Modèles et Heuristiques pour le traitement spatio-temporel de l'information environnementale

Mireille Batton-Hubert

Habilitation à Diriger des Recherches

Université Jean Monnet

Ecole Doctorale de l'Ecole Nationale Supérieure des Mines de Saint-Étienne et de l'Université Jean Monnet

Spécialité : Sciences et Génie de l'Environnement

Modèles et Heuristiques pour le traitement spatio-temporel de l'information environnementale

Présentée par

Mireille BATTON - HUBERT

Soutenue le 04 décembre 2006 devant le jury composé de :

José RAGOT	Professeur, Institut National Polytechnique de Lorraine, Nancy	Rapporteur
David W. PEARSON	Professeur, Université Jean Monnet, St Etienne	Rapporteur
Michel MAINGUENAUD	Professeur, Institut National des Sciences Appliquées, Rouen	Rapporteur Président de jury
Jean-Jacques CHEVALLIER	Professeur, Université Laval, Québec	Examinateur
Pierre DUMOLARD	Professeur, Université J. Fourier, Grenoble 1	Examinateur
Roland DECHOMETS	Professeur, Ecole Nationale Supérieure des Mines de St Etienne	Examinateur
Didier GRAILLOT	Directeur de recherche, Ecole Nationale Supérieure des Mines de St Etienne	Directeur HDR

Préambule

Ce mémoire a pour objectif de faire le bilan qualitatif et quantitatif de mes activités de recherche depuis ma nomination à l'Ecole nationale supérieure des mines de Saint-Étienne en 1996.

La première partie sera consacrée à la pertinence scientifique du concept de la connaissance d'un phénomène continu simultanément dans l'espace et dans le temps dans un Environnement physique et anthropique. Cette problématique de recherche est centrée sur un domaine d'application, l'optimisation et la simulation d'impacts environnementaux. Volontairement elle ne se scinde pas entre le réalisé et le futur. Car les travaux de recherche se déroulent dans un continuum temporel du point de vue des problématiques d'application ou au niveau des avancées historiques. Ce qui impliquerait que les travaux passés sont exhaustifs, et sans amélioration possible. Ce travail est une réflexion approfondie d'une préoccupation commune se déclinant au travers de 2 grands axes de recherche (volet 1 et volet 2) qui débouchent sur une proposition de thématiques de recherche pour les années à venir (volet 3).

Les 3 volets sont :

- volet 1 : Structuration et Modélisation de l'information spatialisée environnementale - aide à la décision en hydrodynamique et Ressources en eaux
- volet 2 : Analyse *espace - temps* de l'information environnementale
- volet 3 : Justification d'heuristique de la composante espace - temps de l'information environnementale

Un rapport annexe contient le syllabus:

- des activités d'enseignement et de responsabilités diverses assurées,
- des thématiques de recherche abordées avec leurs principaux résultats issues :
    - des activités d'encadrement doctoral
    - de la réalisation de projets de recherche (académique et appliquée)
    - des publications réalisées à ce jour.

Ce document est établi en vue de l'obtention de l'Habilitation à Diriger des Recherches.

Table des matières

Avant propos

L'objet même, que constitue un mémoire de HDR est en quelque sorte un fil d'Ariane qu'il faut tisser puis, dérouler dans le labyrinthe de ses recherches. Il constitue un enjeu pour sa propre conception du futur en terme de recherches et constitue également une rétrospective, même si elle est à court terme, sur son passé de chercheur. Il est à la fois le présent et le futur : il doit prouver que les propositions passées sont fructueuses en terme d'applications même modestes en Recherche appliquée ou pour l'aide à la décision (cette dernière pourra être judicieuse si en adéquation avec des préoccupations de développement durable, par exemple) mais également, que les perspectives et les propositions futures le seront également. Paradoxe, dilemme ou sophisme du *Ménon* de Platon, le chercheur doit-il connaître ou ne pas connaître ce qu'il veut trouver, *"peut-on trouver ce que l'on ne cherchait pas ?"*

S'attaquer à la rédaction d'un tel document n'est donc pas une chose aisée. Classiquement un chercheur peut se rattacher à un domaine *mono* disciplinaire or des choix et des opportunités successives en ont décidé autrement : après une thèse de doctorant en Informatique Appliquée aux Géosciences (Université P et M. Curie - Institut de Physique du Globe) j'ai pu intégrer le domaine des Systèmes Informatiques de l'Environnement Urbain et tout particulièrement celui de l'hydrologie urbaine (INSA). Bien que l'on ne parlât pas encore de Sciences de l'Environnement, les Géosciences, l'Informatique et l'Environnement constituaient des compétences en adéquation avec ma nomination à l'Ecole Nationale Supérieure des Mines de Saint-Étienne en 1996 au Centre *Sciences, Information et Technologies pour l'Environnement.*

Pour toutes ces raisons j'ai choisi d'établir le fil conducteur d'une réflexion transversale et pluridisciplinaire de la *Connaissance* d'un processus ou phénomène continu (la circulation atmosphérique de polluant, l'écoulement hydrodynamique ou hydraulique, l'érosion) perçu simultanément dans l'espace et dans le temps.

Mais regardons de plus près ce qui nous préoccupe : soit un phénomène physique évoluant dans un environnement naturel fortement *anthropisé* pouvant être à l'origine d'effets et/ou impacts éventuels:

- un écoulement souterrain perturbé par des crues/étiages ou bien par des prélèvements intensifs en période d'étiage dénoyant un puits,
- une pollution atmosphérique urbaine aggravée lors de beau temps et de fortes circulations automobiles ayant des impacts sanitaires,
- une activité industrielle produisant des nuisances olfactives autour d'un site …

L'aide à la décision consiste à:

- prévoir la côte de dénoyage du puits,
- calculer un risque sur la santé des populations sensibles et le dépassement d'un seuil d'alerte,
- diminuer l'occurrence des nuisances olfactives autour d'un site industriel.

On agit pour et sur un environnement fortement "*anthropisé*" et socio-économique où le processus décisionnel n'est ni l'optimum (que l'on ignore), ni un compromis, ni une solution satisfaisante d'une optimisation multiobjectif (Collette et Siarry 2002).
Pour cette raison fondamentale, il s'agit de fournir la meilleure information disponible et quantifiable pour l'aide à la décision, qui peut être une valeur scalaire (issue d'un modèle) ou une carte spatialisée (construite par un SIG) ou toute autre donnée mais qui doit être transcrite

et adaptée à l'acteur concerné soit l'industriel, l'élu, ou l'aménageur (Roche et Batton-Hubert, 1998). Pour fournir la meilleure donnée où information, il faut se doter d'outils numériques permettant de calculer, simuler ou extraire cette donnée. Classiquement on se référera à des outils numériques de mécanique des fluides, statistiques ou probabilistes: on reviendra largement sur ce point ultérieurement.

Usuellement, qu'est-il préconisé ? On dispose d'un modèle mathématique déterministe (simulation des écoulements) ou non (probabiliste, statistique…) qui consomme des données et produit en sortie une valeur de la variable recherchée sans oublier la boucle de rétroaction sur l'ajustement de l'erreur et la prise en compte de la sensibilité et la robustesse du modèle . Ceci suppute :

- qu'il existe un modèle mathématique et/ou physique
- que les données nécessaires au modèle sont disponibles.

Un modèle mathématique/physique quel qu'il soit, est une représentation d'une certaine réalité. Il est par conséquent construit avec des hypothèses, selon des théories et pour une connaissance perçue.

N'y aurait il pas un autre moyen de rechercher une autre perception du phénomène physique concerné ? Plaçons-nous à un niveau d'abstraction supérieur celui de la *Connaissance*. Mais avant tout qu'entend - t - on par *Connaissance* ?

Etat de l'esprit de ce qu'il connaît et discerne dit (*le Littré*).

C'est *être capable de former une idée, un concept, une image de quelque chose* ce qui s'oppose à la notion d'information qui est un élément quantitatif ou qualitatif, utilisée pour bâtir une connaissance de quelque chose. La donnée est une quantité connue, mesurable, *instanciation*[1] de paramètres et caractéristiques (*var.* ou *paramètres* d'un point de vue mathématique) de l'objet processus concerné.

Mes travaux de recherche s'articulent autour de cette notion d'identification de connaissance complémentaire d'un processus physique, dynamique par laquelle on cherche à extraire l'information pertinente non explicite a priori qui permettra d'asseoir mais surtout de discerner d'autres particularités sur la perception globale du processus considéré. On admet alors que toute forme d'information disponible pourra être exploitable. On admet aussi les limites explicites des modèles utilisés (comme la non linéarité, la non existence d'une solution analytique, l'optimum par heuristique) et de ne pas forcement s'attacher à construire à tout prix un nouveau modèle mais de l'améliorer en exploitant ses lacunes (retranscription sur l'erreur).

Les bases de notre argumentaire posées, voyons précisément ce dont on dispose : en priorité, ce sont des données descriptives, mesurées et /ou géométriques ainsi sont elles de nature différente.

Partant du constat de l'importance en quantité et en variabilité de l'information spatialisée concernant les Géosciences mais plus généralement de l'information disponible en Sciences de l'Environnement, la première étape a consisté à proposer une schématisation conceptuelle avancée de l'information en Géosciences basée sur une approche géométrique et topologique de l'espace topographique (Hubert 1993). Toute cette réflexion sur les schémas conceptuels de données et leur utilisation pour l'interopérabilité (au delà du couplage) entre plusieurs modèles déterministes et une base ou des bases de données géographiques permet d'établir les

[1] *Instanciation* : ce qui fait référence à la notion d'instance usuelle en informatique

fondements de notre approche dite *espace - temps* de l'information et de la modélisation de processus environnentaux. Le volet 1 *Structuration et Modélisation de l'information spatialisée environnementale* fournit les bases de cette approche et un cas d'application en *Aide à la décision en hydrodynamique et Ressources en eaux.* Une proposition au problème non totalement résolu actuellement, de l'interopérabilité structurelle entre une/ou des BDD et, un/ou des codes numériques de simulation de processus hydrodynamique et hydraulique est développée au paragraphe 2. Elle constitue les premiers résultats du transfert de ces techniques appliquées à l'étude d'impacts et de la ressource en eaux. D'autres modèles sont envisagés afin d'enrichir une base d'outils (modèles physiques + code numérique associé) nécessaires aux interfaces existantes entre plusieurs types d'écoulement. L'argumentation de choix d'un modèle en adéquation avec la connaissance du phénomène, est entreprise à travers l'élaboration de critères quantifiés par des variables ou indicateurs géométriques, physiques ou mathématiques.

Que ce soit pour un domaine plus académique comme l'hydrodynamique ou pour des Sciences de l'Ingénieur, il est fondamental de savoir comment organiser, structurer l'information en tant qu'objet : on part d'une donnée et on veut reconstruire l'objet avec ses propriétés, ses relations d'un point de vue sémantique. Dans cette approche, on suppose qu'il y a un sens univoque entre la donnée et l'objet et, que la donnée doit être une image un "peu *idéale"* de l'objet. Mais, quelle est l'image *idéale* de l'objet (objet qui sous-entend le processus physique qui nous préoccupe) ? Peut-on approcher cette image ?
A priori, on peut supposer que oui s'il existe un modèle déterministe ou stochastique du processus. Regardons ce qui se pratique dans les techniques de diagnostic les plus récentes : elles sont basées sur le concept de redondance de l'information qui utilise un test de cohérence entre un comportement observé du processus et le comportement attendu par une représentation mathématique du processus. Le diagnostic de défaillance comporte classiquement 3 étapes fonctionnelles : i)génération des résidus, ii) génération de la signature des défaillances, et la détection et la localisation des défaillances (soit le diagnostic), iii) puis entreprise d'une action. Or cette démarche impose qu'il y ait suffisamment de données pour construire un état initial et un modèle du processus pour exploiter l'erreur entre l'information attendue et l'observée.

On intègre le fait que la donnée est entachée d'erreur, mais que l'on ne recherche pas forcement à connaître une image idéale ou attendue de l'objet : la connaissance de l'objet est recherchée. On avance comme lemme qu'il existe dans la donnée et/ou information collectée, la propriété fondamentale de contenir une information pertinente qui permet de faire émerger d'autres formes d'information qui conduisent à accroître la connaissance intrinsèque du processus. L'objet de cette recherche est la caractérisation d'un processus physique continu dont on veut prévoir un certain état ou du moins des évènements qui lui sont liés : le cadre méthodologique est établi au volet 2.
Deux types d'approches ont été proposées en fonction de la disponibilité des données sur le phénomène ou le dispositif, de l'état de fonctionnement du système et de la faisabilité d'une simulation déterministe classique.
La première approche a pour cadre celui des réseaux d'adduction en eau potable en réseau urbain (chapitre 1.); la seconde s'intéresse au problème de la qualité de l'air, la pollution atmosphérique et les nuisances olfactives (chapitre 2). Bien que les domaines d'application soient différents, la préoccupation commune est la perception du processus, appelée encore l'identification. Lorsque le système est mal connu, un réseau d'adduction en eau potable usé avec peu de mesures directes, l'originalité de la démarche consiste à exploiter une forme d'observation de l'erreur du système, que sont les interventions. Elle contribue à obtenir un

ordonnancement des aléas sur le réseau et renseigne le gestionnaire sur le fonctionnement de son réseau en mode de diagnostic ou de prévision.
Lorsque le système permet d'envisager un modèle physique déterministe relativement réaliste, l'exploitation des résultats de simulation identifie un ensemble possible d'observations qui peuvent prétendre être des mesures de l'état du système (dans un certain espace géométrique). On doit intégrer cette nouvelle dimension dans la prévision d'un scénario de qualité d'air au lendemain.

Quelle que soit la nature de la donnée utilisée, d'ordre qualitative ou quantitative, il est possible de reconstituer une certaine connaissance du processus. Les mots clés sont *interpolation, estimation, simulation, prévision*. Le volet 3 a pour objet de proposer des éléments méthodologiques pour tenter de résoudre la prise en compte de la dimension *temp*s et *espace* de la donnée environnementale. Il s'appuie sur les résultats et l'analyse du volet 2. et aboutit à une heuristique de ces 2 dimensions particulières de la donnée. Il décrit les aboutissements partiels des précédents volets et les perspectives de recherche, il constitue la conclusion générale de ce mémoire.

Pour montrer l'étendue du domaine des Sciences pour l'Environnement et qu'il constitue un point de convergence de toute une famille de modèles mathématiques appliqués qui sont liés aux problèmes de l'optimisation en continu, en nombre discret, linéaire ou non, le dernier paragraphe cite quelques exemples de projets de recherche opérationnelle en Environnement sur lesquels j'ai pu travailler comme la gestion des bassins de production de déchets ménagers et les sites des unités de traitement.

Convaincue de l'intérêt, de ces approches à base de représentation, où les mesures portent en elles un modèle du système, de la transdisciplinarité pour une ingénierie de l'Environnement, ce mémoire eût eu pour intention d'apporter, même si partielle, une contribution au développement d'un thème novateur qu'est le *data mining et la Reconnaissance de Forme* en Sciences de l'Environnement.

"Le Temps des Hommes est de l'éternité pliée"
Jean Cocteau

Volet 1 : Structuration et Modélisation de l'information spatialisée environnementale - aide à la décision en hydrodynamique et Ressources en eaux

Il s'agit d'établir l'importance d'une approche conceptuelle de la donnée géo-graphique et de l'information associée et de démonter comment l'informatique permet de se doter d'outils adaptés à la simulation d'impacts en Géosciences. Les éléments précurseurs de cet axe de recherche ont été établis durant ma thèse de doctorat d'Université dans le domaine des Géosciences puis développés sur la première thématique *Hydrodynamique et Ressources en eaux* à l'Ecole nationale supérieure des mines de Saint-Étienne.
Trois aspects sont développés :

- l'établissement d'un modèle conceptuel de la Connaissance
- l'interopérabilité entre modèles déterministes pour l'aide à la décision
- le développement et l'adaptation de modèles de simulation déterministes en hydrodynamique

La méthode posée au paragraphe 1. est appliquée au domaine de la Ressource en eaux avec notamment le problème de l'interopérabilité entre modèles. Pour l'Hydrodynamique, il s'agit également d'intégrer de nouveaux phénomènes physiques (loi de parois) et de faire une brève incursion dans les modèles de type potentiel pour les écoulements souterrains afin d'identifier des critères de choix optimal d'un type de modèle.

1. Elaboration de modèles de connaissance en Géosciences - géophysique

1.1 Objectif et méthodes

L'information spatialisée ou géo-graphique est associée à une donnée cartographique dont la carte IGN au 1/25000ème est un excellent exemple associant des informations sur la topographie, la végétation, l'hydrographie, l'habitat, le réseau routier etc.… Elle représente un espace ou une parcelle du monde réel (relief) ou anthropique (population) auquel est associé un modèle géométrique de type discret, le mode raster, ou de géométrie euclidienne, le mode vectoriel.
Pour éviter les travers d'une approche cartographique trop souvent liée à un découpage linéaire en couvertures appelées *layers* ou *covers* dans les SGBD et les SIG, nous avons alors proposé un formalisme conceptuel avancé de l'information en Géosciences basé sur une approche géométrique et topologique de l'espace topographique, défini par un Modèle Numérique et Topologique de Terrain, le MNTT (Hubert 1993). Le concept de CLASS, utilisé est hérité directement des schémas conceptuels objets structurés (Guttag 1977) mais surtout du modèle HBDS (Hypergraph Based Data Strutures) (Bouillé 1977). Précurseur des langages de programmation objet, le concept de type abstrait de données (Dahl 1966), permet de définir 6 nouveaux type de données : la Classe, l'Objet de la classe, le Lien entre classes, le Lien entre objets, l'Attribut de classes et l'attribut d'objet. A chacun de ces 6 types sont associées des procédures de traitement qui leurs sont propres ainsi que des requêtes d'accès à leur contenu et notamment à leur état. Le schéma conceptuel final utilise la notion de graphes et hypergraphes développée par F. Bouillé (Bouillé 1977). Chaque entité géomatique physique ayant une géométrie, est décrite par un modèle vectoriel associé à des entités de types "A*rc, Sommet, Domaine*" (ex. graphe de la topographie, graphe de la géologie, graphe de l'hydrographie) pouvant porter des TAD[2] de données factuelles.

Conjointement il s'agissait de proposer un modèle conceptuel qui intègre l'évolution de cette information géomatique par des processus d'accroissement, de transformation, de modification et migration, en terme de limites et de continuité ; soit, l'extension géo-graphique du phénomène physique modélisé. Le contexte était celui des Géosciences, avec l'évolution d'un modèle numérique associé à un processus d'érosion d'un relief conjointement à un processus d'écoulement en rivière.
Un Modèle Dynamique de Terrain est alors défini comme l'association entre un modèle numérique et topologique de terrain et un ensemble de *processus* informatiques, proches de la notion de transitions d'état des graphes Pert et ou des réseaux de Pétri.

Un processus informatique est un "objet informatique", entité algorithmique, dynamique pouvant interagir avec d'autres processus et sur lui-même (notion d'interruption du processus par lui-même et il existe une différence fondamentale entre un état *suspendu* et un état *terminé*). Il est donc caractérisé par son état en fonction du temps. On associe alors un phénomène physique (érosion, transport de sédiment, hydraulique fluviale, tectonique…) à un ensemble de processus transformant les objets modélisant l'espace géo-graphique contenus dans et par le modèle numérique de terrain.
Un gestionnaire de simulation utilisant les éléments du Sequencing Set de OJ. Dahl (Dahl et Nygaard 1968) et les coroutines du langage Simula67 (Dahl 1966), prend en charge les changements d'état de chacun des processus informatiques. Le SQS (SeQuencing Set) est une

[2] TAD : type abstrait de données

liste circulaire de processus, ordonnée n'ayant qu'un processus actif en tête. Il permet un enchaînement dynamique reproduisant des phénomènes conjoints et simultanés de phénomènes réels en mode séquentiel sur un calculateur ordinaire. Ce système est donc portable, adaptable et manipule l'information intrinsèque, portée par les TAD contenant l'information descriptive et géométrique.

