-statistiques de l'université d'Oran.

# 2 Travaux connexes

La fouille de données dans le domaine pharmaceutique, grâce aux informations collectées[1] à partir des enregistrements des données, a ciblé deux entités potentielles pour obtenir des modèles de fouilles répondant à différents aspects : la première entité concerne le comportement du *patient* (ou client) et la deuxième concerne l'étude du *produit*. On peut classer les différents modèles, obtenus à partir de l'application des différentes techniques de fouille de données, selon les objectifs de chaque étude :

*La santé publique* : Souvent confrontée aux pressions du marketing. Elle essaie de suivre la prescription médicale pour déterminer quels médicaments les médecins préfèrent pour des diagnostics spécifiques. Ceci va permettre de créer les portraits de prescription détaillés de chaque médecin [6].

*La gestion et la surveillance de certaines pathologies* : plusieurs études ont été menées pour comprendre et surveiller certaines pathologies, par exemple l'asthme [4], l'avancement dans des thérapies de cancer [15], et enfin le développement des systèmes de bio-surveillance qui peuvent être utilisés pour identifier les manifestations des maladies.

Ces études, qui sont menées afin de comprendre et de surveiller les maladies, sont étroitement liées à la compréhension de la vente de certains produits [4] pour des patients choisis. Autrement dit, les patients sont ciblés au préalable et l'étude surveillera leur régime de consommation de certains produits dans une période donnée. Ceci permettra ainsi de déterminer le comportement des patients vis-à-vis de la maladie étudiée.

[1] Données provenant des ventes directes aux clients et de sources telles que les hôpitaux et les rapports médicaux.

*Système d'aide à la prescription médicale* : L'objectif est d'améliorer la prescription médicale chez les praticiens, en se basant sur les résultats d'une fouille de données sur les dossiers médicaux et les données des pharmacies. L'apport de ce système d'aide à la prescription est multiple : d'une part, il permet de sensibiliser les médecins prescripteurs sur des recommandations (lettres, séminaires, etc.), d'autre part, de l'encourager à la prescription des produits qui peuvent être moins chers et plus efficaces [16], et enfin, l'exploration des réactions de médicaments défavorables afin d'alerter les médecins sur les effets nuisibles potentiels [5].
*La gestion optimale des stocks et sa modélisation* : Gérer un stock multi-produits signifie obtenir le bon mix de produits, en termes de quantité et de disponibilité au moment voulu, tout en prenant en compte le comportement du client.
Il est nécessaire de répondre aux questions primordiales liées à la gestion des stocks : « combien commander ? » et « quand commander ? ». Plusieurs modèles sont obtenus, cités dans [3], pour étudier la dépendance d'achat dans la vente. A cette fin, des algorithmes génétiques ont été utilisés dans la classification d'inventaire, et les réseaux de neurones ont été employés pour la classification pour des unités de stockage [3].
*La réalisation des bénéfices dans les ventes de médicaments* : Enfin, le volet commercial est un domaine très consommateur des techniques de fouille de données afin de réaliser des profits financiers. Des entités commerciales importantes dans le domaine pharmaceutique ont utilisé les techniques de fouille pour augmenter leurs chiffres d'affaires d'une manière significative et ainsi réaliser des bénéfices remarquables, en observant certains attributs de produits. C'est le cas de *Pharma*, la chaîne de pharmacies au Japon [7].

# 3 Démarche proposée

La démarche adoptée pour la conception et la réalisation de SARESM est issue de la démarche globale du processus d'extraction de connaissances à partir des données.

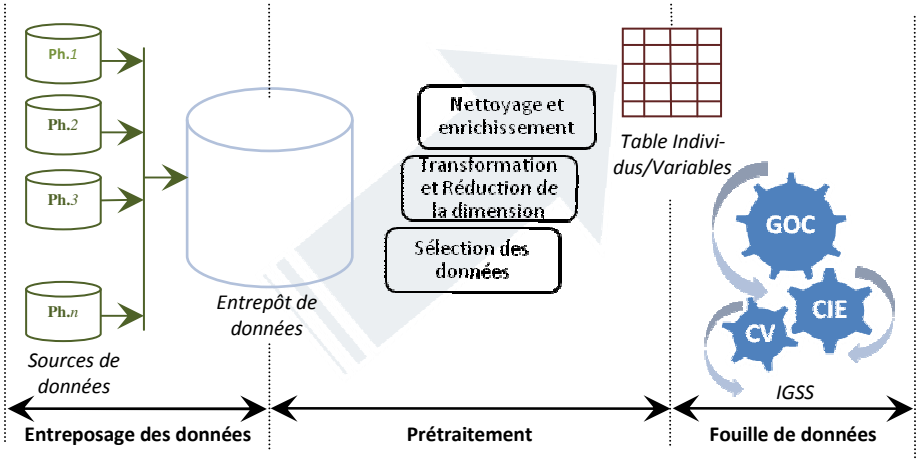



*Figure 1 : Schéma global de la démarche proposée pour SARESM*

On retient principalement dans cette étape de l'étude, trois phases dérivées du processus global : entreposage des données, prétraitement et fouille de données.

## 3.1 L'entreposage des données

La première phase dans notre démarche est la conception de l'entrepôt des données, dans le but d'obtenir une source unique de données pour effectuer les tâches de fouille.