1.2 Résultats

L'enjeu des années 80 était le problème de stockage, la visualisation 2D et 3D, l'interrogation de base de données de type SIG et d'un SGBD et enfin les premiers pas de l'analyse spatiale intégrée à ces *softwares*. Une approche conceptuelle d'une quantité très importante de données cartographiques et textuelles était une nécessité. Elle permettait en particulier, de prendre en compte toute l'information disponible et surtout prévoir une modélisation 3D de terrain complémentaire des représentations par maillage régulier ou non (*Digital Terrain Modelling, Triangular Irregular Network*) . Notons toutefois qu'à ce jour, bien qu'il existe des schémas conceptuels orientés objets sophistiqués notamment pour le modèle de réseau hydrographique ou autre, modèle *Network* d'ArcGIS 3.2, il demeure encore difficile d'intégrer de véritables entités géo-graphiques 3D de géométrie vectorielle dans un SIG pour un utilisateur *lambda*.

Une des premières applications de ces travaux concerne le projet Thetys (1990), collaboration de recherche entre l'IFP, l'IFREMER, Elf Aquitaine , le BRGM et l'Université Paris 6. Ce projet avait pour objet de reconstruire l'histoire de la mer primitive, la Thétys, afin de retrouver ses descendants directs, les couches à hydrocarbures. Projet d'intérêt interdisciplinaire, la centralisation et la structuration de la Base de Connaissance des données de divers domaines de la Géophysique - allant de la prospection sismique et gravimétrique au stockage souterrain, et intégrant la Géologie, l'hydrographie - était au cœur du projet.

L'application de ces méthodes dans un second temps, au domaine de l'Environnement et tout particulièrement à l'hydrologie urbaine, constitue en quelque sorte un *test de validation.* Ce travail a été effectué au département Génie Civil et Urbanisme de l'INSA de Lyon. Il s'agissait d'élaborer un schéma conceptuel commun pour des données de gestion et d'entretien de réseau d'assainissement de la ville de Lyon prenant en compte des interfaces avec divers outils de conception hydraulique de réseau d'égouts, de dimensionnement d'ouvrage, et de bases de données géographiques de la Ville de Lyon. L'objet était de définir le schéma d'une nouvelle génération de logiciels de conception et d'élaboration d'un réseau d'égout en milieu urbain (CEDRE), évoluant vers un outil intégrant l'interopérabilité vers des bases de données urbaines de type quantitative et qualitative des eaux de réseau d'assainissement (logiciel CANOE).

1.3 A retenir

Partant de la nécessité d'avoir une idée de *l'ordre* ou plutôt de structure dans la donnée quantitative de l'information géographique des Géosciences, ces travaux ont fourni les bases d'une approche conceptuelle numérique et topologique d'un espace géographique. Puisque l'on voulait reproduire une simulation[3], on se dotait de la puissance de langage de

[3] au sens de déroulement

programmation objet pour déclarer, déclencher, un processus modifiant les objets contenus dans la base et suivre leur évolution.
L'intérêt a été prouvé dans l'application au domaine des Géosciences et en hydrologie urbaine ; mais l'amorce de cette démarche vers le domaine les Sciences de l'Environnement (conjointe à mon arrivée au Centre SITE de l'ENSM.SE) va nous faire prendre conscience d'un fait remarquable : l'idée même d'une structure conceptuelle sophistiquée " idéale" c'est à dire dotée des propriétés de la théorie des ensembles, indépendante de l'échelle et du mode de stockage, accessible, partageable, impliquait dans la démarche même de l'informaticien, de ne pas remettre en question l'information manipulée.
Bien sur il existait et existe toujours l'erreur associée à une donnée (écart type, intervalle de confiance, problème d'arrondi), le problème est plus en amont avec 3 hypothèses sous-jacentes :

- 1ère hypothèse : on disposerait d'un modèle connu programmable (ce qui ne veut pas dire exact ni valable pour le secteur donné) pour modéliser de façon approchée un phénomène physique (ex. écoulement souterrain, recul d'érosion de versant, écoulement canalisé)
- 2ème hypothèse : on disposerait des données nécessaires au modèle : les données alimentant les variables en entrée, existent ainsi que les paramètres du modèle physique
- 3ème hypothèse : la base de données est suffisante et contient toute l'information.

Il existe une lacune importante non seulement sur les données mais surtout sur la connaissance du processus que l'on cherche à simuler. L'information disponible n'est plus suffisante : il faut se décaler et se placer au niveau de la Connaissance de l'objet considéré.
Une application directe de ceci est la suivante : on se propose d'établir l'interopérabilité entre 2 systèmes :

- une base de données de gestion d'infrastructure qui contient l'ensemble des canalisations et des ouvrages d'un réseau d'assainissement en milieu urbain
- un outil de dimensionnement de réseau qui modélise les écoulements attendus en chaque section du réseau.

A priori toute l'information utile au modèle est contenue dans la base ; un modèle déterministe d'écoulement impose 2 choses : un système mathématique à base d'équations différentielles (équations de Barré de Saint-Venant) et de la géométrie du dispositif à savoir le réseau sous forme d'un graphe où chaque arc est caractérisé par sa longueur, les diamètres et des paramètres d'écoulement (friction, rugosité..). Or pour passer d'une information très détaillée contenue dans le SGBD de gestion, à la structure de données nécessaire au modèle on utilise l'expertise humaine. Le schéma de données est établi en fonction de certains critères : choix des nœuds du réseau ayant un impact hydraulique remarquable (raccordement hydraulique, pompe…) et suppression des canalisations négligeables. On peut donc imaginer un système d'assistance à la modélisation (système expert ou analyse multicritère).
Or, ce qui constitue pour nous un véritable enjeu demeure : il existe une lacune importante non seulement sur les données mais surtout sur la connaissance du processus que l'on cherche à simuler.
La question devient alors : l'information disponible est-elle suffisante, pour envisager le niveau "connaissance" de l'objet considéré ?

Les chapitres et volets suivants s'attacheront à répondre (du moins partiellement) à :

- peut-on structurer toute donnée et /ou information imparfaite et/ou incomplète afin d'accroître la connaissance du système ? Ce premier point constitue un enjeu majeur de l'axe de recherche particulièrement intéressant du volet 2,

- n’existe-t-il pas un moyen d’identification ou d’extraction de toute forme d’information à partir de donnée et /ou d’information imparfaite et/ou incomplète afin d’accroître la connaissance du système ? le volet 3 tentera de fournir des éléments de réponse substantiels et objectifs à cette question.

Publications et rapports produits :
- 7 articles en colloque avec actes,
- Thèse de doctorat M. Batton-Hubert (Hubert 1993),
- DEA M. Batton-Hubert (Hubert 1993)

2. Interopérabilité entre modèles déterministes en Hydrodynamique -Hydraulique

Au printemps 1998, une association de protection de la nature, la Frapna Loire nous a sollicité pour développer une proposition de recherche sur l'évaluation des échanges hydrauliques existants sur un site post-industriel d'exploitation de granulats afin d'améliorer ou favoriser l'intérêt écologique du site. Ce projet soutenu par la Région et la société Morillon Corvol (Groupe RMC) s'est déroulé de 1998 à 2001 sous ma responsabilité. Il a suscité de nombreux contacts et collaborations avec des partenaires locaux comme la DDE 42, l'université Jean Monnet (Saint-Étienne), l'université Claude Bernard (Lyon1)
Une présentation brève de l'ensemble du projet permet de nous focaliser sur les éléments prépondérants liés à la thématique de recherche du Volet 1. Nous avons choisi de taire l'enjeu environnemental, remarquable que constitue le site de l'Ecopôle du Forez, précurseur en terme de réaménagement post-industriel réussi sur l'ensemble du territoire français ainsi que l'aspect aménagement du territoire au travers des différents acteurs cohabitant autour du site.

La formulation de ce projet, à savoir évaluer et quantifier des flux hydrauliques, appartient aux Sciences de l'Ingénierie de l'Environnement et fixe le domaine d'application d'une recherche thématique autour de la gestion de l'information spatialisée et son adéquation avec la modélisation déterministe adoptée.

2.1 Objets de l'Etude : aide à la Réhabilitation d'un site post-industriel

Le contexte de cette étude est celle de l'aide à la réhabilitation d'un site post-industriel d'exploitation de granulats, conditionnée par 2 enjeux : i) limiter les perturbations provoquées par des écoulements souterrains, ii) éviter l'eutrophisation à terme des plans d'eau en fin d'exploitation. Le site de Ecopôle du Forez a été créé dans les années 1990, lors de l'arrêt de l'exploitation des granulats dans le lit d'inondation. L'industriel a rétrocédé un site post-industriel composé des plans d'eau (anciennes gravières exploitées), à une association de protection de la nature la FRAPNA Loire. Pour développer l'intérêt patrimonial et écologique du site, le gestionnaire de l'Ecopôle doit pratiquer une gestion des niveaux d'eau dans les gravières durant l'année pour :
- assurer l'accueil hivernal des limicoles nécessitant un niveau d'eau bas afin de découvrir les îlots,
- et favoriser la nidification des oiseaux d'eau au printemps-été en maintenant les niveaux d'eau hauts dans les bassins.

De part le contexte hydrologique, la demande en gestion des niveaux d'eau est en contradiction avec les variations saisonnières de la nappe alluviale. De plus, les plans d'eau, enrichis en nutriments lors du passage des crues, connaissent un colmatage de leur berge qui menace à terme leur alimentation souterraine. L'alimentation et le maintien du niveau d'eau nécessaire à chaque période posent un certain nombre de questions:
- l'apport d'eau peut-il être suffisant et maintenu malgré des cycles d'alimentation de la rivière décalés sur l'année ?
- les alimentations en eau sont-elles seulement liées à la rivière ? Quelle est la contribution de la nappe alluviale ? Quelle est l'influence des variations saisonnières (étiage, hautes eaux, crue) sur les courants d'infiltration qui conditionne le mouvement de la nappe alluviale et l'écoulement chenalisé ?

- comment se fait la vidange des gravières : chenaux, drainage de la nappe et/ou de la Loire ? Quelles sont les lois de remplissage ? Comment quantifier les phénomènes d'inertie volumique lors d'une remontée ou d'une baisse du niveau piézométrique ?
- la création de nouvelles gravières entraîne-elle des modifications des écoulements souterrains sensibles au niveau du site ?

Il existe une source de données : le suivi des niveaux d'eau dans les gravières et dans la nappe alluviale imposée durant l'exploitation.

Le problème peut s'énoncer ainsi : pour évaluer les flux hydrauliques et hydrodynamiques existant sur le site, peut-on proposer une simulation intégrée basée sur la modélisation déterministe de phénomènes ayant des échelles de temps différentes (écoulement lent souterrain de quelques m/jour et débit instantané en rivière de ~1000m/s en eau de surface, variant sur quelques heures) et évoluant sur des espaces et volumes reliés par des lois de parois ou de frontières ?

2 .2 Méthodes

Classiquement, on utilise le couplage entre une base de données de type SIG et un *code*[4] numérique, implémentation d'un modèle physique déterministe pour l'hydrodynamique et/ou hydraulique (figure 1.2.1). Prenons le cas d'un modèle de type différences finies de l'écoulement souterrain (Modflow, Newsam.) : les conditions initiales et aux bords ainsi que la géométrie du système sont les données d'entrée du modèle, fournies via le SIG sous la forme de matrices de données calquées sur le schéma de discrétisation. Une fois les données en entrée du modèle fixées, le code de calcul prend le relais sur le SGBD. Lorsque les résultats convergent, le *code* de calcul fournit finalement le résultat, les hauteurs d'eau en chacune des mailles, qui seront finalement visualisées et archivées dans le SIG. Ce couplage sans rétroaction n'est pas satisfaisant : en effet, il existe une évolution espace-temps des fronts d'échanges hydrauliques. Ces échanges sont liés aux périodes de basses eaux ou hautes eaux du fleuve. Lors de crue, les anciens chenaux de la rivière sont activés et les gravières capturées au niveau de seuils de débordements.
Ce qui entraîne :
- une modification de la géométrie du front qui s'agrandit avec l'importance de la crue (figure 1.2.1),
- une perturbation des écoulements 3D du sous-sol en imposant un potentiel de charge hydraulique ponctuellement sur la nappe liée à la modification de la loi physique de son comportement.

Il existe entre deux temps de simulation (t + dt) du processus d'écoulement souterrain des variations perturbant le bilan de quantité de mouvement et des grandeurs de densité (masse, énergie..) établi en chacune des mailles qui doivent être intégrées à la modélisation.
Il n'existe pas d'interopérabilité mais seulement un déroulement séquentiel entre les modules de calcul et le SGBD. Nous avons proposé un schéma de simulation interactif de processus dynamiques travaillant à des pas de temps et d'échelle de temps, non homogènes présenté ci-dessous.

[4] Code : terme souvent utilisé en analyse numérique et mécanique des fluides pour désigner un programme ou ensemble de programmes informatiques, schéma de la résolution numérique d'un modèle mathématique

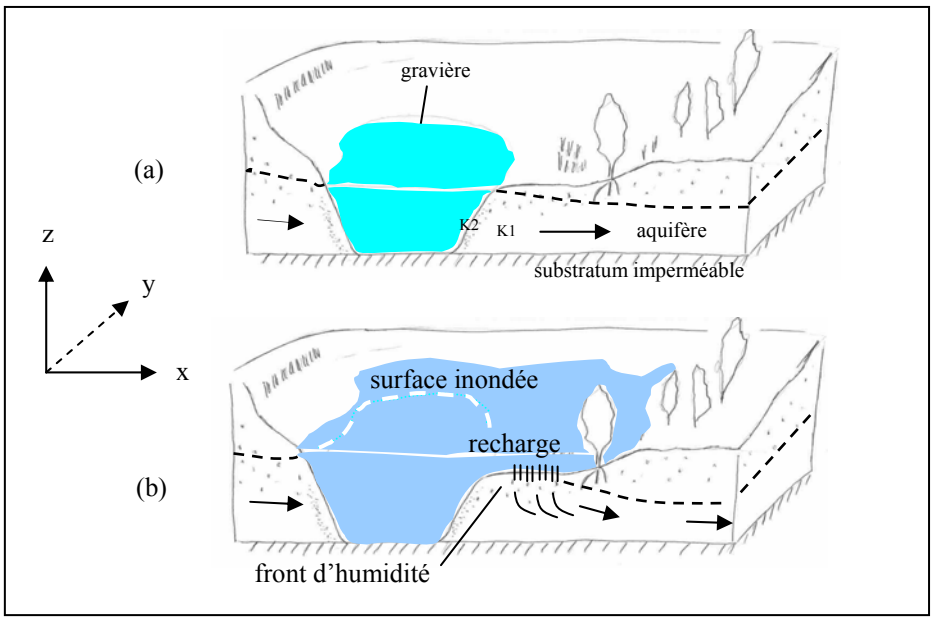


Figure 1.2.1. Structure d'écoulement à proximité d'une gravière : situation basses eaux (a) et hautes eaux (b) (Mimoun 1999)

2.2.1 Choix d'un Schéma conceptuel de données pour l'interopérabilité

L'adoption d'une modélisation topologique et numérique à ce système hydraulique "compliqué" (Hubert 1993) a permis d'établir le Schéma conceptuel et fonctionnel du modèle intégré *hydro* suivant (figure 1.2.2) : le système dynamique est composé de trois entités [écoulements souterrains, les écoulements en rivières, gravières]. Chaque entité hydrologique est identifiée par son comportement et sa contribution dans le bilan des échanges en eau (Batton-Hubert *et al*. 2000) :

- l'entité nappe alluviale soumise à la loi de Darcy,
- l'entité rivière, pour les écoulements de surface (équations de Barré de Saint Venant),
- l'entité *gravière et bassin*, dont le bilan hydrologique entre les précipitations, l'évaporation et le stockage d'eau dans la gravière,
- l'entité *front d'échange* entre les eaux de surface et l'eau souterraine au niveau des transferts bilatéraux localisés le long des interfaces nappe-rivière.

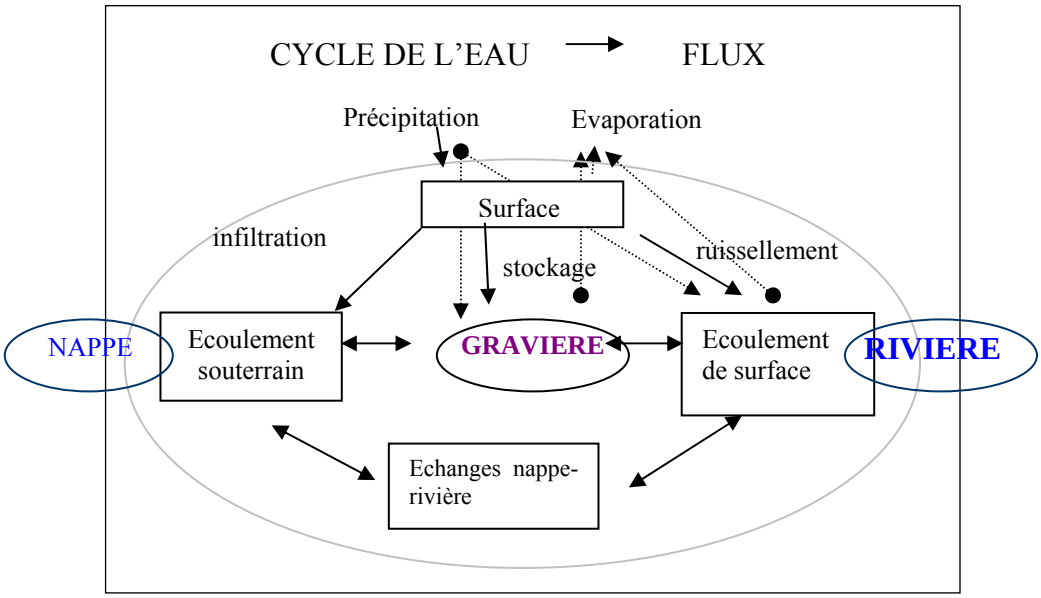


Figure 1.2.2**.** Schéma fonctionnel du modèle intégré : fronts d'échanges hydrauliques de type nappe- gravière et de type gravière- rivière

A chacune des entités *hydro* (rivière, nappe, gravière) est associé un modèle physique d'écoulement. Les données contenues dans le SIG peuvent alimenter un code, qui est l'implémentation d'un modèle physique déterministe pour l'hydrodynamique et l'hydraulique (et plus généralement en mécanique des fluides).

2.2.2 Identification de 2 types de gravières

Il existe deux types de comportements caractéristiques des bassins et de leurs échanges avec la nappe et la rivière : les bassins fermés non reliés aux écoulements de surface et les bassins ouverts en communication avec la rivière.

- Gravière de type fermé :
En période de basses eaux, le bassin a globalement le même comportement que l'aquifère. En période de hautes eaux, l'absence de corrélation entre le niveau d'eau dans le bassin et la piézométrie indique l'existence d'échanges aquifère-rivière au travers de l'interface des 2 milieux. La rivière, à partir d'un débit donné, recharge la nappe d'accompagnement. Consécutivement, cette recharge de l'aquifère des alluvions provoque l'élévation du niveau d'eau dans la gravière, et induit un écoulement dirigé de la rivière vers la gravière. La figure 1.2.3 indique que la recharge sera d'autant plus importante que les épisodes de hautes eaux seront maintenus dans le temps : dans ce cas, la gravière est hydrauliquement liée à l'aquifère des alluvions (Mimoun 1999). L'alimentation en eau de ces bassins ouverts dépend des précipitations et des apports d'eau souterrains latéraux en période de basses et moyennes eaux (nappe aquifère) (figure 1.2.1.a)

- Gravière de type ouvert :
Certains bassins sont plus ou moins reliés au fleuve. S'il s'agit d'anciens chenaux actuellement déconnectés du fleuve à la suite d'endiguements, la rivière en période de hautes eaux peut entraîner des ruptures ou remontées des eaux au niveau des points bas de la digue. Certains bassins sont équipés de buses qui assurent la connectivité hydraulique avec la

rivière ; en période de hautes eaux, la rivière se déverse directement par surverse dans la gravière. L'alimentation en eau de ces bassins ouverts dépend :

- des précipitations et des apports d'eau souterrains latéraux en période de basses et moyennes eaux (nappe aquifère),
- des volumes d'eau déversés par la rivière en période de hautes eaux, par ruptures ou remontée des eaux au niveau des points bas de la digue.

En fonction de l'époque hydrologique, les flux échangés vont dépendre (en quantité scalaire et vectorielle) du gradient de charge hydraulique en chaque point de l'aquifère et du bassin. Lors des épisodes de crue, le bassin alimenté par les eaux de rivière va se comporter comme un bassin d'infiltration qui va recharger l'aquifère (figure 1.2.1.b). Le front d'échange entre la gravière et la nappe s'agrandit et son comportement évolue vers une recharge de nappe; ce dernier étant la conséquence directe de son extension sur la surface topographique.

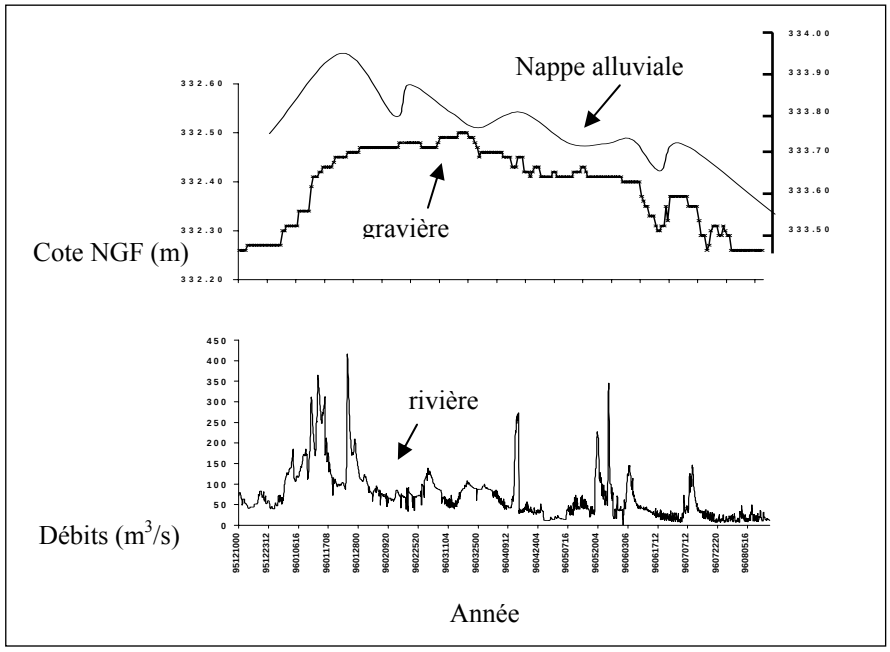


Figure 1.2.3. Gravière de type fermé : courbes des hauteurs d'eau dans la nappe, le bassin principal et les débits de la rivière (12/1995 à 08/1996)

2.2.3 Modèle objet de l'entité *hydro*

D'un point de vue conceptuel, une entité *hydro* est associée à une classe d'objet, décrit par des types abstraits de données et possédant les propriétés d'une structuration de type objet. L'objet doit porter à la fois ses propriétés (attributs) et ses relations (liens entre class).
On va distinguer la forme, un espace 3D dans lequel les propriétés physiques sont identifiables par un même comportement hydraulique. Un objet est décrit par sa géométrie et par un modèle physique/mathématique :

Une gravière est définie par :

- la géométrie de la gravière, assimilable à un graphe orienté composé d'arcs, et de sommets délimitant le domaine,

- une loi de remplissage/vidange caractérisant le réservoir :

$$q(t) = K H^n$$

Avec
- K perméabilité des berges, du fond du réservoir (m/s)
- H^n variation du niveau d'eau dans le bassin (m)
- q(t) débits d'échange (m^3/s)

- un comportement fixant les débits d'échange q(t) en fonction des apports souterrains et les écoulements de surface (bassin relié à la rivière par une buse, débordement de la rivière lors de hautes eaux).

La rivière est définie par :
- la géométrie, portée par le graphe associé (Hubert 1993),
- un modèle hydraulique en 1D, utilisé pour évaluer les vitesses et les hauteurs au niveau de section en travers placées aux nœuds du graphe,
- des sections en travers, associées chacune à un nœud du graphe de la rivière, portant la hauteur d'eau et le débit,
- le débit, calculé à partir du système des équations de Barré de Saint-Venant :

$$\frac{\partial S}{\partial t} + \frac{\partial Q}{\partial x} = q$$

$$\frac{\partial Q}{\partial x} + \frac{\partial}{\partial x}\left(\beta \frac{Q^2}{S}\right) + gS\frac{\partial z}{\partial x} = -g\frac{Q^2}{K^2 S R h^{4/3}} + kq\frac{Q}{S}$$

Avec : S section mouillée (m^2)
Q debit(m^3/s)
q débit d'apport latéral
g accélération de la pesanteur
β , k coefficients de quantité de mouvement
Rh rayon hydraulique
K coefficient de Strickler

On peut définir de la même façon tous les objets concernés de type linéaire ou surfacique : la rivière, l'interface nappe-aquifère (front d'échange) (Batton-Hubert et al. 2000).

- Une nappe aquifère est un volume de sol ayant des caractéristiques de perméabilité et de transmissivité qui vont conditionner la loi d'écoulement des eaux souterraines (loi de Darcy). Les hétérogénéités sont liées aux structures géologiques souterraines, son extension est donnée par le toit de la nappe et son substratum. L'objet nappe est une entité 3D dont on peut proposer schéma de type 3D (Hubert 1993) mais qui malheureusement ne peut être implémenté directement en tant qu'objet 3D dans un SIG (2,5) de par les lacunes des SIG concernant le 3D.

Les équations de l'écoulement souterrain (équation de continuité et Loi de Darcy) sont résolues par un schéma numérique implicite (ou explicite) qui utilise un maillage régulier sur lequel on évalue la charge piézométrique à partir des bilans évalués en chaque maille, par l'équation de Boussinesq :

$$\frac{\partial}{\partial_x}\left(T_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial_y}\left(T_y \frac{\partial}{\partial} \frac{h}{y}\right) = S \frac{\partial h}{\partial t} + q$$

avec :

	T	transmissivité (m^2/s) suivant les 2 directions de l'espace
	h	hauteur piézométrique ou potentiel (m)
	S	coefficient d'emmagasinement spécifique (m^{-1})
	q	débit échangé avec le milieu extérieur par unité de surface (m^3/s)

En entrée du modèle discret, le mode raster des SIG permet d'instancier chaque couche nécessaire : à savoir le toit de l'aquifère par la topographie (nappe libre), la profondeur du substratum, la perméabilité de la ou des formations géologiques (en modèle 2D d'un aquifère ou d'un empilement d'aquifères). Chacun de ces *layers* permet de reconstruire les volumes de la nappe. La topologie implicite portée par un maillage régulier (raster) contribue à conserver ce mode matriciel afin d'alimenter directement le module numérique d'hydrodynamique. Il n'y a pas de véritable instance de l'objet nappe, il est systématique reconstruit lors de l'appel au modèle. L'interface avec la nappe aquifère se fait par le débit latéral pouvant alimenter la rivière (basses eaux). Le débit obtenu en chaque nœud Q (ou la hauteur d'eau) sera injecté dans les conditions limites du modèle hydrodynamique. Tous deux sont instanciés pour un sommet du graphe de la rivière.

2.2.4 Interopérabilité gérée par un échéancier de type Sequencing Set

Devant s'adapter à des codes de calcul (Newsam, Modflow, Caredas) déjà existants et à des bases de données sous un SIG (Arcview, ArcGIS) pour lequel le schéma numérique devait être conservé (matriciel), il s'agissait d'obtenir l'interopérabilité de ces outils en élaborant un échéancier de type de Sequencing Set (Dahl 1966, Hubert 1993), des processus dynamiques (algorithmes - semblables aux méthodes de la programmation *objet* actuelle) modélisant l'impact structurel sur les TAD de l'entité *hydro*.

Une CLASS de processus informatique a été développée : chaque module peut se comporter comme (figure 1.2.7):

- le calcul d'une nouvelle géométrie de l'objet conjointement à un algorithme de propagation du cheminement hydraulique dans l'unité. Il fournit les paramètres et les conditions hydrauliques de la limite (module zone_inondable),
- une interpolation linéaire ou non pour passer d'une donnée ponctuelle à une donnée continue (hauteur d'eau sur les arcs de rivière, module Int_Sect),
- une fonction d'analyse spatiale : comme la fonction *overlays* d'un SIG permettant de superposer de l'information spatiale et reconstruire les données d'entrée à un modèle hydrodynamique, dans ce cas il reconstitue l'entité nappe,
- une fonction de transformation de type de donnée : conversion du mode *vectoriel* vers un mode *raster* (module m_Vec_Rast).

Chaque processus est défini par son état et par ses transitions possibles (appel à un autre processus), sa suspension, sa terminaison, sa réactivation, sa passivation. Deux instances particulières de cette CLASS, sont les modules de calcul hydrodynamique et le module de calcul hydraulique (considérés ici comme des boîtes noires dont seules les entrées-sorties sont maîtrisées) et qui sont à leurs tours encapsulés dans un processus d'état dynamique.
Le gestionnaire de simulation comporte une liste circulaire de processus ordonnés dont l'ordre varie en fonction d'un seul processus, le processus actif. Un exemple d'enchaînement de cette

simulation est fournie en figure 1.2.7. Le déroulement du SQS produit la simulation de différents évènements (bilan d'eau sur une année avec des épisodes de crue/étiage).

2.3 Résultats

Les principaux résultats numériques ont permis de quantifier les aménagements nécessaires (vanne d'arrivée d'eau, hauteur d'inondation annuelle pour la pisciculture et l'avifaune) sur le site de l'Ecopôle du Forez décrit en (Annexe 1.1), mais également, ont servis à calculer le potentiel écologique et hydraulique d'une ancienne gravière par analyse multicritère (Mimoun et al. 2004).

2.3.1 Choix du site d'installation des points de suivi des niveaux d'eau

La première phase de réalisation technique a consisté en l'équipement de ce site en matériel de suivi de niveaux d'eau et de la qualité de l'eau à moyen terme avec le choix des sites de mesure (annexe 1.2.2). Une difficulté souvent rencontrée dans le domaine de l'Environnement est de trouver l'emplacement optimal des points de mesure afin d'obtenir un échantillonnage représentatif du phénomène observé, avec peu d'informations disponibles, soit : [peu de données et de mesures ; pas de schéma de fonctionnement du système ; encore moins un modèle]. La solution est de suivre une heuristique par amélioration successive : on propose un premier modèle rustique, du système caractérisant ce que l'on connaît du site, puis on itère un réajustement jusqu'à obtenir une solution satisfaisante opérationnelle.

Ces résultats du prototype d'interopérabilité *hydrodynamique - hydraulique* de surface après validation ont permis de définir des critères du choix de l'emplacement des stations de mesure puis de spécifier la fréquence d'acquisition de la mesure des hauteurs d'eau au niveau des piézomètres, et enfin mettre en place un protocole d'installation (diamètre de puits, référence de côte NGF...) et d'acquisition (fréquence des relevés, nombre de personnes, suivi de l'état des stations, transfert de données par format compatible avec la base de données géographiques ...). Outre ces critères hydrauliques, des critères de proximité pertinents tels que : les remontées de substratum, les anciens chenaux actifs ou non, la configuration des piézomètres déjà existants (sur site industriel d'exploitation de granulats et du site d l'Ecopôle ont été intégrés.

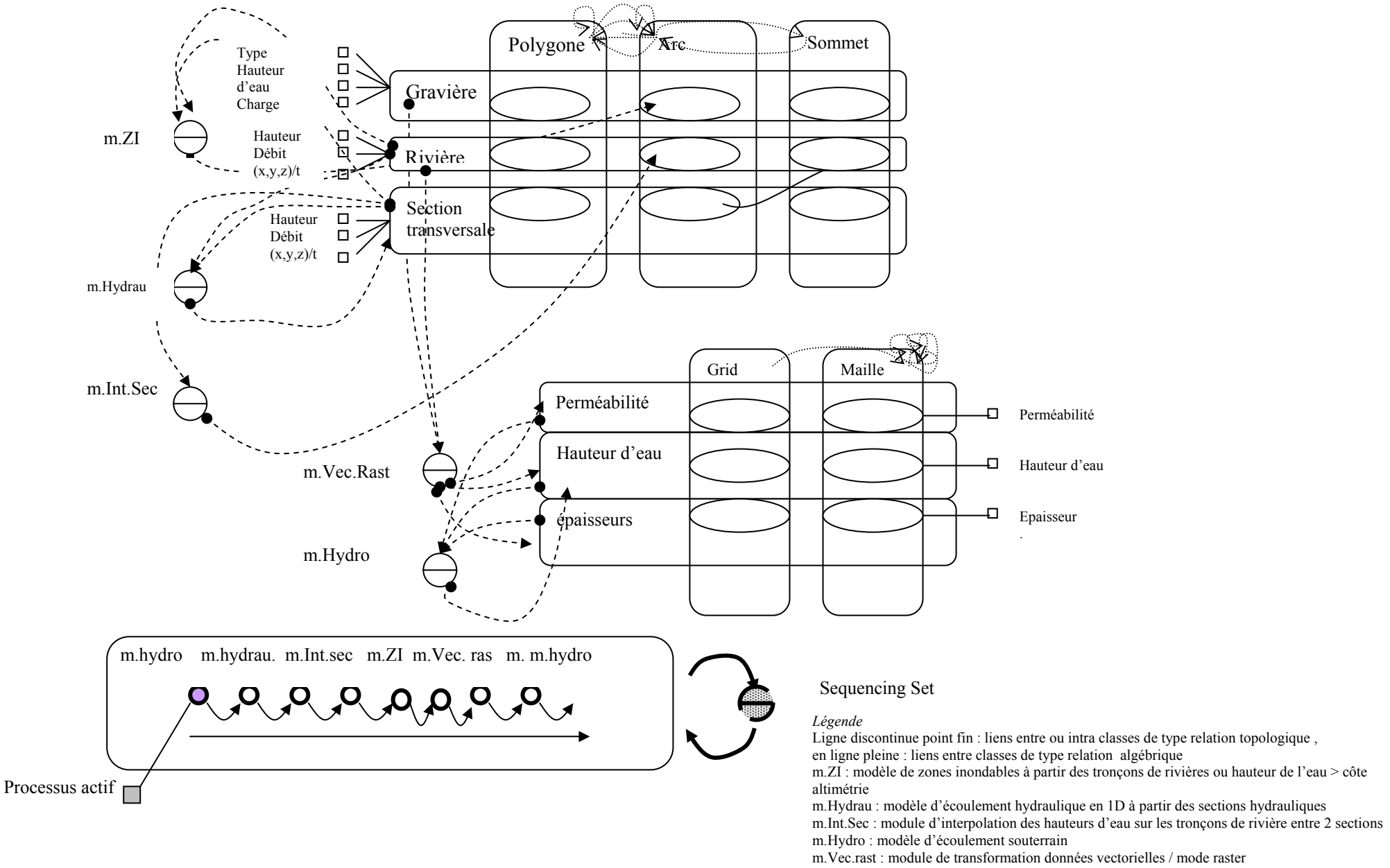


Figure 1.2.7 Interopérabilité pour une modélisation hydraulique et hydrodynamique de type SQS : un ensemble de processus algorithmiques est pris en charge par un gestionnaire de simulation de type SQS , qui entraînent une modification de la structure du modèle de données avec création dynamique ou non d'objets de type arcs, sommets et de leurs topologies, des processus de type codes numériques de calcul des hauteurs d'eau . Exemple de séquence instantanée dans le SQS

2.3.2 Suivi « quasi-temps » réel des niveaux d'eau

Les résultats de la modélisation hydrodynamique ont permis de simuler les écoulements souterrains et le suivi en quasi temps réel des niveaux d'eau (figure 1.2.7 - Annexe 1.2). Ils mettent en évidence des lacunes déjà bien connues comme : la méconnaissance de la proximité des zones d'influence de la rivière et des gravières, des observations incomplètes lors de crue et d'étiage notamment quantitatives (débit, hauteur) pour la validation, des erreurs et imprécisions sur les données de hauteur (mauvaise côte non NGF), le problème de calage difficile sur données ponctuelles et anciennes.

D'un point de vue opérationnel, la simulation d'un tel environnement hydraulique permet d'anticiper une gestion à court terme (fermeture de vanne et protection contre les crues et lâchers de barrage) et à plus long terme, lors de réaménagement afin de favoriser les zones piscicoles par ouverture et fermeture de chenaux et la mise en communication de bassins.

2.4 A retenir

Outre l'aspect opérationnel et l'application de la méthode sur site industriel ou post industriel pour évaluer des indicateurs de suivi du plan de réaménagement pour le graviériste, qui est motivant pour le chercheur, on retiendra trois points importants :

- la reproductibilité et l'adaptation de cette démarche de la modélisation et de simulation de processus physiques dans un environnement *anthropique.*
- la nécessité d'identifier et d'intégrer d'autres modélisations et notamment de phénomènes d'échange entre différentes entités.
- enfin, faire une rétrospection sur le problème soulevé au §1.3 (chapitre1.) concernant l'approche conceptuelle de l'information et de la donnée.

Travaux et bibliographie produits :

- DEA D. Mimoun - Etude des échanges Loire/nappe alluviale à différentes échelles à l'aide d'un SIG et de modèles hydrauliques et hydrodynamiques, DEA « Sciences et Techniques du Déchet », Ecole Nationale Supérieure des Mines de Saint-Étienne, 1999
- Doctorat D. Mimoun - Spatialisation de l'information : une aide à l'analyse hydraulique et paysagère développée lors de la réhabilitation des sites post-industriels - cas de réaménagements des gravières en milieu alluvionnaire, Ecole nationale supérieure des Mines de Saint-Étienne - Université Jean Monnet, 2004
- 3 articles en colloques avec actes + 1 article en revue
- 1 contrat de recherche FRAPNA Loire-ARMINES, 1998 - 2001

3. Amélioration de modèles de simulation déterministes en hydrodynamique

3.1 Contexte

Revenons au domaine d'application de ces recherches et à leurs applications en Ingénierie de l'Environnement : on s'attache à fournir au décideur des éléments de réponse en terme de quantification d'impacts et d'indicateurs par rapport à un problème donné. Le paragraphe précédent (§2.) fournit un cas d'utilisation de l'interopérabilité entre modèles pour répondre à l'évaluation des perturbations des écoulements souterrains durant et après l'exploitation de granulats en milieu alluvionnaire.

Il existe une préoccupation scientifique sous-jacente sur le concept de la connaissance, de son obtention via la modélisation déterministe du processus d'écoulement: quelle fiabilité peut-on donner réellement à l'information fournie par le modèle ?

Il existe bien identifiées 2 facettes :
- l'objectif "*application en Sciences de l'Ingénieur*" de la démarche de recherche au cours de ces années,
- et une préoccupation scientifique sur le concept d'information et de connaissance, de sa manipulation au travers des outils informatiques mais surtout de son obtention.

Illustrons ce propos :
- lorsque un moteur de simulation assure et restitue les résultats de la simulation de processus physiques synchrones d'écoulement souterrains et superficiels, on obtient la répartition des hauteurs d'eau et des débits (sur une carte et pour une période donnée). L'application *Sciences de l'Ingénieur* aura pour objet de quantifier ces valeurs en terme de valeurs seuils admissibles et lorsqu'il existe plusieurs décisions possibles avec plusieurs décideurs, ces indicateurs seront transformés en critères de préférence par une Analyse multicritère adaptée (Mimoun 2004). Bien que le choix de la méthode d'analyse multicritère avec la scission en critères et l'adaptation de la pondération de ces méthodes est délicat, elle reste pour nous un outil à adapter à notre contexte d'étude (Mimoun, Batton-Hubert 2004).
- lorsque l'on utilise des modèles déterministes en écoulement souterrain, on se heurte aux problèmes bien connus de la modélisation. A savoir : le choix de la discrétisation et de la résolution, le problème de la validité de la modélisation, l'amélioration de la modélisation par couplage et interopérabilité entre modèles. On doit s'attaquer à des problèmes plus fondamentaux en modélisation des écoulements souterrains ; c'est une incursion (ponctuelle et ciblée) dans la théorie des écoulements et leur modélisation.

Restant modeste devant une telle tâche nous nous limitons à deux aspects de la modélisation en Hydrodynamique :

1. l'amélioration de la modélisation en fonction des données et information disponibles,
2. l'aide au choix de la pertinence d'un modèle en fonction de la connaissance du site.

Ce détour ponctuel dans la modélisation même de la Physique nous permet d'améliorer la notion de structuration et de modélisation de la connaissance (qui ne se limite plus à l'information géo-graphique).