### 3.1.1 Les sources des données

Nos sources de données sont les enregistrements de ventes en détail des officines pharmaceutiques privées.
Les pharmacies algériennes se sont équipées depuis plusieurs années des outils logiciels et matériels pour leur gestion commerciale, depuis plus d'une dizaine d'années où il a été impératif d'utiliser les outils informatiques pour prendre en charge des systèmes de ventes par convention (Assurances Sociales). Tous ces régimes prennent *entièrement* en charge les assurés présentant des *maladies chroniques*.
D'autre part, des logiciels de gestion commerciale de pharmacies ont été développés. Ces derniers continuent à aider les pharmaciens à enregistrer toutes les opérations d'achats et de ventes au quotidien, la tenue de plusieurs autres fonctions mais aussi la génération de rapports. Dans cette étude, nous avons utilisé les données de ventes générées par ce type de logiciels.
Ce type de logiciels a été choisi pour des raisons multiples : parmi elles, nous soulignons l'esprit collaboratif des concepteurs et des utilisateurs de ce logiciel, pour améliorer la phase de l'entreposage. D'autres sources de données[2] ont été utilisées, notamment, les référentiels médicaux ; nous avons choisi celui qui concerne l'affectation des médicaments, à travers leurs dénominations communes internationales, par rapport aux classes de maladies (Les médicaments METFORAL et GLUCOPHAGE ont la même DCI : METFORMINE CHLORHYDRATE soit la classe ANTI-DIABETIQUE), l'endroit où la pharmacie exerce et les informations relatives aux dates. Les informations pertinentes[3] pour la sélection de ces sources externes ont été élaborées suite aux orientations du laboratoire de bio-statistique de l'université d'Oran et en collaboration avec des médecins et pharmaciens.

### 3.1.2 L'entrepôt de données

Les entrepôts de données ou « Data Warehouses » permettent de stocker l'ensemble des données nécessaires à l'interrogation, l'analyse et la prise de décision. Les entrepôts sont alimentés par des extractions de données appliquées à des sources d'informations telles que les bases de données, les fichiers plats, etc. [9].
L'architecture de l'entrepôt de données de SARESM (Fig. 2.), s'articule autour de trois axes :
**Intégration :** Dans cette première étape, le travail consiste à extraire et regrouper les données provenant des différentes bases de données des pharmacies privées et des sources externes [11]. Ces bases de données sont supportées par un même SGBD relationnel, elles sont identiques du point de vue de leurs structures, et elles sont installées dans des sites différents où aucune connexion n'existe entre ces sites, ni l'existence d'un système centralisé. Les bases de données sources récupérées (fichiers) sont codifiées et stockées dans le système de fichiers (le système global).
**Construction :** Elle consiste à extraire les données pertinentes, puis à les recopier dans l'entrepôt de données [11]. Par conséquent, l'entrepôt de SARESM constituera une collection centralisée de données matérialisées et historiques, disponibles pour les applications de fouilles. Les données relatives aux ventes des médicaments et les caractéristiques liées aux produits vendus, sont prises en compte dans ce cadre d'étude, et les autres données, telles que les achats, sont négligées.

---

[2] Sources de données externes au type de logiciel choisi.
[3] Les informations pertinentes concernent principalement le choix des paramètres à prendre en compte dans nos analyses.

**Structuration :** Cette étape consiste à réorganiser les données, dans des magasins afin de supporter efficacement la fouille de données [13] ; nous créons, dans ce cadre, un magasin de données concernant uniquement les informations relatives aux maladies chroniques choisies par rapport aux ventes en détail et les caractéristiques des patients faisant partie des variables indispensables pour la fouille de données.

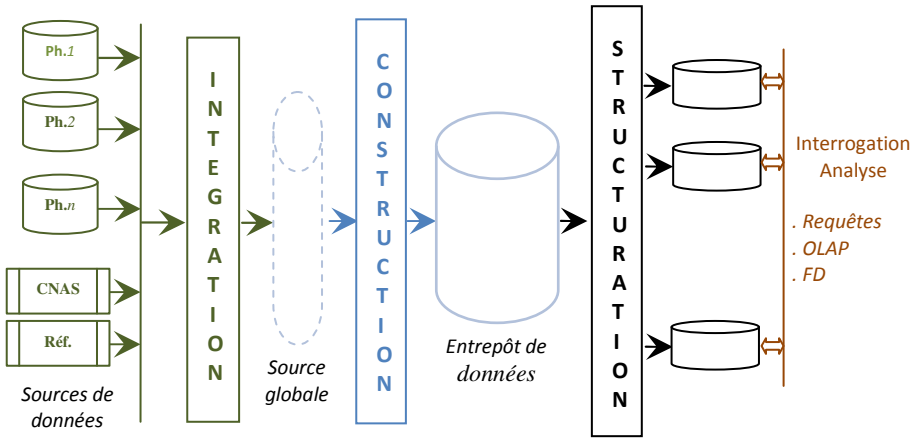



*Figure 2 : Architecture de l'entrepôt de données de SARESM*

### 3.1.3 Modélisation de l'entrepôt de données

La modélisation multidimensionnelle se base sur deux concepts fondamentaux : le concept de fait (ou mesure) et le concept de dimension. Un fait est une information qui contient les données observables, que l'on possède sur un sujet, et que l'on veut étudier. Il peut s'agir du prix des ventes, la quantité vendue, etc. Une dimension représente un axe d'analyse selon lequel on veut étudier des données observables (les faits) [8]. Ces concepts de base ont permis de définir trois schémas classiques. Le premier est le schéma en étoile [10]; il se compose d'une table de faits centrale et d'un ensemble de tables de dimension. Le deuxième est le schéma en flocon de neige ; il correspond à un schéma en étoile dans lequel les dimensions ont été normalisées. Le troisième est le schéma en constellation ; ce schéma permet de faire coexister plusieurs tables de faits qui partagent ou non des dimensions communes.
L'entrepôt de données de SARESM repose sur le modèle en étoile (Fig. 3.) et stocke toutes les informations liées aux ventes en détail, les données concernant les produits et les données concernant la localisation des officines. Les données sources disponibles sont des données commerciales (données simples des enregistrements de ventes) pour effectuer des recherches médicales (épidémiologiques), et suivant les études réalisées par F. Ravat et al. dans [12], nous avons opté pour une modélisation multidimensionnelle classique pour l'entrepôt de données de SARESM. Les données de l'entrepôt sont comme suit :

- La table des faits « VENTES » ; contient les mesures telles que la quantité vendue (brut), le prix de vente, date de péremption, date d'achat, etc.
- Les tables de dimension :
    - La localisation des officines choisies « LOCALISATIONS_OFFICINES ».
    - Une dimension date « DATES ».
    - La table des produits manipulés « PRODUITS » qui comprend le nom commercial, le dosage, etc.

- Les spécialités et les sous spécialités des différents produits existants dans la base « SPECIALITES_PDT » et « SOUS_SPECIALITES_PDT ».
- Les « DCI » et les maladies correspondantes.
- Les laboratoires fabricants les produits « LABORATOIRES ».
- Les patients « ASSURES ». Pris sous un anonymat total : seulement les informations concernant les dates de naissance, le sexe et la situation familiale sont reprises dans l'entrepôt.

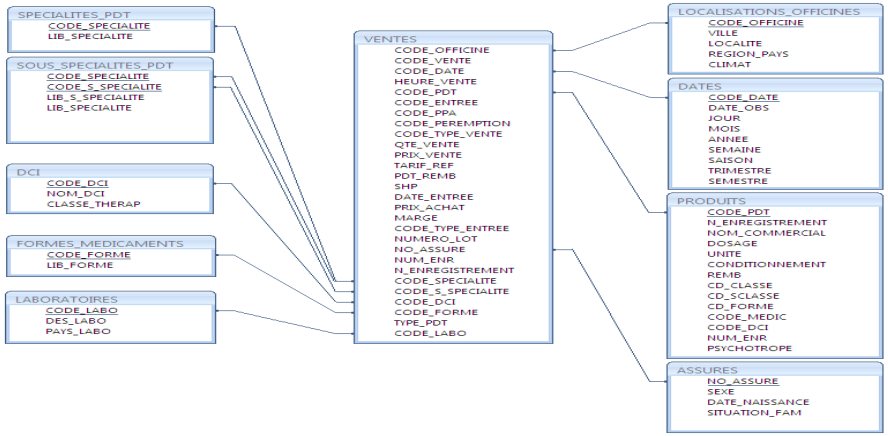



*Figure 3 : Le modèle en étoile adoptée dans SARESM*

## 3.2 Le prétraitement

Les données issues de l'entrepôt sont très variées et ne sont pas nécessairement toutes exploitables par les techniques de fouille de données [14]. La majorité des techniques utilisées ne traitent que les tableaux de données sous formes lignes/colonnes. L'objectif est de préparer des tables lignes/colonnes, autrement dit, des tables individus/variables (cf. Tableau 1), obtenue par les étapes suivantes :

*Tableau 1 : Exemple d'une table individus/variables cible*

| | *Localité* | *Saison* | *Age* | *Sexe* | *Maladie* |
|---|---|---|---|---|---|
| $\omega_1$ | Oran | Hiver | Jeune | Homme | Asthme |
| $\omega_2$ | Oran | Hiver | Jeune | Femme | Asthme |
| $\omega_3$ | Tlemcen | Hiver | Jeune | Homme | Diabète |
| ••• | … | … | … | … | … |

**Sélection des données** : Elle s'effectue sur les données qui existent déjà dans l'entrepôt de données et qui sont sous forme tabulaire. Il s'agit ensuite d'appliquer des filtres qui nous permettront de sélectionner un sous-ensemble de lignes ou de colonnes. La sélection des données repose sur les informations suivantes :

- A partir de la table des faits « VENTES », nous prendrons la mesure concernant la quantité vendue « QTE_VENTE ». Cette dernière sera prise en premier dans son état « brut » et sera par la suite agrégée selon les dimensions choisies.
- De la table « LOCALISATIONS_OFFICINES », l'attribut « LOCALITE ».
- La dimension date « DATES » afin d'effectuer la fouille sur un intervalle de temps. Dans notre cas nous procédons par la période « MOIS ».
- De la table « DCI », nous prenons l'attribut « CLASSE_THERAP » où les différentes maladies sont présentes. Un filtre est ainsi appliqué à ce niveau afin de ne garder que les enregistrements relatifs aux maladies citées.
- Enfin, les patients, pris sous une discrétion totale, présents dans la table « ASSURES » et à partir desquels nous prenons les attributs sexe « SEXE » et âge (recommandations des experts). Etant donné que l'âge est inexistant pour le moment, la donnée date de naissance « DATE_NAISSANCE » sera prise (nous verrons dans la partie transformation l'obtention de l'attribut âge).