3.2 Amélioration de la modélisation déterministe en prenant en compte les frontières des entités hydrauliques

Afin d'approcher une modélisation réaliste d'un processus hydrodynamique, deux étapes complémentaires doivent être effectuées :

- Evaluer s'il est possible d'améliorer le comportement du système en intégrant une modélisation des frontières entre les différents écoulements, comme celui des lois de parois ou de l'écoulement de Couette turbulent.

- Vérifier s'il existe d'autres types de résolution numérique des équations de continuité et de transport (loi de Darcy). Si elles existent, sont-elles applicables et dans quelles conditions. Dans le cadre de la physique newtonienne et on ne remet pas en question à ce niveau les équations qui gouvernent l'écoulement.

Une des faiblesses de la modélisation développée (écoulement 2D par D.F[5] ou E.F[6]) est dans l'approximation des conditions limites imposées (type Dirichlet - hauteur d'eau imposée) ou de type Neumann (débit) qui est consécutive aux changements de régime et de géométrie (débordement de rivière cf. figure 1.2.1). Cette approximation concerne à la fois le choix du type de condition mais également la quantification de cette valeur scalaire ou vectorielle. Il s'agit alors d'améliorer le comportement du système en intégrant une modélisation des frontières entre les différents types d'écoulements, et particulièrement les lois de parois pour un écoulement de type Couette turbulent. La thèse de doctorat de V. Devigne soutenue en 2006, porte sur ce sujet en collaboration avec l'Université de Lyon 1 sous la direction de A. Mikelić, et T. Clopeau (Devigne, Batton-Hubert, Clopeau 2003, Devigne 2006).

3.3 Amélioration de la modélisation en fonction des données disponibles et de la pertinence d'un modèle en fonction de la connaissance du site

Considérons le système hydrodynamique de la zone alluvionnaire comportant une rivière (écoulement canalisé), des plans d'eau (gravière) et une zone saturée pour la nappe aquifère; le site pilote de l'étude est l'Ecozone sur la Loire cité au §1.2.
L'évaluation des écoulement souterrains peut être résolue de différentes manières en fonction de la méthode de résolution des processus d'écoulements et de transport. En général, ces méthodes de résolution peuvent être d'ordre analytiques ou numériques. Des solutions analytiques peuvent être obtenues pour certains problèmes dont les équations sont linéaires ou quasi-linéaires et d'une géométrie simple du domaine. Pour des problèmes contenant des équations non linéaires ou des problèmes à géométrie complexe, des solutions exactes n'existent pas en général. Il s'agit donc d'obtenir une approximation des équations du problème et le résoudre par une méthode numérique. La discrétisation en temps se fait généralement par un schéma aux différences finies. Pour la discrétisation du domaine, plusieurs méthodes discrètes existent dont les méthodes aux frontières et les méthodes aux domaines.

Le principe des méthodes aux frontières est de réduire le problème d'une dimension. C'est généralement effectué par un procédé analytique utilisant une intégration des équations du problème. Les inconnues se trouvent alors situées aux limites du domaine. Ce nouveau

[5] D.F Differences Finies
[6] E.F Elements finis

système d'équations est alors discrétisé pour permettre la résolution du problème. Dans cette catégorie on trouve principalement les A.E.M[7] et les BEM[8].
Pour les méthodes aux domaines, dont celles des FD et des FE, le concept utilisé est la subdivision du domaine calculé en éléments de formes et tailles arbitraires. La seule restriction est que les éléments ne se chevauchent pas et qu'ils recouvrent tout le domaine. Le maillage du domaine calculé est composé d'un ensemble de formes géométriques élémentaires. Le choix de cette géométrie dépend notamment de la géométrie du domaine, de la précision requise, du coût en calcul numérique, etc. Les FE methods requièrent une discrétisation sous forme intégrale des équations du domaine.
Pour chaque méthode numérique, il existe plusieurs façons de discrétiser et de calculer les équations, comme par exemple le choix des fonctions d'interpolations (fonctions de bases). C'est le choix de la discrétisation et du schéma numérique de résolution.

L'existence de différentes méthodes de résolution numérique des équations des écoulements en zone saturée, pose la question de savoir dans quelles conditions peut-on appliquer une méthode au profit d'une autre (Le Grand P, 2003).
Il s'agit alors d'améliorer la modélisation (rendre une modélisation *réaliste*) en fonction des données et établir l'aide au choix de la pertinence d'un modèle en fonction des données disponibles.
L'objet étant d'évaluer l'erreur commise et d'identifier des critères ou recommandations d'usage de certaines méthodes. Afin d'avoir un référant, on utilise un software industriel (type D.F Modflow) et a été retenue une famille innovante de modèles numériques, dite des Éléments Analytiques, peu répandue en France. Cette méthode est dérivée des méthodes aux frontières basées sur une discrétisation des frontières uniquement. Les résultats obtenus pour les deux méthodes (D.F et A.E.M) sont similaires et la principale source d'erreur observée se trouve dans les conditions aux limites de l'aquifère, non dans la discrétisation des équations physiques. On peut alors faire l'hypothèse que le résultat calculé est la forme exacte (ou théorique) que l'aquifère devrait avoir pour un type de conditions aux limites particulières connues. Pour évaluer l'influence et les erreurs commises il est nécessaire de s'intéresser à des formes élémentaires pour lesquelles on peut faire varier la géométrie, la condition hydraulique. Ces indicateurs de forme, de la qualité du maillage (nombre de segment définissant un arc limite de tronçon de rivière) devront être quantifiés à partir des données fournies par un SIG et évaluer sur le cas réel. La thèse de doctorat de F Dauvergne est en cours sur ce sujet en collaboration avec l'Université de Minesotas (O. Strack) (soutenance fin décembre 2006).

Travaux produits :

- Doctorat, P. Le Grand, Formes curvilinéaires avancées pour la modélisation centrée objet des écoulements souterrains par méthode des éléments analytiques, Ecole nationale supérieure des Mines de Saint-Étienne - Université Jean Monnet, 2003
-1 article en colloque
- Doctorat V. Devigne, Ecoulements et conditions aux limites particulières appliquées en hydrogéologie - théorie mathématique des processus de dissolution/ précipitation en milieu poreux, Ecole nationale supérieure des Mines de Saint-Étienne - Université Jean Monnet, 2006,
- 1 thèse dont la soutenance est prévue en décembre 2006

[7] A.E.M : Analytical Element Method
[8] B.E.M : Boundary Element Method

Références

Colette Y., Siarry P. (2002),Optimisation multiobjectif , éditions Eyrolles 2002, Paris

Guttag J. (1977), Abstract data and the development of data structures Comm. O of the ACM, Vol. 20, 1977

Bouillé F. (1977), Un modèle universel de banque de données simultanément partageable, portable et répartie. Thèse de Doctorat d'Etat de l'Université P et M. Curie , 1977, 550p.

Dahl OJ (1968), Discretes events simulation languages. Programming Languages Ada press 1968

Dahl OJ, Nygaard K.(1966), Simula : an algol based simulation language CACM v9 n°9, 1966 pp671-678

Annexe 1.1 Site expérimental Ecopôle du Forez

Le site Ecopôle du Forez se localise en bordure du fleuve Loire, au cœur de la plaine du Forez (42-France). Délimité à l'Ouest par des formations oligo-miocènes faiblement vallonnées formant des basses terrasses, le site correspond au lit majeur du fleuve. Il couvre 6 km de fleuve pour une largeur de 3 km. Deux unités se distinguent :

- entre le fleuve et les levées de terre, une zone dont l'altitude est inférieure à 5 mètres par rapport au fleuve. Cette zone s'inscrit entièrement dans la zone de débordement à fort courant de la rivière.
- La plaine alluviale délimitée vers l'Ouest par un ruisseau (l'Aillot). D'une altitude voisine de 5 mètres par rapport au fleuve, elle correspond à la zone d'inondation maximale de la Loire à faible courant.

D'un point de vue géomorphologique, les alluvions récentes et actuelles sont constituées de sables et graviers sur une épaisseur moyenne de 3 à 4 mètres reposant sur un substratum imperméable, les marnes vertes. La faible épaisseur d'alluvions fluviales et son caractère hétérogène en granulométrie et en texture conduit les gravièristes à exploiter les sédiments jusqu'au substratum imperméable. Ce type d'exploitation est une particularité par rapport au mode d'exploitation sur d'autres sites (ballastières sur le Rhône, Nord de la France).

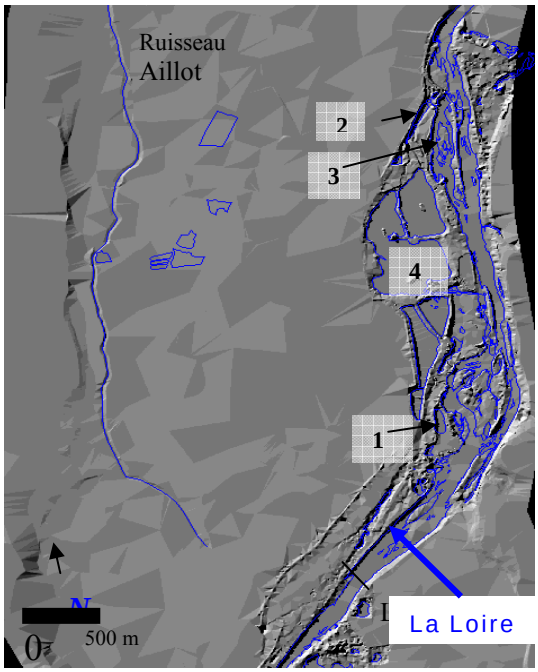


Figure 1.2.4 : Site de l'Ecopôle du Forez : les gravières en 4, la Loire et ses anciens chenaux (1- 2 - 3)

Annexe 1.2 Résultats de simulation et choix de sites de suivi de hauteurs d'eau

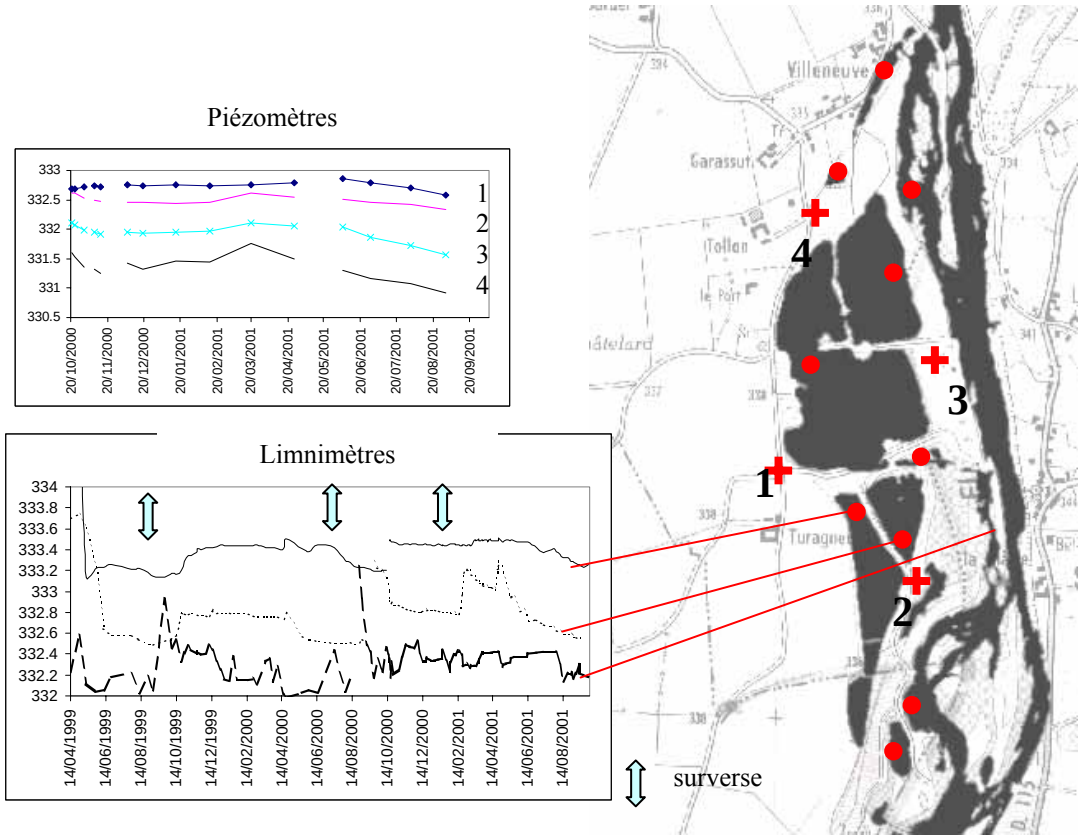


Figure 1.2.5. Suivi et restitution en quasi- temps réel les niveaux d'eau sur le site

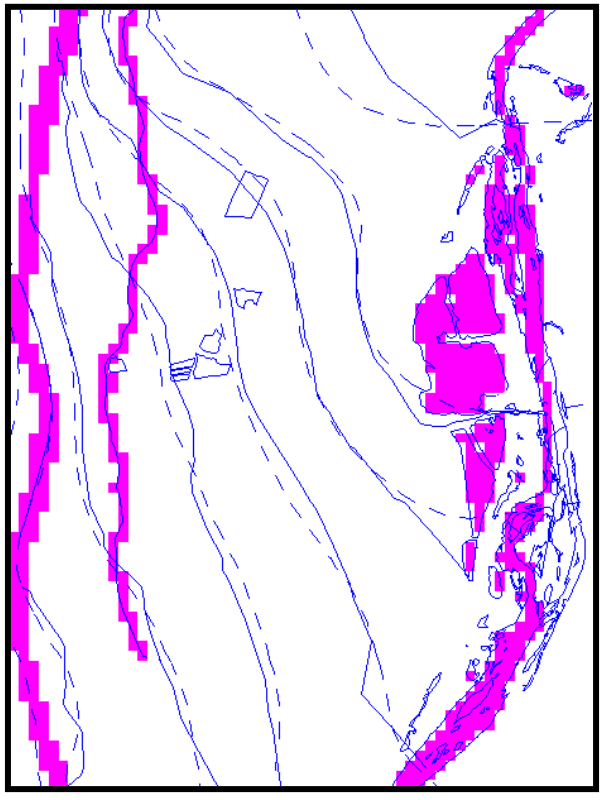


Figure 1.2.6 **:** Résultat d'une simulation hydrodynamique : en trait plein, les iso pièzes calculées, en tiret les iso pièzes observées - maille 50m (Le Grand 2003)

Volet 2 : Analyse *"temps-espace"* de l'information environnementale - Bilan des Recherches

Ce volet 2 s'intéresse aux difficultés liées au développement de modèles déterministes ou statistiques pour la simulation et la prévision d'impacts environnentaux et du diagnostic d'un processus continu concernant l'eau ou l'atmosphère. Qu'il s'agisse d'évaluer une valeur scalaire comme une concentration en polluant ou une valeur vectorielle comme la vitesse, ou bien de diagnostiquer la variation d'un état moyen du système, les données disponibles sont souvent entachées d'imprécisions diverses. Une approche complémentaire consiste à exploiter toute forme d'information disponible permettant d'accroître la connaissance du processus ou du système : une méthodologique est proposée, exprimée sous la forme synthétique d'un prédicat.

Deux types d'application sont développées :

- l'analyse du fonctionnement d'un réseau d'adduction en eau potable au chapitre 1,
- l'analyse de la qualité de l'air urbain et des nuisances olfactives autour d'un site industriel, chapitre2.

Bien qu'*a priori* travaillant sur des domaines environnentaux différents, l'approche fournit les bases qui permettront de proposer une généralisation du traitement des composantes *espace* et *temps* de l'information environnementale au volet 3.

Introduction

Ce volet 2 marque le basculement de mes recherches pour et dans le monde des Sciences de l'Environnement et s'inscrit dans le champ disciplinaire de l'optimisation et la prévision des impacts environnementaux qui me préoccupent depuis mon arrivée à l'ENSM.SE. Evaluer une pression d'eau dans un réseau d'eau potable ou évaluer la concentration de polluant atmosphérique en milieu urbain n'est pas révolutionnaire puisqu'il s'agit d'une préoccupation journalière d'autant plus d'actualité avec les préoccupations environnementales et sanitaires qui marquent ce début du 21ème siècle. Mais qu'y a-t-il de nouveau en soi ? Et quels sont les liens entre ces 2 questions ? Une solution se trouve dans l'approche et la manière de construire et d'améliorer la Connaissance du phénomène pour lequel on doit émettre une quantité qui le caractérise (ou une quantification partielle au moins).

La pratique pour simuler (interpoler ou estimer) la valeur d'une quantité attendue couramment utilisée, est la suivante : à partir d'une description du site et du processus physique intéressé on s'empresse (parfois un peu trop) de fournir à un modèle numérique connu, les données recensées dans la première étape, puis après calage (car il existe souvent trop d'imprécision sur les données en entrée) autour des données (conditions limites, paramètre du système), le modèle dans le meilleur des cas donnera une valeur d'une pression, d'une vitesse température etc.. Ce type de modèles dit de connaissance est basé sur l'interprétation physique du comportement du système : l'identification de ce type de modèles a besoin d'une connaissance exhaustive, qui suppose :

- qu'il existe un modèle mathématique et/ou physique avec un schéma explicite de discrétisation,
- que les données nécessaires en entrée du modèle sont suffisantes, relativement fiables (indice de confiance), continues (*sans trou*) et disponibles dans une base de données plus ou moins sophistiquée.

Or ces 2 hypothèses ne sont pas toujours vérifiées (manque de données, difficulté de calage d'un modèle, données sporadiques) et les moyens de vérifier le réalisme d'un modèle non suffisants. Un modèle mathématique/physique est une représentation d'une certaine réalité. Il est construit avec des hypothèses, selon des théories et pour une connaissance perçue.

Le second type de modèles, dit modèles de représentation établit les caractéristiques d'un système en utilisant des relations mathématiques dépendantes entre les attributs du système (paramètres et variables) Appelés encore modèles statistiques ou de *data mining* (approche floue et neuronale, régression..) ils sont construits à partir des observations et de l'exploitation de l'information intrinsèque à la donnée et à l'ensemble des données. Ils donnent un sens aux données.

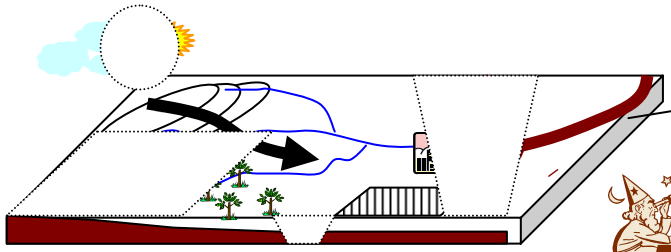
On dispose de deux moyens d'appréhender la connaissance du système : soit on tente de trouver le modèle physique et mathématique basé sur des lois déjà connues, soit on essaye de faire émerger de la connaissance à partir des données.
Cette dualité entre déterminisme et statistique ne doit pas être exclusive mais au contraire un argument de complémentarité : nous cherchons à extraire l'information pertinente qui permettra d'asseoir et de discerner les propriétés du processus considéré. N'est-il pas possible d'utiliser l'analyse de données pour progresser sur le déterminisme et de recourir au déterminisme pour paramétrer un modèle de représentation ?

On pose comme hypothèses fortes que :

- toute forme d'information disponible peut être exploitable,
- il existe un sens univoque entre la donnée et l'objet pour obtenir une certaine image de l'objet (nous parlerons plus loin du terme état du système qui intègre les 4 dimensions),
- il est possible de reconstruire l'objet avec ses propriétés, ses relations d'un point de vue sémantique,
- il existe un modèle déterministe ou stochastique du processus possible,
- on admet les limites explicites des modèles utilisés (comme la non linéarité, la non existence d'une solution analytique, l'optimum est obtenu par heuristique).

En s'inspirant de ces modèles développés en *data mining* et statistique, j'ai retenu et propose 3 points :

- comme lemme, qu'il existe dans la donnée et/ou information collectée, la propriété fondamentale de contenir une information pertinente qui permet de faire émerger d'autres formes d'information qui conduit à accroître la connaissance intrinsèque du processus,
- comme objectif, qu'il ne s'agit pas de construire à tout prix un nouveau modèle mais d'améliorer l'existant en exploitant ses lacunes,
- comme fait, qu'il existe de nombreuses méthodes et outils de traitement de la *donnée* que l'on utilisera à bon escient pour faire émerger cette information complémentaire.