**Nettoyage et enrichissement des données** : Une étape de nettoyage des données est indispensable afin de traiter les données manquantes (suppression d'enregistrements). Par contre, l'enrichissement par des sources externes a été effectué lors de la création de l'entrepôt des données.

**Transformation et réduction de la dimension** : Il s'agit de transformer un attribut A en un autre A' qui serait plus approprié aux objectifs de l'étude. Dans cette étape, l'unique transformation qui est effectuée concerne les informations relatives aux dates de naissance des assurés. Ces dernières sont des valeurs numériques, prises par différence entre date de naissance et date de vente. Ensuite, la donnée obtenue (*âge)* est mise dans une tranche d'âge. Différentes méthodes existent comme la discrétisation ; dans notre cas, la transformation est établie par rapport aux tranches d'âge recommandées par les experts du domaine.

## 3.3 Fouille de données (IGSS)

Comme nous l'avons signalé dans la section 3, après l'entreposage et le prétraitement des données, nous entamons la phase de fouille de données. Pour la modélisation booléenne des règles de prédiction épidémiologique, nous avons opté pour IGSS comme module de fouille de données. IGSS (Fig.1.) a été développé[4] afin d'intégrer le principe cellulaire [2] et enrichir l'environnement graphiques de la plateforme WEKA[5].

La démarche adoptée par ce système s'appuie sur la méthode cellulaire d'extraction de règles à partir des données nommée CASI [1] et qui se base sur les graphes d'induction (GI) produits par la méthode SIPINA [17][18]. Il prend en entrée l'échantillon d'apprentissage sous forme de table individus/variables afin de fournir en sortie une base de règles de prédiction épidémiologiques en binaire en appliquant le principe booléen de la machine cellulaire.

[4] Equipe de recherche « Simulation, Intégration et Fouille de données (SIF) », Laboratoire d'Informatique d'Oran (LIO), Université d'Oran.

[5] WEKA (Waikato Environment for Knowledge Analysis) est un outil de fouille de données open-source (licence GNU) développé en Java, http://www.cs.waikato.ac.nz/ml/weka.

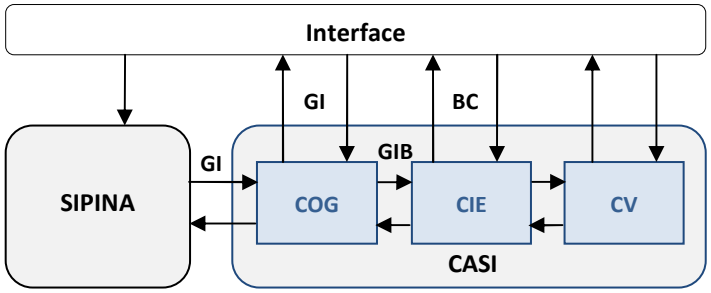



*Figure 4 : Architecture générale du système IGSS*

### 3.3.1 La machine CASI

CASI [1] est un automate cellulaire qui simule le principe de fonctionnement de base d'un moteur d'inférence. Il assure l'optimisation du graphe d'induction (GI), la génération des règles conjonctives (« *Si* cond1 [*ET* $cond_2$] … [*ET* $cond_n$] *Alors* conclusion »), et la validation du modèle dégagé. A partir d'un échantillon d'apprentissage, un traitement symbolique commence pour la construction du GI en utilisant l'algorithme SIPINA. Le module CASI est décomposé en trois sous modules :

- COG (Cellular Optimization and Generation) consiste à assister le traitement effectué par l'algorithme SIPINA pour générer le graphe d'induction booléen (GIB).
- CIE (Cellular Inference Engine) est capable, à partir du GIB, de générer une base de connaissance booléenne (BCB). Ce module simule le fonctionnement du cycle de base d'un moteur d'inférence en utilisant deux couches finies d'automates finis ; *CELFAIT*, pour la base des faits et *CELREGLE*, pour la base de règles. Les états des cellules se composent de trois parties : *EF, IF et SF*, respectivement *ER*, *IR* et *SR*, sont l'entrée, l'état interne et la sortie d'une cellule de *CELFAIT*, respectivement d'une cellule de *CELREGLE*.
- CV (Cellular Validation) est consacré au processus de validation du modèle élaboré.

### 3.3.2 Exemple d'illustration de la modélisation booléenne

Le tableau 2 représente un échantillon d'apprentissage de 14 exemples de détection des maladies 'Asthme' et 'Diabète' dans différentes localités (villes). Chaque exemple ou individu est décrit par quatre attributs : Localité, Saison, Age et Sexe.