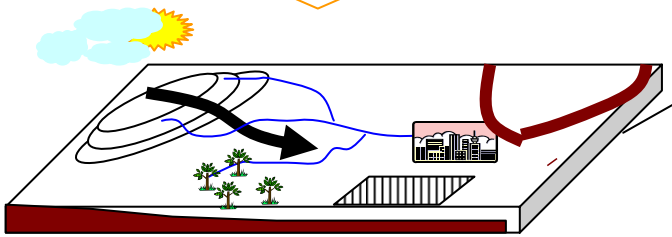


Ce que l'on sait
$f(x_1, x_2, x_3, \ldots x_n)$: Modèle mathématique

Figure 2.0.1 : Cadre méthodologique caractérisation d'un processus physique continu comme étant l'état d'un système en un point donné et à une date donnée par analyse de la connaissance disponible

Phénomène – processus Physique

Connaissance



Ce que l'on voit

❶

S'il n'existe pas de modélisation fiable ET Il existe des Données/observations Ayant certaines propriétés
Alors
On peut générer de l'information

Modèle conceptuel du système 'OBJET

SC Processus

Modèle - Module algorithmique)

BDD Data

BDD Modèles

Données /mesures/observations Quantitatives - temporelles – localisées dans l'espace

❷

Il existe une modélisation ET
Il existe dans les données une information qui caractérise l'Etat du système
ALORS
∃ Aide à la simulation par la connaissance récurrente de certains états du système

Le schéma, de la figure 2.0.1 fixe le cadre méthodologique : le problème est celui lié à un phénomène physique continu (mouvement de l'atmosphère ou écoulement en réseau) dont on veut caractériser un état du système en un point donné pour une date donnée[9], par conséquent prévoir une quantification des caractéristiques du processus sachant qu'il est délicat de l'obtenir par un schéma classique de modélisation déterministe.

Le schéma 2.0.1 introduit la vision de notre préoccupation : à savoir comment approcher la vision temps - espace d'un processus physique et par conséquent de sa connaissance.

Un phénomène ou processus environnemental n'est connu que par ce que l'on sait de lui, d'un point de vue physique et, de par ce que l'on voit de lui soit les observations, composées de données, de mesures directes ou indirectes. Tout ce qui est su finalement de lui, va constituer la *Connaissance* perceptible et actuelle. On peut assigner un modèle conceptuel *objet*, qui prendra sa réalisation en un système composé d'outils informatiques et numériques.
Il existe un lien entre une modélisation physique et la Connaissance du processus : en effet un modèle numérique déterministe peut quantifier le processus et permettre sa perception (vitesse, hauteur, température) ; il accroît le niveau de connaissance du processus en lui-même, sous réserve qu'une connaissance nécessaire, suffisante et disponible lui soit fournie - comment ce seuil d'information minimale peut-il être défini et surtout admissible pour avoir une simulation acceptable d'un point de vue erreur ?

Deux cas se présentent pour la modélisation :

- soit, la connaissance est insuffisante pour faire de la modélisation : alors il faut trouver un autre moyen d'accroître la connaissance du système,
- soit la modélisation est *coûteuse*: en données, en temps de calcul, dans la complication des phénomènes (cas de l'atmosphère), dans le faible gain résultats/investissement.

Ce qui nous a conduit à développer deux approches se formulant de la manière suivante:

> - IL n'existe pas de modélisation fiable ET IL existe des Données/observations ayant certaines propriétés ALORS On peut générer de l'information par analyse de la *Donnée.*
>
> - IL existe une modélisation déterministe ET IL existe dans les données une information qui caractérise l'Etat du système ALORS on peut accroître la connaissance de certains états du système par simulation récurrente.

La première approche est développée dans le contexte des réseaux urbains, avec un réseau d'adduction en eau potable, la second est consacrée à la qualité de l'air en milieu urbain, la pollution atmosphérique et les nuisances olfactives.

Bien que le domaine d'application soit différent, il s'agit toujours d'un problème de phénomène continu (mouvement de l'atmosphère ou écoulement en réseau) dont on peut caractériser un état du système en un point donné pour une date donnée. L'objet est alors un *objet* attributaire et dynamique dont on formalise son évolution. L'objectif à terme est de prévoir une quantification représentative de ce processus mais qui ne peut être obtenue par un schéma classique de modélisation. (Figure 2.0.1)

[9] on parle alors d'approche *espace-temps d'information et processus environnementaux*

Chapitre 1. Analyse du dysfonctionnement d'un réseau d'Adduction en eau potable

Dans le cadre de programmes européens, le centre SITE a été sollicité dés 1998 pour élaborer un projet de recherche sur la conception et l'élaboration d'un *Outil d'aide à la gestion d'infrastructures urbaines pour la ville de Chisinau (Moldavie*), dont la première phase était ciblée sur l'aide au diagnostic du réseau d'AEP par analyse spatio-temporelle des dysfonctionnements et aléas. Le montage de ce projet de recherche pour un Ex pays de l'URSS, démarche difficile et à rebondissements, a abouti au soutien et à une collaboration entre la Région Rhône Alpes, le Conseil Général 42, la Mairie de Saint-Étienne et la Mairie de Chisinau pour la première phase du projet qui nous intéresse directement. Ce projet démarré en 1997 a permis de construire une collaboration scientifique avec l'université Technique de Moldavie.

Rappelons l'enjeu technique d'un réseau d'adduction en eau potable : il doit assurer la distribution d'une eau répondant aux normes de qualité en perpétuant la continuité du service. Une bonne gestion du réseau passe toujours par une bonne connaissance des infrastructures, du fonctionnement hydraulique et l'entretien du réseau. Or avec le temps le réseau vieilli, engendrant des perturbations qui s'observent par une dégradation possible de la qualité du l'eau et/ou de la quantité disponible. Différents facteurs sont à l'origine de ces dégradations. Ils sont d'origines internes (matériaux, diamètre, la résistance à la corrosion du matériau, …), externes (le sol, les fuites, le mouvement des sols…) et liés à l'exploitation du réseau (débit, pression, vitesse, nature de l'eau…).
Le réseau d'adduction en eau potable sur le Ville de Chisinau (Moldavie), pour lequel le vieillissement des infrastructures avec son cortège de symptômes caractéristiques est particulièrement significatif : un nombre élevé d'interventions sur le réseau, une augmentation des pertes de charge, des plaintes importantes concernant la qualité de l'eau. Repérer et diagnostiquer ces dysfonctionnements constitue donc un réel challenge. Rappelons, qu'outre la quantité nécessaire aux 750 000 hab., l'état de la distribution de l'eau est tel qu'il engendre des risques sanitaires importants (mortalité infantile élevée due à des malformations, hépatites chroniques,…).

1. Objet de l'étude : réseau d'Adduction en eau potable de Chisinau

La responsabilité technique du projet, deux missions en Moldavie (Bacle et al. 1997), la collaboration avec les experts locaux, et l'étude plus spécifique d'un quartier de la ville (le quartier de Riscani), m'ont permis d'identifier les particularités du réseau dés 1998 (Batton-Hubert 1998) dont :

- les infrastructures du réseau sont gérées sur des plans cartographiques, des planchettes au 1/500 et 1/2000 non informatisées en 1997,
- le suivi du comportement hydraulique du réseau se fait à partir de douze points de mesure ce qui s'avère insuffisant sur les 1200 km des conduites,
- la mise à jour de la banque de données sur les travaux et la maintenance technique du réseau, se fait avec des pertes d'informations et avec un risque d'erreur important,

- en l'absence de moyens informatiques suffisants, le stockage de l'information complète des paramètres du réseau se fait uniquement sur quatre jours avant d'être archivé d'où un faible historique disponible,
- un faible historique des données sur le suivi des travaux et la maintenance du réseau (6 années),
- le réseau se trouve dans une zone séismique ([6°-7°] sur l'échelle de Richter). Sur une période de 25 ans, il y a eu trois tremblements de terre importants qui ont fragilisé le réseau et ses infrastructures,
- le réseau est surdimensionné,
- les pertes d'eau sur le réseau s'élèvent à 37%,
- sur les 115 km de canalisations à réhabiliter, 100 km nécessitant un renouvellement immédiat,
- il n'existe pas encore de compteur individuel chez tous les consommateurs ;
- l'absence d'un modèle hydraulique de suivi du réseau ;

Une analyse des principales causes et manifestations de vieillissement du réseau a fourni un modèle conceptuel de l'objet *réseau*, et des éléments dynamiques influençant l'état du réseau (Batton-Hubert 1998, Blindu 1999). Dans la figure 2.1.1, le fonctionnement d'un réseau d'AEP peut être représenté par un système dynamique avec :

- des manifestations en surface : observations diverses : plaintes, mesures, visites.
- des causes : vieillissement, pannes, phénomènes extérieurs : inondations, accidents, autres réseaux, environnement urbain,
- des conséquences sur le fonctionnement du réseau : augmentation de la production, perte d'eau, perte d'énergie, durée de pompage, coûts…
- des interventions : réparations, vidanges, nouvelles constructions...

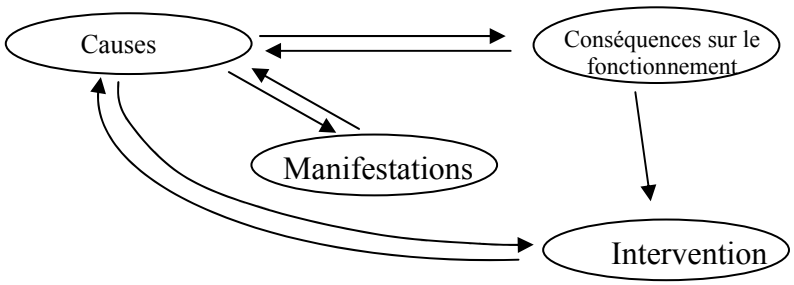


Figure 2.1.1 : Système dynamique de l'analyse des dysfonctionnements

Il apparaît que le vieillissement d'une conduite ou d'un secteur de réseau n'est pas lié à un seul phénomène mais est l'expression conjointe de plusieurs causes entraînant d'autres conséquences, devenant elles-mêmes des facteurs d'influences (causes) plus ou moins directes du mauvais fonctionnement entre l'état attendu (soit la distribution de l'eau prévue en fonction du dimensionnement initialement imposé) et celui observé sur le réseau quotidiennement. Les interventions faites par la suite ne sont que la réalisation d'une prise de décision pour limiter ces conséquences. Ces interventions peuvent être des causes de dysfonctionnement pour lesquelles certaines observations et manifestations de dysfonctionnement seront à leur tour, observées (figure 2.1.1)

2. Méthode développée : analyse spatio-temporelle des dysfonctionnements

2.1 Incapacité et inefficacité de la modélisation hydraulique

Les techniques de diagnostic les plus récentes sont basées sur le concept de redondance de l'information qui utilise un test de cohérence entre un comportement observé du processus et le comportement attendu par une représentation mathématique du processus (Brunet et al. 1990, Ragot J. et al. 1990). Une analyse de cohérence de ces informations permet de détecter des dysfonctionnements puis d'en rechercher les causes. Ainsi le couplage entre une base de données décrivant le réseau et un modèle hydraulique simulant le fonctionnement hydraulique (pression, débit, vitesse) en chaque maille du réseau devrait permettre de mettre en évidence les canalisations ayant des pertes de charge importantes en fonction de l'état observé ou mesuré du réseau ou au contraire les secteurs ayant une forte pression susceptible de provoquer des risques de ruptures.

Or cette démarche impose qu'il y ait suffisamment de données pour construire un état initial et un modèle du processus (ou d'un observateur) pour exploiter l'erreur entre l'information attendue et l'observée.

Les résultats de la simulation hydraulique effectuée, montrent que les erreurs dépassent les tolérances admissibles de 10 à 12 % sur les débits, et ~ 3 m sur les pressions. Or l'état actuel du réseau avec les incertitudes existantes sur l'état des canalisations, sur l'état de fonctionnement du réseau, l'absence de dispositifs de mesures (compteurs, débitmètres, manomètres) aux mailles du réseau, les fortes pertes du réseau, l'indisponibilité de plans côtés numérisés des infrastructures ne permettent pas d'utiliser la modélisation hydraulique. Un important problème de calage initial du modèle (coefficient de perte de charge, diamètre réel, débit initial..) en chaque point fil d'eau du réseau, ne peut être résolu et est rendu impossible dans certaines zones.

2.2 Choix d'un modèle objet

Nous avons émis le lemme que, la donnée et/ou information collectée a la propriété fondamentale de pouvoir contenir une information pertinente, ce qui permet de faire émerger d'autres formes d'informations et qui conduit à accroître la connaissance intrinsèque du processus.

Nous proposons l'approche suivante : en déformant la notion d'erreur, écart entre un comportement attendu et une mesure ou une donnée, on s'attache à identifier une sorte de donnée spécifique qui soit la manifestation d'un comportement erroné du système. Par exemple, [SI inondation ($h \rightarrow \infty$) ALORS $\exists$ rupture de canalisation ou ouvrage] : c'est une observation d'erreur du système. Nous ne recherchons plus à calibrer "l'état normal" du système, mais un état, une image de l'objet. Il s'agit de reconstruire la connaissance de l'objet à partir d'observations complémentaires autres que les mesures *via* un capteur.
Considérons la figure 2.1.1 : les causes ainsi que les conséquences sont des processus dynamiques agissant *via* et par le réseau. Ils sont assimilables à un processus événementiel faisant évoluer la structure de données de l'objet *réseau*.
Les principales manifestations du vieillissement d'un réseau d'AEP sont : les chutes de pression, les fuites diffuses, les ruptures, la détérioration de la qualité de l'eau. Il n'y a pas de données directes ou indirectes (simulées) de Δh, Δp, Δq. Mais il existe une information

indirecte des manifestations de dysfonctionnement, ce sont les observations et surtout les interventions (des fontainiers) sur le réseau.
D'autre part, le réseau est structuré par un schéma conceptuel de type graphe ainsi que son environnement (réseaux urbains, sous-sol..) dans une base de données géo-référencée avec un modèle de données géométriques, de type vectoriel.

Le schéma fonctionnel étant proposé en figure 2.1.2, un lien est recherché entre les causes possibles de dysfonctionnements, et des observations ou des interventions disponibles. Ce lien est établi par l'analyse qualitative et quantitative de tous les aléas survenus sur le réseau uniquement par le biais de leurs manifestations (des plaintes et des interventions).

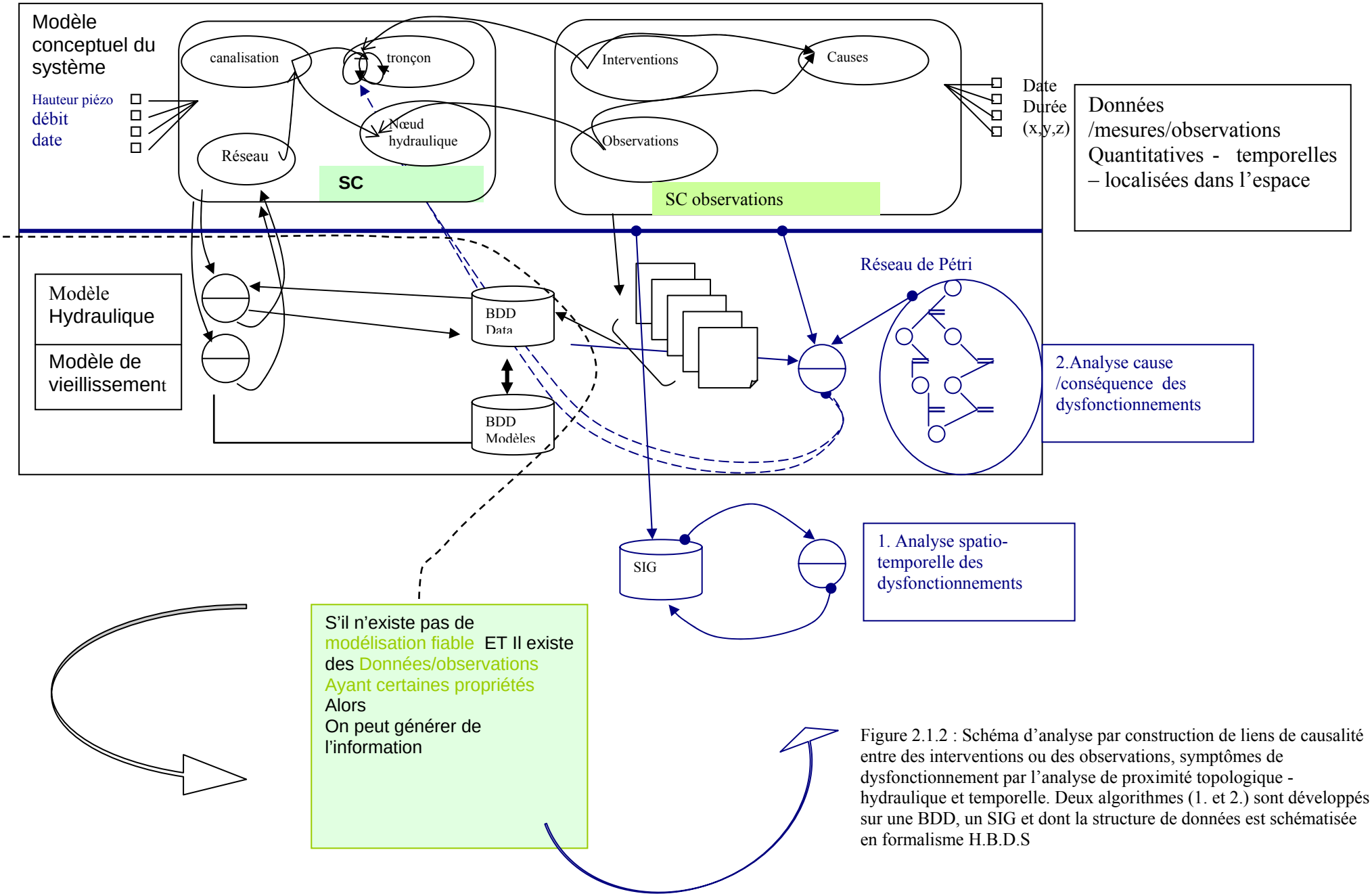


Figure 2.1.2 : Schéma d'analyse par construction de liens de causalité entre des interventions ou des observations, symptômes de dysfonctionnement par l'analyse de proximité topologique - hydraulique et temporelle. Deux algorithmes (1. et 2.) sont développés sur une BDD, un SIG et dont la structure de données est schématisée en formalisme H.B.D.S

2.3 Données disponibles

La maintenance technique journalière du réseau d'eau potable de la ville de Chisinau se fait par un Système de Gestion de Bases de Données de type Oracle. Un fichier informatisé répertorie toutes les opérations effectuées sur l'ensemble des canalisations depuis 1995. Un système de codification permet de différencier différents types d'accidents et travaux survenus sur le réseau. Cette base de données permet de :

- localiser les accidents (nom et numéro de rue),
- connaître les caractéristiques des conduites sur lesquelles l'accident est survenu (diamètre, type de matériau),
- connaître la nature de l'évènement ; par exemple, pour les accidents distinguer une rupture, d'un écoulement en regard, du manque d'eau,
- identifier les travaux effectués (montage du collier, soudure du manchon …)
- dater les différents types d'événements (date de commencement des travaux, date de mise à jour, date de localisation d'accident).

Afin de caractériser les dysfonctionnements, on définira *un événement* comme étant un objet calculé ou observé sur le réseau. Il correspond à toute intervention et/ou accident qui a eu lieu sur le réseau d'eau potable, comme :

- les plaintes concernant les fuites, les pertes de pression, l'absence d'eau, un débit ou une pression trop forte, une mauvaise qualité de l'eau...
- les observations faites au cours d'interventions : fuites diffuses ou ruptures,
- des mesures de pression, de débit, de qualité de l'eau ou des données sur la surproduction de pompage, ou l'arrêt de pompe.

Ces événements sont la manifestation d'un dysfonctionnement.

2.4 Modèle d'analyse spatio-temporelle des dysfonctionnements

Un modèle d'analyse spatio-temporelle des dysfonctionnements a été proposé à partir du schéma fonctionnel (figure 2.1.1), et qui établit un lien entre les causes possibles de dysfonctionnements et des observations (ou des interventions) disponibles afin d'établir le cheminement possible entre les observations, les causes possibles, et d'évaluer alors les conséquences possibles.

L'analyse se focalise sur les différents dysfonctionnements se produisant sur le réseau et qui sont recensés dans la base de données des interventions. Ces événements ont des relations:

- temporelles qui permettent d'identifier parmi les événements répertoriés dans le temps via la date d'apparition, ceux qui se produisent dans un même laps du temps, soit une même date, une même heure, mais à des endroits différents sur le réseau,
- topologiques pour lesquelles la localisation d'évènements permet de répertorier et d'identifier les événements se produisant dans un même quartier, un même secteur, une même proximité (par exemple de requête : recherche d'évènements survenus dans un rayon de 100 m autour du lieu de plainte où une dégradation de la qualité de l'eau a été signalée),
- hydrauliques, lorsque le graphe du réseau et les paramètres physiques du réseau permettent de propager ces événements sur les branches du réseau. Ainsi, l'interdépendance est établie entre des événements se produisant sur un sous-graphe du réseau ayant une source d'alimentation commune (station de pompage, réservoirs, puits, …), ou sur la même maille du réseau, ou sur la même artère, ou sur la même branche, ou encore sur la même conduite.

Ces trois types des relations ont permis de bâtir une méthode d'analyse spatiale et temporelle des dysfonctionnements hydrauliques sur le réseau.

L'algorithme proposé est le suivant : une analyse successive et récursive cherche à détecter la simultanéité de 2 ou plusieurs observations et/ou évènements assimilables à des manifestations de dysfonctionnement, se produisant dans un même laps de temps. Elle met en évidence les relations topologiques et hydrauliques pouvant exister entre les points du réseaux où sont observés les dysfonctionnements.
L'algorithme proposé fait référence à l'information stockée (SIG et BDD), aux bases de données temporelles (à l'aide de requêtes temporelles), à l'analyse spatiale et au raisonnement cognitif.
La simultanéité dans l'espace et dans le temps (date avec décalage) d'un aléa sur le réseau (soit un incident, une plainte, une intervention) repose sur la propriété de filtrer des événements {observations, plaintes, interventions) repérés à la fois :

- temporellement : la date de survenue de chaque aléa est connue,
- spatialement : chaque aléa est affecté à un arc du réseau.

2.5 L'algorithme

Pour chaque manifestation ***i*** {plainte, intervention, observation}, une requête temporelle, permet de rechercher les manifestations de dysfonctionnement se produisant dans un même laps de temps que la manifestation *i,* puis des requêtes spatiales permettent de rechercher parmi ces manifestations celles ayant un lien topologique et hydraulique avec la manifestation *i* (figure 2.1.3) :

- le lien topologique est recherché en effectuant une analyse de proximité basée sur la distance euclidienne (les manifestations survenues dans un rayon de *x* mètres autour du lieu de la manifestation *i,* sont isolées),
- le lien hydraulique, quant à lui, s'appuie sur la structure du réseau. Il consiste à partir de l'arc où est survenue la manifestation *i*, à parcourir les différents arcs sur une longueur de *x* mètres afin de rechercher les manifestations survenues sur ces arcs,
- ainsi, toutes les manifestations de dysfonctionnements ayant une relation temporelle, spatiale et hydraulique avec la manifestation *i* peuvent être isolées. Cette association doit fournir une liste d'événements ayant un lien potentiel avec la manifestation *i* et qui permet de l'expliquer.

Le squelette de l'algorithme comporte 3 étapes (figure 2.1.3) :
- 1) choix de l'événement **i** dont on veut identifier les causes possibles
- 2) pour tout évènement j (j ≠ i), identification de $\mathbf{j_k}$, évènement synchrone à **i** [date_**j** = date_**i** +- *dt*) avec *dt* variant de quelques heures (synchrone) à + ou – 2 jours, retard] : ces requêtes temporelles fournissent la liste $\mathbf{\Gamma j_{k(temps)}}$ des événements $\mathbf{j_k}$ ayant des proximités temporelles avec l'événements **i**
- 3) parmi $\mathbf{\Gamma j_{k(temps)}}$, on recherche de relations topologiques et hydrauliques existant entre $\mathbf{j_k}$ et **i** [distance euclidienne ; puis distance topologie amont/aval, ; topologie de réseau maillé - ligne de courant], identification de $\mathbf{\Gamma j_{k(top\text{-}hydr\text{-}temps)}}$,liste des événements possibles $\mathbf{\Sigma j_{k(top\text{-}hydr\text{-}temps)}}$, à l'origine de l'événement **i**.

Une étape supplémentaire de post-traitement, justifiée au paragraphe 4., a été intégrée au traitement.

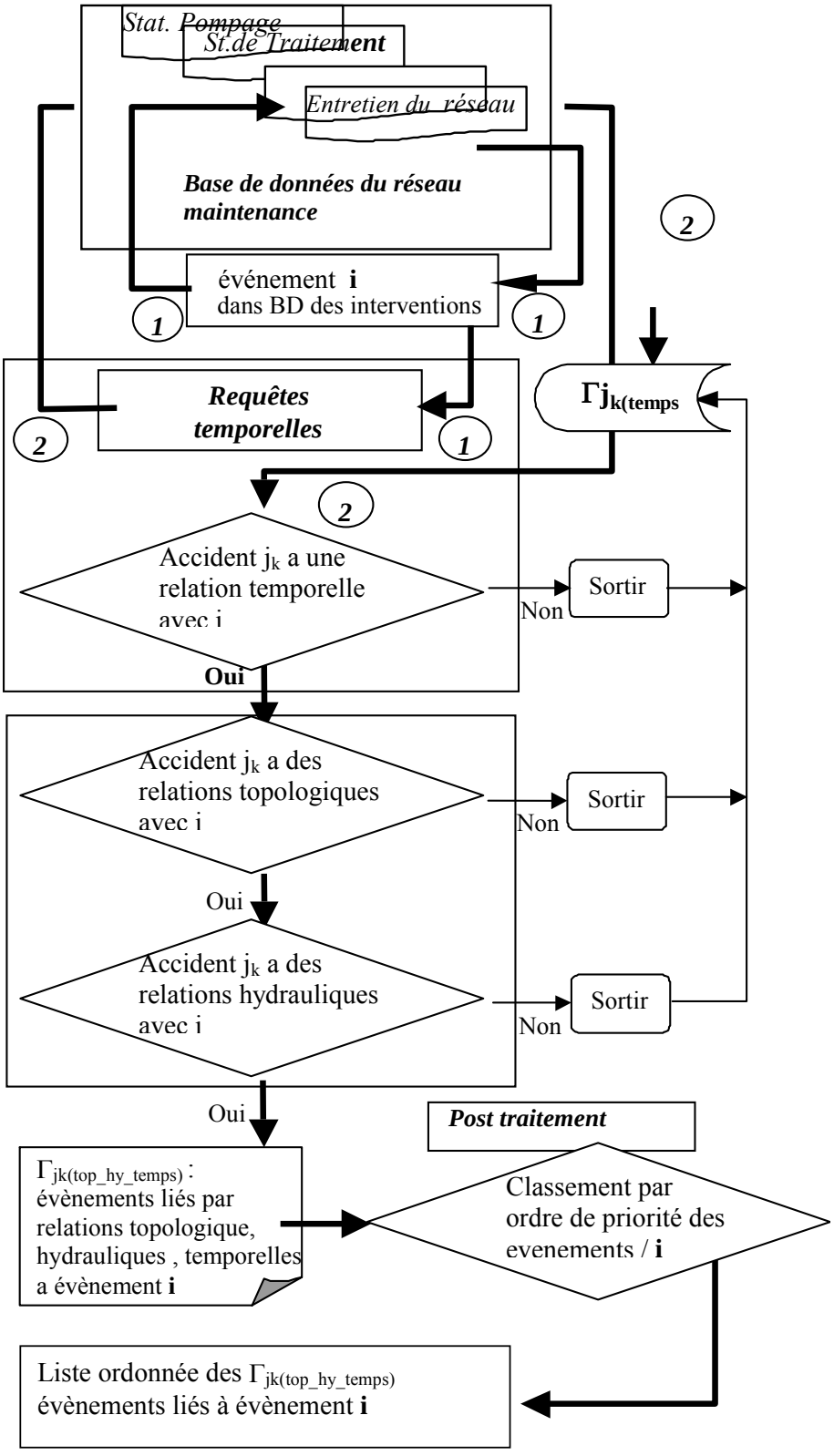


Figure 2.1.3 : Analyse spatio-temporelle : requêtes temporelles et spatiales (topologiques et hydrauliques) (Blindu 2004)

3. Un exemple

Dans ce paragraphe, les résultats obtenus concernent le cas de la plainte associée au manque d'eau signalé le 4/07/1998 dans le bâtiment numéro 2, rue Florilor (figure 2.1.4).

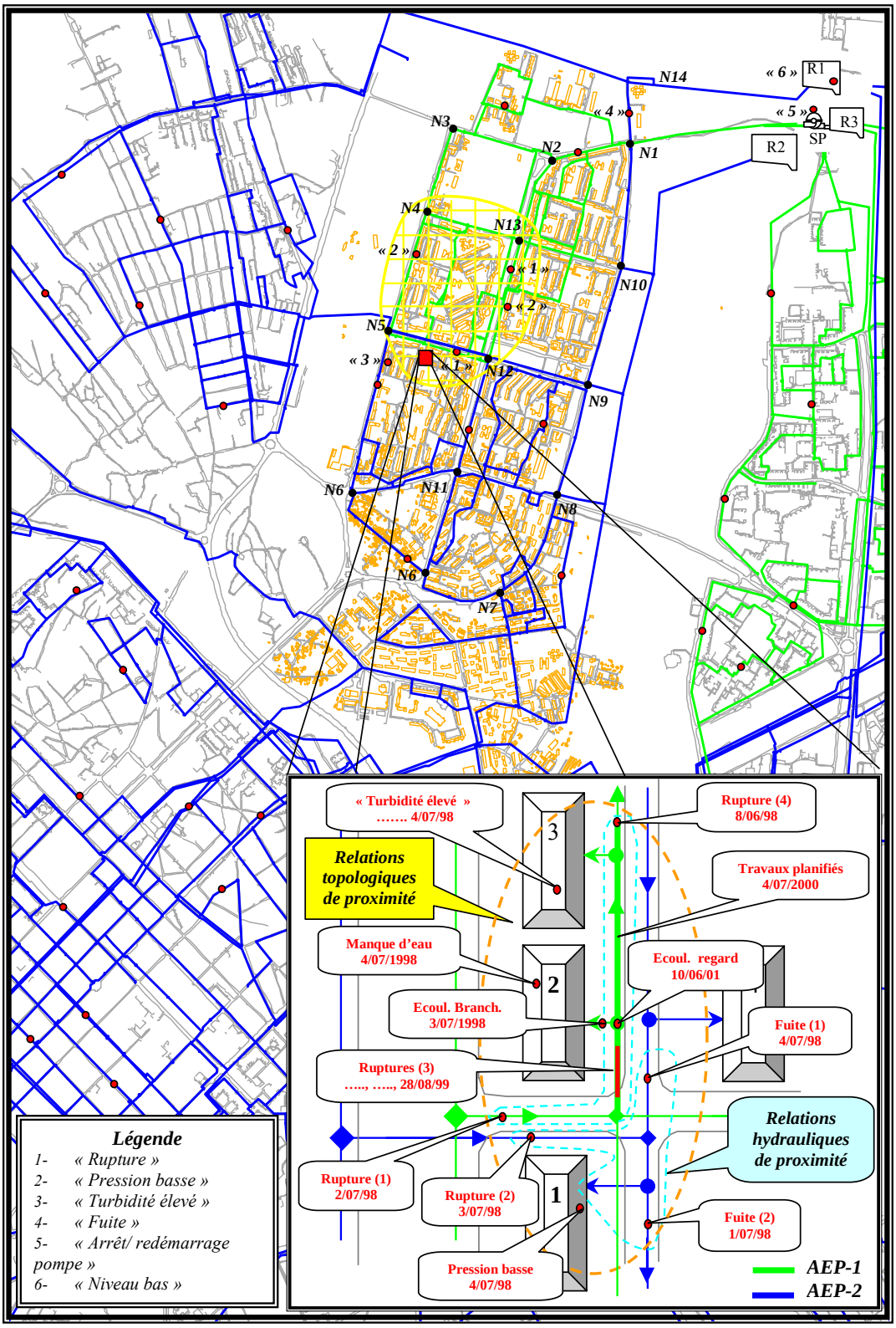


Figure 2.1.4 : Localisation de la plainte du 04/07/1998.

L'étape temporelle identifie la liste des évènements ayant été simultanés ou consécutifs (date de décalage due à la propagation de l'indicent) : les requêtes temporelles sont effectuées sur une période encadrant le jour j, soit j-1 et j+1.
Différents types d'événements sont alors recensés sur le réseau (voir figure 2.1.4) et regroupés dans une liste $\Gamma j_{k(temps)}$. Deux événements sont identifiés : un arrêt et un redémarrage accidentel d'un agrégat de pompes au niveau de la station du pompage, et des fluctuations anormales de niveau d'eau dans les réservoirs.
Dans le cas de la plainte du 04/07/98, les étapes n°1 et n°2 permettent d'identifier différents types d'événements sur le réseau et notamment aux alentours du lieu de la plainte du manque d'eau (événements en grisé sur la figure 2.1.5). Le nombre des événements identifiés, dépendra de l'intervalle de temps pris en compte. Cet intervalle varie de quelques minutes, si la mise à jour des informations se fait en temps réel, à plusieurs jours en fonction du type d'accident, des conditions hydrauliques (pression, vitesse, débit, diamètre, âge de la conduite, …), de sa localisation géographique, et de la gravité de l'impact.

L'avis des experts locaux a permis de fixer de manière empirique, les deux paramètres de la méthode d'analyse spatio-temporelle :

- la durée de la requête temporelle a été fixée à 2 jours,
- la distance pour la requête spatiale a été imposée à 100 mètres.

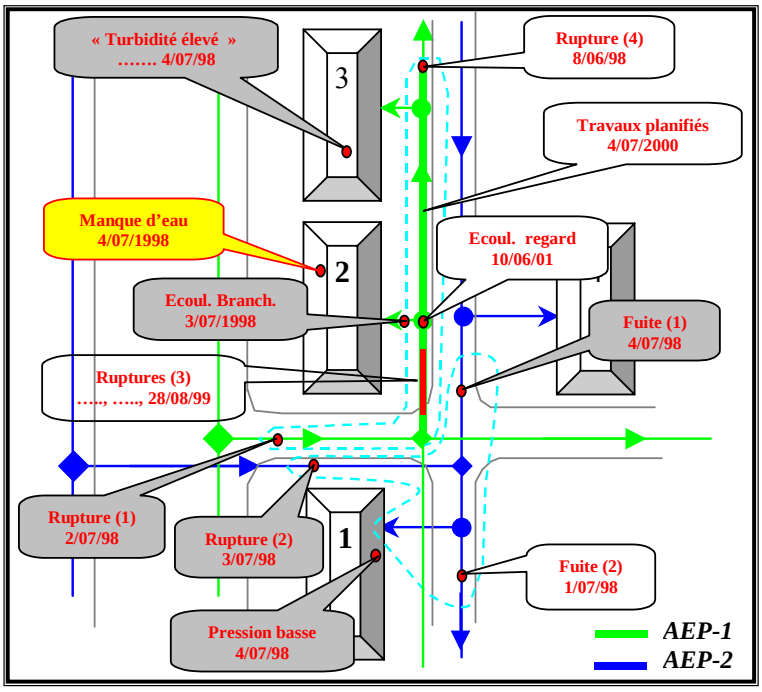


Figure 2.1.5 : Zoom du résultat de la requête temporelle

L'étape suivante n°3 identifie les relations topologiques et hydrauliques existant entre 2 points d'observations d'événements, détectés lors de l'étape 2.

La requête temporelle du 04/07/98 est alors complétée par une requête spatiale en utilisant une zone 'tampon' de 100 mètres, et isole les événements en grisé sur la figure 2.1.6.

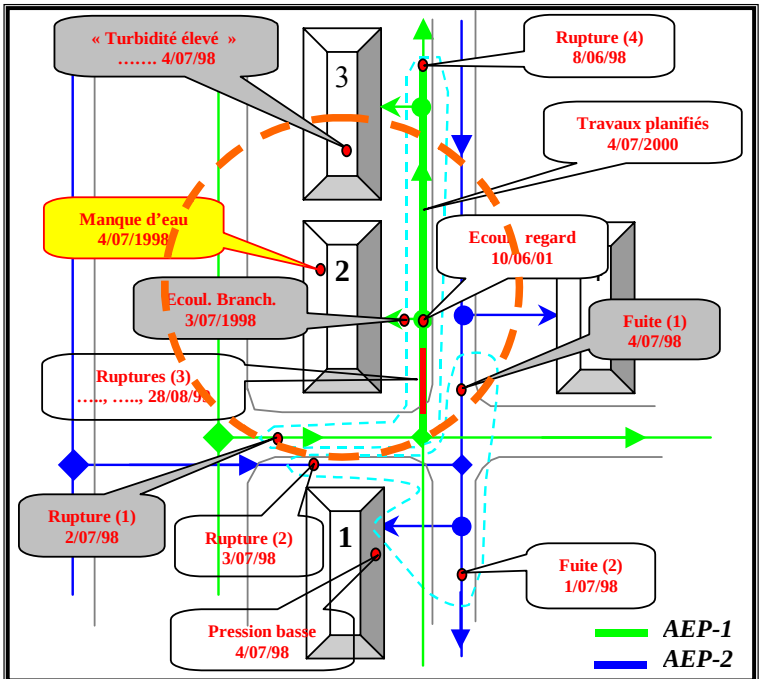


Figure 2.1.6 : Etape d'analyse spatiale des évènements corrélés chronologiquement

La première étape d'analyse spatiale, de type proximité (un ellipsoïde de proximité centré sur le site où survient l'évènement, en figure 2.1.6) repère plusieurs événements synchronisés à l'événement. Parmi ces évènements on note la rupture n°2 au niveau du branchement se produisant la veille de la plainte le (3/07/98), la rupture n°1 du (2/07/98) et la rupture n°2 du (3/07/98), et une plainte concernant une augmentation de la turbidité. La localisation de ces évènements, montre qu'ils apparaissent sur deux branches différentes de réseau. La rupture n°2, les fuites n°1 et n°2, ainsi que la plainte concernant l'insuffisance de pression, dépendent hydrauliquement de la branche d'adduction AEP-2, alimentée par des réservoirs, tandis que la rupture n°1 ainsi que l'écoulement du branchement, dépendent de la branche d'alimentation AEP-1, alimentée par la station du pompage. Les deux branches du réseau (AEP-2 et AEP-1) ne sont pas interconnectées. Leur fonctionnement dépend du régime de la station de pompage, pour la branche AEP-1, et de la fluctuation de niveau d'eau dans les réservoirs, pour la branche AEP-2. Donc, il ne peut pas avoir d'interdépendance hydraulique entre ces manifestations.

Il est alors nécessaire d'ajouter un filtre qui analyse la relation hydraulique entre 2 évènements. Ce qui permet, d'éliminer la fuite n°2 du 03/07/98 liée à une autre branche du réseau (réseau AEP-2) mais qui est corrélée chronologiquement et survient dans un rayon inférieur à 100 mètres.

Cette 1ère analyse spatiale de proximité faite, l'étape suivante discrimine les événements hydrauliquement liés sur l'ensemble du réseau : il s'agit de parcourir le graphe en identifiant les évènements en amont ou en aval du point concerné, connectés sur un même réseau (le graphe est connexe) et dépendant de la même source d'approvisionnement (soit la station de pompage ou le(s) réservoir(s)).

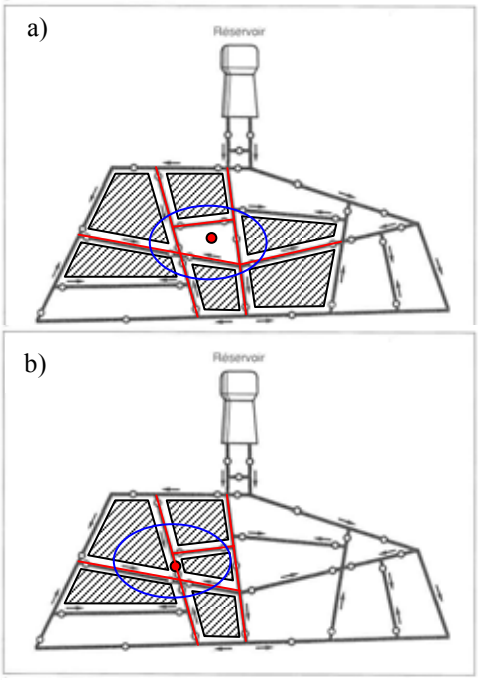


Figure 2.1.7 : Relations topologiques

On remarque alors que cette analyse de proximité aboutit à une liste d'événements survenus sur la partie du réseau alimentée par une même source. Le nombre des événements dépendra de l'étendue du réseau (de quelques événements à quelques dizaines d'événements).