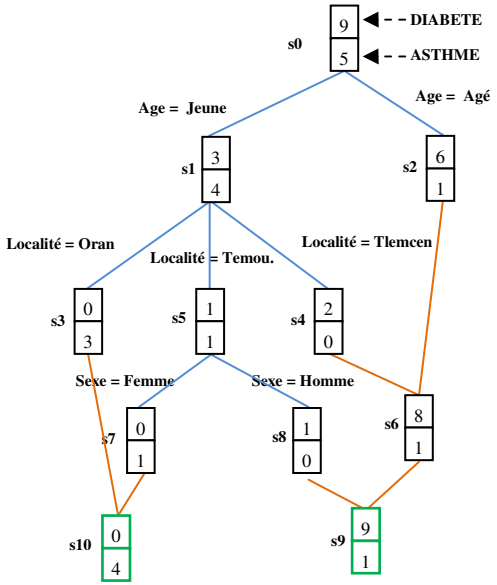
*Tableau 2. Exemple d'un échantillon d'apprentissage – Les maladies*

| | *Localité* | *Saison* | *Age* | *Sexe* | *Maladie* |
|---|---|---|---|---|---|
| $\omega_1$ | Oran | Hiver | Jeune | Homme | Asthme |
| $\omega_2$ | Oran | Hiver | Jeune | Femme | Asthme |
| $\omega_3$ | Tlemcen | Hiver | Jeune | Homme | Diabète |
| $\omega_4$ | Témouchent | Printemps | Jeune | Homme | Diabète |

| | | | | | |
|---|---|---|---|---|---|
| $\omega_5$ | Témouchent | Eté | Agé | Homme | Diabète |
| $\omega_6$ | Témouchent | Eté | Agé | Femme | Asthme |
| $\omega_7$ | Tlemcen | Eté | Agé | Femme | Diabète |
| $\omega_8$ | Oran | Printemps | Jeune | Homme | Asthme |
| $\omega_9$ | Oran | Eté | Agé | Homme | Diabète |
| $\omega_{10}$ | Témouchent | Printemps | Agé | Homme | Diabète |
| $\omega_{11}$ | Oran | Printemps | Agé | Femme | Diabète |
| $\omega_{12}$ | Tlemcen | Printemps | Jeune | Femme | Diabète |
| $\omega_{13}$ | Tlemcen | Hiver | Agé | Homme | Diabète |
| $\omega_{14}$ | Témouchent | Printemps | Jeune | Femme | Asthme |

Pour illustrer l'architecture et le principe de fonctionnement du module *CIE*, nous considérons le graphe d'induction (Fig. 5.) obtenu par IGSS.

CASI nous permet d'obtenir les partitions S0 (sommet (*s0*)), S1 (Age = Jeune (*s1*), Age = Agé (*s2*)), S2 (Localité = Oran (*s3*) , Localité = Tlemcen (*s4*), Localité = Témouchent (*s5*)), S3 (Localité = Tlemcen et Age = Agé (*s6*)), S4 (Sexe = Femme (*s7*), Sexe = Homme (*s8*)) et S5 (Localité = Oran et Sexe = Femme (*s9*), (*s6*) et Sexe = Homme (*s10*)).




| Faits | EF | IF | SF |
|---|---|---|---|
| s0 | 1 | 0 | 0 |
| Age=Jeune | 0 | 1 | 0 |
| s1 | 0 | 0 | 0 |
| Age=Agé | 0 | 1 | 0 |
| s2 | 0 | 0 | 0 |
| Localité=Oran | 0 | 1 | 0 |
| s3 | 0 | 0 | 0 |
| Localité=Tlemcen | 0 | 1 | 0 |
| s4 | 0 | 0 | 0 |
| Localité=Témouchent | 0 | 1 | 0 |
| s5 | 0 | 0 | 0 |
| s6 | 0 | 0 | 0 |
| Sexe=Femme | 0 | 1 | 0 |
| s7 | 0 | 0 | 0 |
| Sexe =Homme | 0 | 1 | 0 |
| s8 | 0 | 0 | 0 |
| s9 | 0 | 0 | 0 |
| s10 | 0 | 0 | 0 |

CELFAIT

| Règles | ER | IR | SR |
|---|---|---|---|
| R 1 | 0 | 1 | 1 |
| R 2 | 0 | 1 | 1 |
| R 3 | 0 | 1 | 1 |
| R 4 | 0 | 1 | 1 |
| R 5 | 0 | 1 | 1 |
| R 6 | 0 | 1 | 1 |
| R 7 | 0 | 1 | 1 |
| R 8 | 0 | 1 | 1 |
| R 9 | 0 | 1 | 1 |
| R 10 | 0 | 1 | 1 |

CELREGLE

*Figure 5 : Graphe d'induction de l'échantillon du tableau 1 réalisé par IGSS et initialisation de CELFAIT et CELREGLE*

La figure 5 montre comment la base de connaissances extraite à partir de ce graphe est représentée par les couches *CELFAIT* et *CELREGLE*. Initialement, toutes les entrées des cellules dans la couche *CELFAIT* sont passives (*EF = 0*), exceptées celles qui représentent la base des faits initiale (*EF(1) = 1*). Notons que dans cette étape les deux matrices d'incidence d'entrée ($R_E$) et de sortie ($R_S$) de CASI sont générées.

La dynamique de l'automate cellulaire CIE, pour simuler le fonctionnement d'un *Moteur* d'*Inférence*, utilise deux fonctions de transitions $\delta_{fait}$ pour évaluation, sélection et filtrage et $\delta_{règle}$ pour exécution.

- La fonction de transition δ fait :
  $(EF, IF, SF, ER, IR, SR) \rightarrow \delta fact\ (EF, IF, EF, ER + (R_E^T \cdot EF), IR, SR)$

- La fonction de transition δ règle :
  $(EF, IF, SF, ER, IR, SR) \rightarrow \delta rule(EF + (RS \cdot ER), IF, SF, ER, IR, \overline{ER})$

Où la matrice $R_E^T$ désigne la transposé de $R_E$.