Dans un réseau maillé, l'eau arrive à un même endroit par plusieurs chemins possibles, cependant à un moment donné, on ne peut connaître avec exactitude le sens d'écoulement. La discrimination utilise les relations topologiques pouvant exister entre les différents sites d'observation en récupérant le sous-graphe connexe des nœuds hydrauliques. Lorsque l'événement a lieu à l'intérieur d'une maille, il s'agit de récupérer tous les événements survenus dans les mailles (figure 2.1.7, a). Dans le cas où un événement a lieu sur un tronçon, les mailles voisines du tronçon sont analysées (figure 2.1.7 b).

4. Apport d'un post-traitement

Il s'avère cependant difficile de relier tous ces événements entre eux. La connaissance de certaines variables, comme la pression, la vitesse de l'eau dans les conduites, l'intervalle de temps qui sépare les 2 événements, pourrait permettre de trouver une solution optimale. Or, toutes ces variables et causes n'étant pas connues il n'existe pas de solution optimale.

Afin de fournir une information exploitable pour les gestionnaires du réseau un post traitement a été proposé utilisant des indicateurs, qui permettent de quantifier l'influence des différents événements. Un classement des événements est proposé selon leur importance :

- liée à l'influence ; celle-ci s'exprime par *la distance* entre l'événement *j,* dont la conséquence aurait pu entraîner l'événement *i* . Plus cette distance est faible, plus les effets de l'événement *j* seront importants. Ainsi les événements survenus à l'intérieure de la zone jaune (figure 2.1.8) auront des effets hydrauliques plus importants sur l'événement *i* (manque d'eau) que ceux survenus en dehors de cette zone ;

- liée à la localisation de l'événement *j* :

  - à des emplacements *« stratégiques »* : les stations de pompage, les réservoirs, les puits et captages, la station du traitement d'eau potable, les conduites principales : les anomalies observées au niveau de fonctionnement de la station du pompage (arrêt/redémarrage accidentels des pompes) entraînent un coup de bélier, et une augmentation de taux des ruptures et fuites.
  - en des points en amont ou en aval de l'événement *i* bien que la direction de l'eau dans un réseau change selon la consommation en eau.

- liée au diamètre et à l'importance de la conduite sur laquelle est survenu l'événement *j* :

  - o en fonction de l'importance de la conduite :
    - + les conduites d'alimentation du point *i*
      les artères,
      les conduites de service,
    - - les branchements.
  - o en fonction du diamètre :
    - + D > 500 mm
      D : 250 – 500 mm
      D : 100 – 250 mm
    - - D < 100 mm

- liée à l'intervalle du temps entre les événements *j* et *i*. L'importance de l'événement *j* sera inversement proportionnelle à cet intervalle.

Ces indicateurs, permettent de classer les événements par ordre de priorité. Sur l'exemple précédent, l'ordre de priorité des événements survenus à l'intérieure de la zone jaune ainsi que ceux survenus dans des endroits stratégiques (station de pompage, réservoirs…) est donné dans le tableau 2.1.1:

- des événements ayant des relations temporelles (même laps du temps (instant, heure, jour)), topologiques et hydrauliques voisines de l'événement *i,*
- des événement identifiés comme étant à l'origine de l'événement *i* ayant des relations topologiques et hydrauliques décalées dans le temps (quelques jours auparavant).

Liste des accidents ***tous ordres confondus***	***L'ordre de priorité***
1. l'écoulement du branchement survenu le 3/07/1998 ;	***« 5 »***
2. la plainte concernant l'augmentation de la turbidité de l'eau signalée dans la maison voisine ;	***« 1 »***
3. la rupture n°1 survenue sur une conduite alimentant la maison de l'intérieur du quartier survenue sur la maison voisine le 2/07/98.	***« 7 »***
4. la rupture survenue sur le tronçon N5-N12 ;	***« 2 »***
5. la diminution de la pression observée sur le tronçon N4-N5 le 3/07/1998;	***« 6 »***
6. la rupture sur le tronçon N12-N13 ;	***« 3 »***
7. la diminution de la pression signalée dans les maisons dépendant du tronçon N12-N13.	***« 8 »***
8. des anomalies au niveau du fonctionnement de la station de pompage (« n°5 ») ;	***« 4 »***

Tableau 2.1.1 : Evènements classés selon l'ordre de priorité

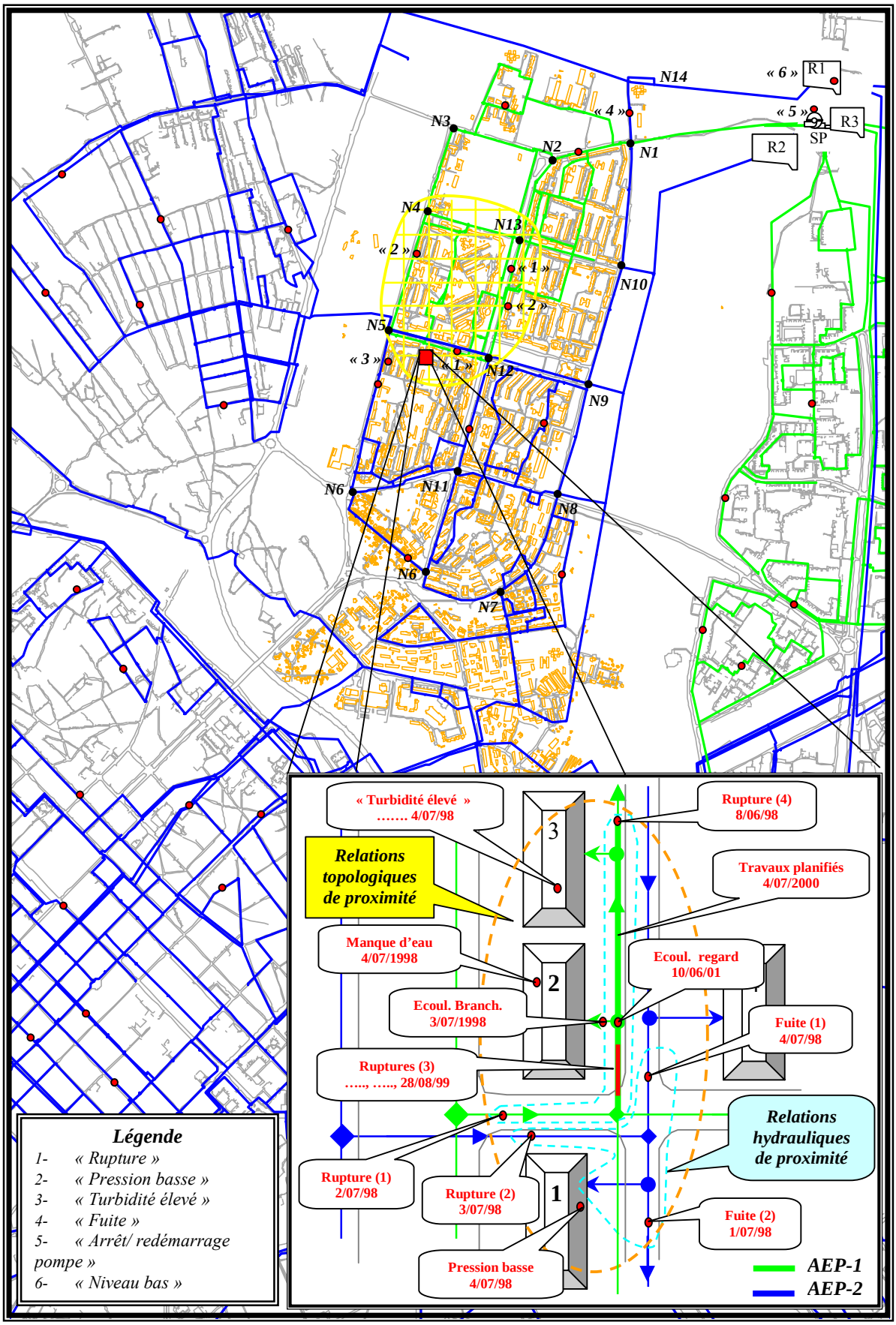


Figure 2.1.8 : Etude de la plainte concernant le manque d'eau

Une extension de la méthode est faite par le passage d'une analyse espace-temps entre ***2*** ou ***n*** événements à une analyse intégrant la causalité entre 2 événements en utilisant un réseau de Pétri. Ce type de graphe a permis de représenter les relations d'accessibilité entre les différents états du réseau d'AEP, de définir les chemins entre ces différents états et les conditions de transition entre ces états.

La complexité des enchaînements entre causes et effets à l'origine des dysfonctionnements impose le regroupement des événements par familles caractérisant le même état du réseau, ce qui simplifie la représentation graphique de changements d'états. Lorsque des phénomènes et des processus susceptibles de provoquer des désordres sur le réseau interviennent, l'état de fonctionnement du réseau est modifié. L'ensemble des variables d'états ***(****Objet_o, Attribut_a, Valeur_x)* (Blindu 2004) permet de caractériser l'état de fonctionnement du réseau au moment de l'incident (les valeurs de la pression, le débit, la turbidité…). L'introduction des variables temporelles aux transitions, aux états et aux arcs, permet de prendre en compte des conditions temporelles pouvant exister entre les états et/ou les transitions.

Le parcours du graphe d'état orienté permet alors de faire :

- du *diagnostic* (si on parcourt le graphe d'aval vers l'amont) : soit, trouver les *causes* qui modifient l'état du réseau,
- de la *prédiction* (si on parcourt le graphe d'amont en aval) : soit, connaître les *conséquences* que pourraient avoir différents facteurs de désordre sur le fonctionnement du réseau.

Cette exploration via un réseau de Pétri, transcrit les relations d'accessibilité entre différents états du réseau d'AEP et pourrait ainsi expliquer les causes d'apparition des dysfonctionnements "instanciés" sous forme d'événements (rupture, pollution bactériologique, turbidité élevée) et de prévoir les conséquences de ces événements sur l'ensemble du réseau dont le graphe associé est non planaire et devant rester connexe. L'implémentation efficace de ce type d'exploration constitue un véritable enjeu algorithmique, notamment il pose le problème d'une exploration sur un graphe (non forcement exhaustif) pouvant être exponentielle si la recherche est déterministe. L'utilisation d'heuristiques peut être une solution, reste alors à déterminer quel type d'heuristique est le plus adapté.

5. Résultats

• La méthode d'analyse spatio-temporelle a été appliquée uniquement sur les données concernant la maintenance journalière du réseau d'AEP. Malgré l'introduction de simplifications par rapport à l'algorithme initial, les résultats d'application des requêtes spatio-temporelles ont permis d'aboutir aux conclusions suivantes :

- le laps de temps choisi (2 à 3 jours) s'est avéré être trop court, compte tenu du nombre des événements regroupés et du temps de réponse du système à un événement,
- les résultats obtenus montrent l'importance des facteurs de désordre liés au fonctionnement du réseau comme les interventions, les réparations perturbant le bon fonctionnement et entraînant à la longue « une fatigue mécanique » du réseau et l'apparition des défaillances. Il s'agit des nombreux coups de béliers qui apparaissent lors des travaux de réparations (source exploitant du réseau) et qui ne sont pas forcement répertoriés dans la base de données en tant qu'incident.

Pour ajuster le choix de l'intervalle de temps (ici 2-3 jours), l'application de lois statistiques peut améliorer le calcul de la durée de synchronisation entre les évènements i et j, variable d'entrée des requêtes spatio-temporelles, même si d'autres paramètres interviennent comme : le type d'événement (ruptures, fuites, turbidité élevée…) et les paramètres de la conduite.

Bien que tout à fait prometteur ce prototype reste très sensible au fait qu'il n'existe pas une base d'interventions suffisante et souligne deux choses :

- la nécessité de disposer d'une base de donnée opérationnelle sur l'ensemble du réseau pour améliorer les résultats ;
- une restructuration de la base de données sur les interventions est nécessaire afin de mieux caractériser les différents types d'événements. La base de données doit prévoir : l'introduction de l'emplacement exact de l'intervention ou incident, le signalement de l'importance de la conduite, la présence ou non d'une protection interne et externe sur les conduites, les conditions climatiques particulières (basse température entraînant le gèle de la conduite), le temps exact de l'apparition de l'incident ou de l'intervention (~ 5 minute).

Ces derniers points ont permis d'établir une liste de recommandations et de procédures au niveau des services techniques du réseau d'AEP de la ville.

• Un outil d'aide à la décision (Blindu 2004) a été développé en parallèle afin de fournir au gestionnaire de la RAC un moyen quantitatif de mise en place des programmes de renouvellement du réseau d'eau potable, à l'échelle de la conduite, sur différents horizons temporels. Compte tenu des difficultés présentées précédemment et des contraintes d'exploitation d'un réseau d'AEP, il s'est avéré que les modèles statistiques et économiques ne sont pas adaptés et nécessitent une base de données exhaustive et couvrant une longue période (plus de 10 ans) sur la maintenance du réseau. En considérant l'expérience de la Stéphanoise des Eaux, concessionnaire des services publics de l'eau potable et de l'assainissement de la ville de Saint Etienne, dans la gestion du réseau, et des données disponibles au niveau de la RAC, une méthode d'analyse multicritère a été adoptée.
La Méthode Hiérarchique Multicritère, MHM développée par (Saaty 1984) retenue s'est avérée être intéressante afin de quantifier et hiérarchiser des critères et des sous-critères caractérisant la prise de décision pour le renouvellement du réseau d'eau potable et tout particulièrement pour obtenir une pondération qui intègre l'expertise des techniciens du réseau de la régie « Apa Canal Chisinau » (RAC).

L'objet de la décision est un *niveau d'urgence*, paramètre permettant de classer les conduites dans les programmes de réhabilitation à court, à moyen et à long terme, tout en considérant les différents critères caractérisant le réseau d'AEP.

L'avis des experts de la RAC concernant les tronçons devant être inclus dans des programmes de réhabilitation, correspondant aux périodes d'intervention immédiates, à court terme (entre 2 et 5 ans), à moyen terme (entre 5 et 10 ans) et à long terme (plus de 10 ans) nous a permis de valider la méthode. La validation se base sur la comparaison du classement, issu de l'expertise du gestionnaire, aux valeurs de niveau d'urgence calculées à l'aide de la MHM. Il s'agissait, autrement dit, de vérifier si les arcs classés par le gestionnaire dans un niveau *n*, possédaient des niveaux d'urgence calculés situés dans un même intervalle de valeur. Les seuils (sur les niveaux d'urgence calculés) de chaque intervalle ont été établi par des courbes cumulées des niveaux d'urgence. Ces courbes fournissent la répartition des niveaux d'urgence calculés pour chaque niveau « d'expertise » donné par la RAC.

Les seuils appliqués pour chaque niveau d'urgence, nous ont permis d'obtenir un taux de coïncidence avec l'avis de l'expert : pour le *premier* niveau de 73%, pour le *deuxième* 74% et pour le *troisième* de 79%.
Après avoir appliqué ces seuils sur l'ensemble du réseau pour l'année 2001, il apparaît que 20.3 % des conduites se trouvent au niveau d'urgence le plus critique (niveau d'urgence I - intervention Immédiat). Parmi ces tronçons, environ 66 % représentent des diamètres supérieurs à 300 mm. Ces conduites représentent, ni plus ni moins, les artères principales du quartier. Ceci coïncide avec les préoccupations des gestionnaires à savoir que ce sont les artères principales qui sont prioritaires dans un programme de renouvellement.

Une première estimation, assez *rustique* en fonction des résultats obtenus et du prix de mètre linéaire des conduites donne un ordre de grandeur des investissements nécessaires pour changer toutes les conduites se trouvant dans le niveau I d'urgence. Ces résultats montrent la nécessite de tels outils dans un contexte de renouvellement des infrastructures, engagé depuis 5 ans sur la ville, ou des priorités d'actions s'imposent.

6. A retenir

L'approche métier de la gestion et du vieillissement d'un réseau susceptible de fortement mal fonctionner est un enjeu majeur pour certains pays européens et nécessite la mise à disposition un certain nombre d'outils associés à des bases de données actualisées.
On retiendra plusieurs points "recherche" :

- la limite de la modélisation hydraulique lorsque les données du réseau (état des conduites et données sur les variables 'écoulement) ne sont pas disponibles, impose de se doter d'autres approches complémentaires de type «*explicatif*" ou «*boite noire*" ou modèle de type *déductif ou comparatif à base de connaissance,*.
- l'exhaustivité des causes et états possibles d'un système n'est pas forcément atteinte notamment lorsque les données disponibles sont peu représentées ou approximatives
- la notion forte de similarité ou dissimilarité entre 2 objets ici 2 évènements dynamiques : il s'agit alors d'identifier cette relation mixant des paramètres hydrauliques, le temps, et la topologie, et de la quantifier.
- le choix entre développer un graphe de causalité ou au contraire exploiter la similarité ou du moins la corrélation pouvant exister entre deux instances.

Travaux produits :

- DEA I. Blindu - Aide au diagnostic du réseau d'AEP pour la ville de Chisinau par analyse spatiale et temporelle des dysfonctionnements, DEA "Sciences et Techniques du déchet ", Ecole Nationale Supérieure des Mines de Saint-Étienne 1999
- Doctorat I. Blîndu - Outil d'aide au diagnostic du réseau d'eau potable pour la ville de Chisinau par analyse spatiale et temporelle des dysfonctionnements hydrauliques, Ecole nationale supérieure des Mines de Saint-Étienne - Université Jean Monnet, 2004
- 1 conférence
- 2 rapports techniques liés à deux phases de financement du projet - Région Rhône-Alpes - Conseil Général - ARMINES, 1998-2003

Bibliographie

Bremond B., « Mesurer le vieillissement d'un réseau d'eau potable », Courants, mars/avril 1994, vol 26, pp.21-28

Brunet J., Jaume D., Labarrère M., Rault A., Vergé M., « Détection et diagnostic de pannes », Approche par modélisation, Traité des Nouvelles Technologies, Hermès, Paris, 1990, 236 p.

Charon I., Germa A. , Hundry O., « Méthode d'optimisation combinatoire », Paris, 1996

Dupont A., « Hydraulique urbaine », Tome 2, Editions Eyrolles, Paris, 1979

Egenhofer M. and Golledge R., Spatial and Temporal reasoning in Geographic Information Systems", New York, Oxford, 1998, 276 p.

M. Ghallab and A.M. Alaoui. « Managing efficiently temporal relations through indexed spanning trees ». In *11th International Joint Conference on Artificial Intelligence*, pages 1297-1303, Detroit, aug 1989.

M. Ghallab and A.M. Alaoui. « Relations temporelles symboliques : représentations et algorithmes ». *Revue d'Intelligence Artificielle*, 3, 1987.

Laurini R et al., 1993 « Les bases de données en géomatique », Paris :Edition Hermes, 1993, 340 p.

Mourot G., Harkat M.F., Ragot J. (CRAN-ENSG, Nancy), « Détection de défauts de capteurs d'un réseau de surveillance de la qualité de l'air », 2ième Colloque A&E « Automatique et Environnement », ENSM.SE, Centre SITE (Science, Information et Technologies pour l'Environnement), 2001

Ragot J., Darouach M., Maquin D., Bloch G., « Validation de données et diagnostic », Hermès, Paris, 1990, 431 p.

Saaty T.L Décider face à la complexité. Entreprise Moderne d'Edition,Paris, 1984, 231 p.