$R_E$ et $R_S$ sont respectivement les matrices d'entrée et de sortie (Figure 6):

- La relation d'entrée, notée iREj, est formulée comme suit:
  $\forall i = 1..l, \forall j = 1..r, if\ (fact\ i \in Premise\ of\ rule\ j) then\ R_E(i,j) \leftarrow 1$

- La relation de sortie, notée iRSj, est formulée comme suit:
  $\forall i = 1..l, \forall j = 1..r, if\ (fact\ i \in Conclusion\ of\ rule\ j) then\ R_S(i,j) \leftarrow 1$

Les matrices d'incidence $R_E$ and $R_S$ représentent la relation *entrée/sortie* des Faits et sont utilisées en chaînage avant. On peut également utiliser $R_E$ comme relation de sortie et $R_S$ comme relation de sortie pour lancer une inférence en chaînage arrière. Nous notons qu'une cellule du voisinage d'une cellule qui appartient à *CELFAIT* (respectivement à *CELREGLE*) n'appartient pas à la couche *CELFAIT* (respectivement *CELREGLE*).

Pour produire, enfin, des règles conjonctives (Tableau.3), le module COG coopère avec le moteur d'inférence cellulaire (CIE) qui utilise les mêmes fonctions de transition $\delta_{fait}$ et $\delta_{règle}$ avec la permutation de $R_E$ et de $R_S$ du graphe, en partant du nœud terminal vers la racine s0.

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| s0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Age=Jeune | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Age=Agé | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Localité=Oran | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Localité=Tlemcen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Localité=Témouchent | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| s6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Sexe=Femme | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Sexe =Homme | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| s9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**$R_E$ (Relation d'Entrée)**

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| s0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Age=Jeune | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Age=Agé | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Localité=Oran | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Localité=Tlemcen | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| s4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Localité=Témouchent | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| s5 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| s6 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Sexe=Femme | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| s7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Sexe =Homme | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| s8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| s9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| s10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**$R_S$ (Relation de Sortie)**

*Figure 6 : Configuration initiale des deux matrices RE / RS*

| Faits | EF | IF | SF |
|---|---|---|---|
| Age=Jeune | 0 | 1 | 0 |
| Age=Agé | 0 | 1 | 0 |
| Localité=Oran | 0 | 1 | 0 |
| Localité=Tlemcen | 0 | 1 | 0 |
| Localité=Témouchent | 0 | 1 | 0 |
| Sexe=Femme | 0 | 1 | 0 |
| Sexe =Homme | 0 | 1 | 0 |
| s9 (Classe = Diabète) | 0 | 1 | 0 |
| s10 (Classe = Asthme) | 0 | 1 | 0 |

**CELFAIT**

| Règles | ER | IR | SR |
|---|---|---|---|
| R 1 | 0 | 1 | 1 |
| R 2 | 0 | 1 | 1 |
| R 3 | 0 | 1 | 1 |
| R 4 | 0 | 1 | 1 |
| R 5 | 0 | 1 | 1 |

**CELREGLE**

| | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Age=Jeune | 1 | 1 | 1 | 1 | 0 |
| Age=Agé | 0 | 0 | 0 | 0 | 1 |
| Localité=Oran | 1 | 0 | 0 | 0 | 0 |
| Localité=Tlemcen | 0 | 1 | 0 | 0 | 0 |
| Localité=Témouchent | 0 | 0 | 1 | 1 | 0 |
| Sexe=Femme | 0 | 0 | 1 | 0 | 0 |
| Sexe =Homme | 0 | 0 | 0 | 1 | 0 |
| s9 (Classe = Diabète) | 0 | 0 | 0 | 0 | 0 |
| s10 (Classe = Asthme) | 0 | 0 | 0 | 0 | 0 |

**$R_E$ (Relation d'Entrée)**

| | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Age=Jeune | 0 | 0 | 0 | 0 | 0 |
| Age=Agé | 0 | 0 | 0 | 0 | 0 |
| Localité=Oran | 0 | 0 | 0 | 0 | 0 |
| Localité=Tlemcen | 0 | 0 | 0 | 0 | 0 |
| Localité=Témouchent | 0 | 0 | 0 | 0 | 0 |
| Sexe=Femme | 0 | 0 | 0 | 0 | 0 |
| Sexe =Homme | 0 | 0 | 0 | 0 | 0 |
| S9 (Classe = Diabète) | 0 | 1 | 0 | 1 | 1 |
| s10 (Classe = Asthme) | 1 | 0 | 1 | 0 | 0 |

**$R_S$ (Relation de Sortie)**

*Figure 7 : Base de connaissances booléenne issue du graphe d'induction de la figure 5.*

La représentation de la base de connaissance booléenne par CASI est illustrée par *CELFAIT*, *CELREGLE*, $R_E$ et $R_S$ (Fig.7).

*Tableau 3. Les règles conjonctives produites*

| | |
|---|---|
| *1* | Si *(Localité = Oran* ET *Age = Jeune)* Alors *Asthme* |
| 2 | *Si* (Localité = Tlemcen *ET* Age = Jeune) *Alors* Diabète |
| 3 | *Si* (Localité = Témouchent *ET* Age = Jeune *ET* Sexe = Femme) *Alors* Asthme |
| 4 | *Si* (Localité = Témouchent *ET* Age = Jeune *ET* Sexe = Homme) *Alors* Diabète |
| 5 | *Si* (Age = Agé) *Alors* Diabète |

Maintenant, le module CV est prêt à lancer la phase de validation, en employant le même principe booléen de base du moteur d'inférence cellulaire CIE, et les mêmes fonctions de transition $\delta_{fait}$ et $\delta_{règle}$.
Supposons un échantillon de test (Tableau 4.) est composé de 5 cas de détection de maladies appartenant aux classes 'Asthme' et 'Diabète', où la classe 'Asthme' est la classe majoritaire de *s10*, et la classe 'Diabète' est la classe majoritaire de *s6* et de *s9*.

*Tableau 4. Exemple d'un échantillon test.*

| | *Localité* | *Age* | *Sexe* | *Maladie* |
|---|---|---|---|---|
| $\omega_1$ | Oran | Jeune | Homme | Asthme |
| $\omega_2$ | Témouchent | Agé | Femme | Asthme |
| $\omega_3$ | Oran | Agé | Homme | Diabète |
| $\omega_4$ | Témouchent | Agé | Homme | Diabète |
| $\omega_5$ | Tlemcen | Agé | Homme | Diabète |

La figure 8 résume la validation de l'individu ω1.

| $\omega_5$ | $\omega_4$ | $\omega_3$ | $\omega_2$ | $\omega_1$ | →→ |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 | Age=Jeune |
| 1 | 1 | 1 | 1 | 0 | Age=Agé |
| 0 | 0 | 1 | 0 | 1 | Localité=Oran |
| 1 | 0 | 0 | 0 | 0 | Localité=Tlemcen |
| 0 | 1 | 0 | 1 | 0 | Localité=Témouchent |
| 0 | 0 | 0 | 1 | 0 | Sexe=Femme |
| 1 | 1 | 1 | 0 | 1 | Sexe =Homme |
| 1 | 1 | 1 | 0 | 0 | Classe = Diabète |
| 0 | 0 | 0 | 1 | 1 | Classe = Asthme |

| EF | IF | SF |
|---|---|---|
| 0 | 1 | 0→1 |
| 0 | 1 | 0 |
| 0 | 1 | 0→1 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0→1 | 1 | 0 |

CELFAIT

| ER | IR | SR |
|---|---|---|
| 0 | 1 | 1→0 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| | | |
| | | |
| | | |
| | | |

CELREGLE

*Figure 8 : Validation de ω1*

# 4 Expérimentations

Avant de donner les résultats de la phase d'expérimentation, nous tenons à rappeler que l'entreposage des données a représenté une tâche fastidieuse dans la mise en œuvre du projet, notamment dans la collecte des données. Nous avons, tout de même, pu mettre dans SARESM plus de trente millions d'enregistrements de ventes, échelonnés entre janvier 2003 et Avril 2010, et relatifs à 219 pharmacies réparties sur 10 départements. Il est à noter que ces enregistrements représentent des données brutes de ventes sur lesquelles aucune forme d'agrégation n'a été effectuée, pour obtenir enfin, après le prétraitement et la tâche d'identification des caractéristiques des patients (Sexe et Age), près de 500 000 actes de ventes[6] pour les maladies choisies (Asthme, HTA et Diabète). Notre expérimentation a porté sur un échantillon de 78 122 actes de ventes.

Afin de réaliser notre expérimentation, nous avons importé les données dans IGSS, nous avons sélectionné les attributs et la variable à prédire (cf. Tableau 5) et nous avons lancé l'induction en utilisant le principe de SIPINA sur les données d'apprentissage.

*Tableau 5. Représentation des attributs et la classe.*

| ***Attribut*** | ***Signification*** | ***Valeurs possibles*** |
|---|---|---|
| DENSITE_DEMOG | Densité démographique | ***Grande*** *s'il s'agit de la concentration dans une grande ville,* ***Banlieue*** *sinon.* |
| CLIMAT | *Le climat par rapport principalement à l'humidité* | ***Sec***, ***Humide*** |
| WILAYA | *Le numéro du département* | ***13****(Tlemcen),****22*** *(Sidi Belabbes) ,****29*** *(Mascara),****31****(Oran),****46*** *(Témouchent) et* ***48*** *(Relizane)* |
| SEXE | *Le sexe du patient* | ***M*** *: Masculin,* ***F*** *: Feminin* |
| AGE | *L'âge du patient* | *Donné en tranches :* ***AD1*** *(<=40 ans),* ***AD2*** *(entre 41 et 65 ans) et* ***Agé*** *(>65 ans).* |
| SAISON | *La période choisie est la saison* | ***Eté***, ***Hiver***, ***Printemps*** *et* ***Automne***. |
| CLASSE_THERAP | *C'est la classe de la maladie à prédire* | ***AST*** *(Asthme),* ***DBT*** *(Diabète) et* ***HTA*** *(Hypertension artérielle).* |

Nous avons, enfin, obtenu le modèle de fouille de données suivant (Tableau 6) :

[6] Nous appelons actes de ventes les ordonnances qui ne concernent que les maladies étudiées.

*Tableau 6. Exemples de règles conjonctives produites dans l'expérimentation*

| | |
|---|---|
| *1* | Si (AGE = Agé) Alors CLASSE_THERAP = HTA |
| 2 | *Si* (WILAYA = 31 *ET* AGE = AD2) *Alors* CLASSE_THERAP = HTA |
| 3 | *Si* (WILAYA = 22 *ET* AGE = AD2) *Alors* CLASSE_THERAP = HTA |
| 4 | *Si* (WILAYA = 29 *ET* AGE = AD2) *Alors* CLASSE_THERAP = HTA |
| 5 | *Si* (WILAYA = 46 *ET* AGE = AD2) *Alors* CLASSE_THERAP = HTA |
| 6 | *Si* (WILAYA = 13 *ET* SEXE = F *ET* AGE = AD2) *Alors* CLASSE_THERAP = HTA |
| 7 | *Si* (AGE = AD1) *Alors* CLASSE_THERAP = AST |
| 8 | *Si* (WILAYA = 48 *ET* AGE = AD2) *Alors* CLASSE_THERAP = HTA |
| 9 | *Si* (WILAYA = 13 *ET* SEXE = M *ET* AGE = AD2) *Alors* CLASSE_THERAP = DBT |

# 5 Conclusion

Dans cet article, nous avons abordé la fouille de données, dans le domaine de la santé publique et en particulier l'épidémiologie, en fournissant des modèles de prédiction des maladies chroniques par la modélisation booléenne des règles de prédiction épidémiologique. Dans le contexte de l'analyse des maladies chroniques, le graphe d'induction cellulaire engendré est un modèle booléen qui nous permettra de voir de plus près les relations entre la maladie et les personnes exposées à cette dernière par rapport aux caractéristiques physionomiques et à l'environnement. Le graphe d'induction engendré facilitera l'identification des maladies au niveau d'une région donnée en vue de proposer de meilleures mesures de prise en charge des patients.

# Références


[1] ATMANI B. et BELDJILALI B., *Knowledge Discovery in Database : Induction Graph and Cellular Automaton*, Computing and Informatics Journal, Vol.26, N°2 pp 171--197, 2007

[2] BENAMINA M. et ATMANI B., *WCSS: un système cellulaire d'extraction et de gestion des connaissances***,** Troisième atelier sur les systèmes d**é**cisionnels, 10 et 11 octobre 2008, Mohammadia – Maroc, pp 223—234, 2008

[3] BALA P.K., *Advances in Electrical Engineering and Computational Science*. pp 587--598. Xavier Institute of Management, Bhubaneswar, India, 2009

[4] BEREZNICKI B.J., PETERSON G.M., JACKSON S.L., WALTERS H., FITZMAURICE K. et GEE P., *Pharmacist-initiated general practitioner referral of patients with suboptimal asthma management Pharm*. World, 2008

[5] CHEN J., HE H., LI J., JIN H., McAULLAY D., WILLIAMS G., SPARKS R. et KELMAN C., *Representing Association Classification Rules Mined from Health Data*, International Conference on Knowledge-Based Intelligent Information and Engineering Systems N°9, Melbourne, Australia, 2005

[6] FUGH-BERMAN A., *Prescription Tracking and Public Health. Department of Physiology and Biophysics*, Georgetown University Medical Center, Washington, DC, USA. 2008

[7] HAMURO Y., KATOH N., MATSUDA Y. et YADA K., *Mining Pharmacy Data Helps to Make Profits*. Data Mining and Knowledge Discovery 2, 391–398, 1998

[8] HARBI N., BOUSSAID O. et BENTAYEB F., *Propriétés d'un modèle conceptuel multidimensionnel pour les données complexes*. 8èmes Journées Francophones Extraction et Gestion des Connaissances, Sophia Antipolis, 2008

[9] INMON W.H., *Building the Data Warehouse*, 2nd Edition John Wiley & Sons, Inc., ISBN n°0471-14161-5, USA, 1996

[10] KIMBALL R., *The data warehouse toolkit : practical techniques for building dimensional data warehouses*. John Wiley & Sons, Inc., New York, NY, USA, 1996

[11] RAVAT F., TESTE O. et ZURFLUH G., *Modélisation et extraction de données pour un entrepôt objet*. Université Paul Sabatier (Toulouse III), IRIT, équipe SIG, 2000

[12] RAVAT F., TESTE O. et ZURFLUH G., *Modélisation multidimensionnelle des systèmes décisionnels*. Revue des Sciences et Technologies de l'information, Vol n°1-2/200, pp. 201-212, EGC 2001 17-19 2001, Nantes, France, 2001

[13] SELMOUNE N., BOUKHEDOUMA S. et ALIMAZIGHI Z., *Conception d'un outil décisionnel pour la gestion de la relation client dans un site de e-commerce*. SETIT 3rd International Conference, Tunisia, 2005

[14] SOIBELMAN L., ASCE M. et HYUNJOO K., *Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases*. Journal Of Computing In Civil Engineering, 2002

[15] SUMATHI S. et SIVANANDAM S.N., *Introduction to Data Mining and its Applications*. pp 500--543, 2006

[16] MELLEY K. et PETERSEN K., *Fact Sheet : Prescription Data Mining*. Pew Prescription Project : http://www.prescriptionproject.org, 2008

[17] ZIGHED D.A., *Méthodes et outils pour les processus d'interrogation non arborescents*. Thèse de Doctorat, Université Lyon 1, 1985

[18] ZIGHED D.A., Auray J.P. et Duru, G., *SIPINA: Méthode et Logiciel, Lacassagne*, 1992