Chapitre 2 Analyse temps - espace de situations météorologiques : approche par scenarii

Rappelons quelques questions soulevées en introduction:

- la Loi sur l'Air et l'Utilisation Rationnelle de l'Energie (LAURE - 1996), fixe la mission des associations de qualité de l'air, qui consiste à mesurer, analyser et informer le public sur la pollution urbaine. Les réseaux de surveillance fournissent des données de concentration de polluants validées en chaque point du réseau. Le suivi en temps réel des concentrations mesurées permet de détecter les seuils réglementaires (*Directive 2002/3/CE, 2002***)** et les niveaux d'alerte afin de déclencher les procédures et stratégies de réduction prévues par la loi et informer la population. Cette gestion journalière impose qu'il soit indispensable d'avoir une bonne estimation de la répartition de la qualité de l'air instantanée sur la zone urbaine et/ou péri-urbaine concernée, et de pouvoir anticiper la qualité de l'air au lendemain.
- autour des centres de stockage de déchets (CSD) de type Ordures Ménagères (OM) et Déchets Industriels Banals (DIB) il peut apparaître de nombreuses plaintes liées à la présence persistante ou non de nuisances olfactives en fonction d'un certain nombre de paramètres comme la météorologie, les conditions d'exploitation. Il s'agit de prévoir les conditions météorologiques et d'exploitation favorables à l'apparition de nuisances olfactives afin de fournir des recommandations durant la phase d'exploitation du site et l'identification de seuils de déclenchement de systèmes de réduction.

Il s'agit i) d'obtenir la répartition instantanée de la qualité de l'air (soit d'une (ou de) valeur(s) de concentration(s) d'ozone ou autres (NOx..), de molécules olfactives), ii) d'anticiper une valeur de donnée (scalaire ou vectorielle) pour le lendemain. Tout processus physique est perçu par des observations, structuré dans un schéma conceptuel et associé à un ou des modèles physiques (figure 2.0.1). En qualité de l'air (polluant passif ou molécule odorante) la modélisation est *coûteuse*: en données, en temps de calcul, pour un faible gain résultats/investissement.
La prémisse "IL existe une modélisation déterministe ET IL existe dans les données une information qui caractérise l'Etat du système " est alors établie comme étant vraie, ALORS on propose d'accroître la connaissance implicite[10] de certains états du système par la simulation.

1. Approche par scénario de la qualité de l'air

1.1 Contexte

La modélisation déterministe des phénomènes de transport et de transformation des polluants atmosphériques reste très partiellement opérationnelle chez les gestionnaires de réseaux (modèle Chimère, PREV'AIR) et à l'usage de spécialistes (http://www.prevair.org/). On s'intéresse à leur usage en mode pronostic/diagnostic pour des évènements choisis pour lesquels ils restituent la variation des polluants sur une période donnée. L'inertie du phénomène de pollution de l'air fait qu'il est indispensable d'une part, de pouvoir anticiper la qualité de l'air au lendemain, mais également d'avoir une bonne estimation de la répartition

[10] au sens de connaissance implicite par opposition à la connaissance explicite pour celle observable ou connue (mesurée ou contrôlée par des lois connues)

de la qualité de l'air sur la zone urbaine et péri-urbaine concernée. Pour répondre à cette attente, il existe deux types d'approche : les outils de prévision dits à court terme qui permettent de prévoir par exemple un niveau d'ozone le lendemain, et les outils de diagnostic qui permettent de reconstruire la variation sur une période donnée (de 1 à 5 jours raisonnablement) de la qualité de l'air sur l'ensemble du territoire.

- Les outils de prévision font appel à des techniques mathématiques et statistiques telles que les réseaux de neurones, l'analyse discriminante ou les modèles de régression (Thomson et al 2001, US EPA, 1999). Ce type de modèle utilise les données fournies en une station du réseau et prévoit la concentration au lendemain d'un polluant (tel que l'ozone). Ils présentent l'avantage souvent d'être peu exigeants en temps de calcul, mais il reste le problème de la sous-évaluation des pics de pollution et plus globalement le dilemme entre précision et robustesse. On notera que ce type de module fonctionne pour un ensemble de variables explicatives associées à une localisation géographique.

- Le second type de modèle s'appuie sur une modélisation déterministe des phénomènes de transport et de transformation des polluants atmosphériques (Figure 2.2.1). Bien que devenus partiellement opérationnels dans les réseaux (modèle Chimère AIRPARIF, ASPA…), la complexité de ces modèles réside dans la nécessité de disposer de nombreuses données en entrée, de calibrer le modèle, et de valider les résultats. Ces modèles ne peuvent pas être facilement utilisés en mode prévision (à l'échelle micro et méso) mais seulement en mode de pronostic/diagnostic. Ils ont l'avantage de restituer spatialement la variation des polluants sur une période donnée.

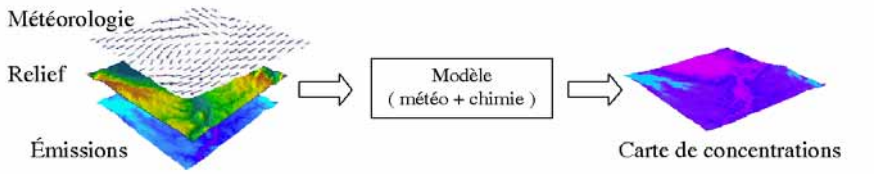


Figure 2.2.1. Modélisation déterministe de la pollution atmosphérique et données d'entrée de type raster (S.I.G)

Afin d'évaluer la qualité de l'air sur la ville (en ozone) de Saint-Étienne, une approche mixte d'identification de scenarii météorologiques caractéristiques propices à un état de pollution atmosphérique a été proposée (Batton-Hubert, 2000) : il s'agit d'identifier et de construire un certain nombre de scenarii types caractéristiques, puis pour chacun de ces scenarii, calculer la répartition de la concentration de la pollution atmosphérique en utilisant un code de mécanique des fluides. Enfin lorsqu'une situation météorologique et/ou de qualité de l'air est observée, il s'agit de pouvoir trouver son (ses) analogue(s) dans la base ainsi constituée (figure 2.2.2) : cette ultime étape qui relève de la prédiction fera l'objet des perspectives présentées au volet 3.

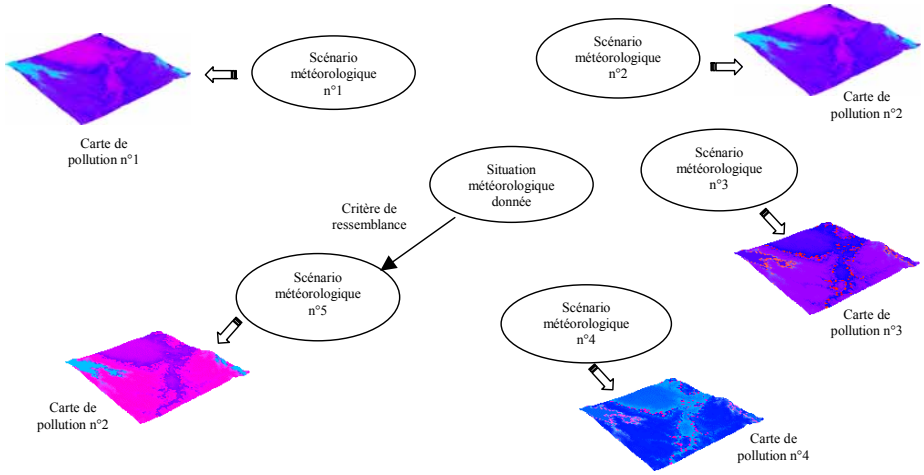


Figure 2.2.2 : Base de scénarii et reconnaissance d'une situation donnée

1.2 Méthodologie développée

L'idée majeure de cette approche mixte, développée au cours de ces travaux, se décline de la façon suivante :

- en l'absence de données au sol suffisantes (au sens d'un semis de points relativement dense et homogène) l'obtention aussi bien des isoconcentrations d'un polluant, des isothermes d'un champ météorologique et encore moins d'un champ vectoriel (vent) ne peut être établie par les méthodes d'interpolation linéaires ou non et/ou d'estimateurs probabilistes ; la modélisation déterministe est envisageable avec son cortège de difficultés, soit l'ensemble suivant : [mise en place et la résolution des équations de la dynamique de l'atmosphère + la disponibilité des données d'entrée (conditions initiales et aux limites) + validation des résultats]. Or, cette modélisation restitue assez fidèlement les mécanismes de transfert et de circulation des polluants mais est trop coûteuse. Nous proposons de la mettre en place ponctuellement sur des évènements majeurs, soient de par leurs concentrations élevées en polluant ou au contraire soient des situations instables pouvant être des états de transition intéressants. Ces évènements constituent des scenarii météorologiques caractéristiques qui doivent être identifiés avant d'être calculés par un CFD[11]. Le polluant concerné est l'ozone.
- L'identification des scenarii météorologiques caractéristiques se base sur l'hypothèse suivante : il existerait une relation univoque entre une situation météorologique sur un jour (*a posteriori* sur 3-5 jours) et les données disponibles, mesures de l'état de l'atmosphère et notamment de la concentration en ozone.
- Par conséquent, il serait possible de caractériser un jour ou une série de jours consécutifs par un ensemble de variables pertinentes du système "*scénario météorologique*" qui, dans un premier temps, sont des données discrètes. Dans le volet 3. qui suit, il sera montré comment cette notion pourra évoluer dans les travaux de recherches futures.

[11] CFD : Computational Fluid Dynamics : code de calcul numérique pour la résolution des équations de l'atmosphère par discrétisation en volumes finis et/ou éléments finis

Il s'agit alors :

- d'identifier les variables pertinentes d'un jour métrologique/état de la qualité de l'atmosphère (pour l'ozone),
- d'établir si ces variables pertinentes permettent d'établir des classes de jours. Une classe peut alors s'interpréter selon (Le Hégarat-Mascle 2005) c*omme un domaine de validité ou d'application d'un modèle physique dont on a l'incapacité immédiate à avoir un modèle universel ;* ce qui est cohérent avec la notion d'ordre et de rangements de la classification mais aussi de la théorie des ensembles,
- de valider la signification physique de ces classes de jours, afin d'identifier un comportement moyen par classe et son intervalle de fonctionnement afin d'obtenir un "*scénario météorologique*",
- la phase suivante consiste alors à décliner la modélisation déterministe (figure 2.2.2) sur un jour, barycentre de la classe, jour le plus significatif de la classe (chapitre 2 - §2.).

1.2.1 Choix d'une classification

Ciblée sur la recherche d'une trace de l'échelle locale du comportement de la ville à partir de données mesurées localement (vent, température, pression, humidité, NO_x, O3…localisées sur une station d'un réseau de surveillance), cette démarche a consisté à adapter plusieurs méthodes de classification afin d'identifier les variables significatives et nécessaires à une classification. En effet, les conditions de la qualité de l'air sont liées au processus physique non linéaire de sa formation qui par conséquent, doit avoir des répercussions sur les variables explicatives utilisées. L'identification de situations météorologiques types a été développée pour certains phénomènes comme les pluies acides, la météorologie et l'hydrologie (Obled et al. 2002) mais non développée pour la pollution atmosphérique et tout particulièrement pour l'ozone.

Différentes méthodes ont été testées, seuls les résultats obtenus par la classification hiérarchique mixte (Lebart et al. 1997) sont exploités dans ce chapitre. Cette méthode mixte regroupe deux techniques de classification : la méthode des centres mobiles et la méthode hiérarchique ascendante afin d'améliorer le résultat final (annexe 2.1).

1.2.2 Choix de données et variables pertinentes

L'idée est d'intégrer des paramètres locaux (via des données de station de mesure au sol) et des facteurs contribuant à la formation de l'ozone comme le profil vertical des températures et le vent en altitude.

Mesures au sol

Les mesures au sol sont fournies par le réseau de qualité de l'air, Ampasel. Elles concernent la température, l'humidité, la direction et la vitesse du vent, la concentration en ozone. Les nombreux travaux sur les paramètres pertinents de la prévision atmosphérique (Kuebler, 2001; Ludwig et al, 1995) ont permis de sélectionner les variables suivantes:

- la température maximale journalière,

- l'amplitude thermique = δ(température maximale, température minimale) /journée,
- la vitesse moyenne du vent le matin entre 6 et 10 heures,
- l'humidité relative moyenne journalière,
- les concentrations en ozone : les journées retenues sont celles dont la moyenne en ozone sur 8 heures dépasse le seuil de 120 µg/m^3 fixé par une directive européenne (*Directive 2002/3/CE, 2002*) pour protéger la santé humaine. Cette moyenne est le maximum journalier de la moyenne sur 8 heures calculée à partir des 24 moyennes horaires glissantes sur 8 heures. La première moyenne est calculée de 17 heure la veille à une 01 heure le jour même et la dernière période de calcul est comprise entre 16 h et minuit. L'échantillon retenu sur les 4 années initiales de 1997-2001 est de ~230 jours.

Données en altitude

Les mesures en altitude, utilisées sont les sorties du modèle américain FNL, disponibles sur le site http://www.arl.noaa.gov/ready/amet.html. Les variables choisies pour intégrer le comportement de l'atmosphère en altitude sont :

- la vitesse et la direction du vent au niveau 850 mbar, vent synoptique (origine de la masse d'air),
- la température à 850 mbar,
- la hauteur du niveau de 500 mbar.

Chaque jour est représenté par 11 variables: la température maximale journalière (Tmax), l'amplitude thermique (ΔT), les composantes zonale et méridienne du vent matinal (Umat, Vmat), les composantes zonale et méridienne du vent de l'après midi (Usoir, Vsoir), l'humidité relative moyenne journalière (Hum), les composantes zonale et méridienne du vent en altitude (Ualt, Valt), la température en altitude (T850) et l'altitude du niveau 500mbar (H500).

1.2.3 Classes obtenues

L'interprétation des classes obtenues, utilise la classification des régimes anticycloniques utilisée par Blanchet en climatologie (Blanchet, 1990,1994), un guide des situations synoptiques (*Roth, 2001*) et les cartes synoptiques au sol et au niveau 500 mbar fournies par le département *Air Ressource Laboratory* (http://www.arl.noaa.gov/ready/amet.html.).

Les variables qui favorisent la formation des classes, sont identifiées par le calcul de la valeur-test[12]. La valeurs-test est évaluée pour chaque variable de chaque classe, soit :

$$t_k(X)=\frac{\overline{X}_k-\overline{X}}{s_k(X)}$$

$$S_k^2(X)=\frac{n-n_k}{n-1}\frac{s_2(X)}{n_k}$$

Avec : X_k est la variable X dans la classe k,
$S_k^2(X)$ est la variance empirique de la classe k,
n_k est le nombre d'individus dans la classe k et
n le nombre total d'individus.

[12] la Valeur test correspond à l'écart entre la moyenne de la variable à l'intérieur de la classe et la moyenne de la variable sur l'ensemble des individus, en tenant compte de la variance de la variable dans la classe considérée.

	Classes						
	1	2	3	4	5	6	7
Omax(µg/m3)	-	-	-	-	-	-	-
Moy8h(µg/m3)	-	3.45	-	-	-	-	-
Tmin(°C)	3.95	-	-	-	-	2.98	-
Tmax°C)	4.35	-	-	-	-	4.19	3.8
deltaT(°C)	-	- 3.2	-	-	-	-	4.53
Hum	2.81	3.23	-	-	-	-	4.01
T850(°C)	3.92	-	-	-	3.32	4.51	3.06
H500(m)	-	-	-	-	3.28	4.96	-
Umat(m/s)	-	-	-	-	2.48	2.92	5.82
Vmat(m/s)	-	-	3.21	-	3.52	3.21	-
Usoir(m/s)	-	2.39	-	2.76	3.89	3.26	-
Vsoir(m/s)	2.48	2.95	-	-	- 2.9	-	4.02
Ualt(m/s)	-	-	2.48	4.12	2.38	-	-
Valt(m/s)	-	-	-	-	-	2.56	4.2

Tableau 2.2.1 Valeurs test significatives par classe

Les valeurs-test des variables (tableau 2.2.1) dont la valeur absolue est importante, caractérisent une classe. L'interprétation des résultats associée à la valeur-test significative, et aux classes obtenues est la suivante :

1ère classe : cette classe est établie selon l'influence des températures maximales et des températures en altitude plus faibles de l'ordre de 2°C que les valeurs moyennes générales sur l'ensemble de l'échantillon. La composante méridienne du vent du soir est élevée. Du point de vue synoptique, cette classe est constituée de situations anticycloniques centrées sur la France ou de situations de marais barométriques (vaste zone sur laquelle la pression varie peu). Localement, le vent vient de l'ouest (secteurs OSO et ONO) le matin.

2ème classe : Une amplitude thermique faible de ~15°C, et une température minimum élevée de ~16,4°C par rapport aux moyennes générales sont les conditions de formation de cette classe. Le vent est orienté O-SO le matin et s'oriente vers le SO-SSO, l'après midi. Les situations de marais barométriques et d'anticyclones situées sur l'Europe du Nord-Est et la Scandinavie sont les plus fréquentes dans cette classe.

3ème classe : la vitesse méridienne le matin et la vitesse zonale en altitude caractérisent ce groupe de jours. Du point de vue synoptique, aucune situation ne se dégage. Le matin le vent du OSO (pour 40 %). On remarque que les roses des vents de l'après midi et en altitude sont similaires. Le vent est légèrement plus rapide en altitude soit 2.5m/s^{-1}, et de 2.0 m/ s^{-1} au sol.

4ème classe : les valeurs test sont importantes pour les vitesses zonales du matin et en altitude. Une dorsale (axe de hautes pressions prolongeant un anticyclone) provenant de l'ouest se retrouve dans la majorité des cas. D'après Blanchet, c'est un type de temps typique du mois de juillet, ce qui est confirmé par la composition de la classe (4 jours sur 7 proviennent du mois de juillet).

5ème classe : Cette classe coïncide avec un anticyclone situé sur le nord de la France et les Iles Britanniques. Ceci est confirmé par le fait que la hauteur du niveau 500 hPa est élevée (haute pression).

6ème classe : Les températures (Tmax et T850) et l'altitude du niveau 500 hPa sont les principales variables à l'origine de ce groupe. Il correspond à un temps anticyclonique advectif du nord-est centré sur la Scandinavie. (Figure 2.2.3) Durant ce type de temps, l'air est sec, modérément chaud et un vent modéré souffle du nord, ce qui coïncide avec les valeurs moyennes de cette classe pour laquelle la température maximale est de 24.24°C alors que la moyenne générale est de 29.43°C. La rose des vents en altitude confirme la direction nord du vent. Un seul épisode de 5 jours est à l'origine de cette classe

7ème classe : Le type de situation le plus rencontré dans cette classe est le régime de marais barométrique. Les fortes températures et la faible humidité sont les caractéristiques de cette classe.

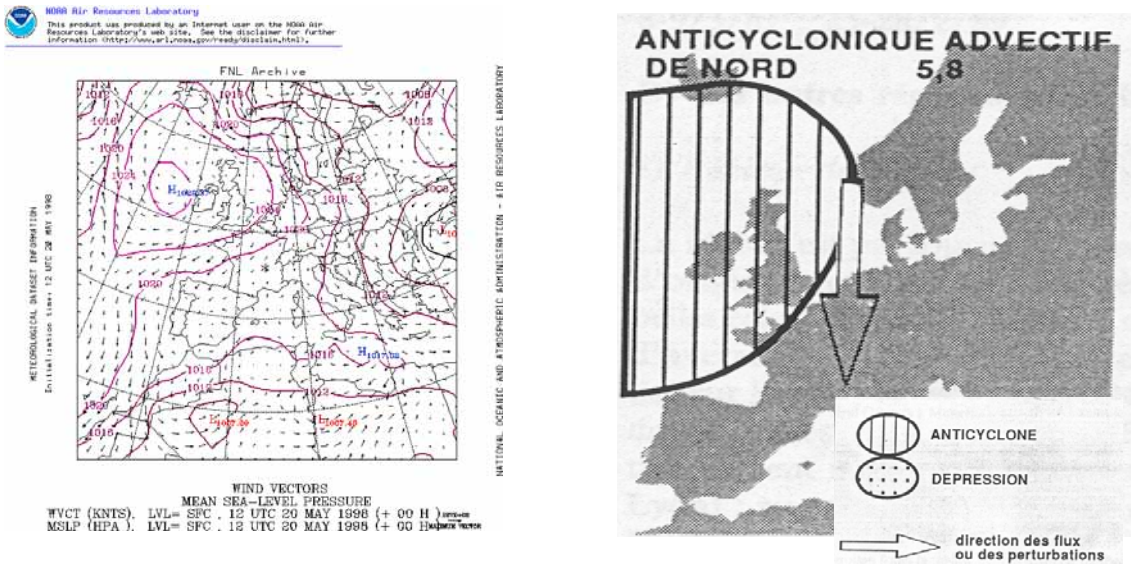


Figure. 2.2.3 : Exemple d'une carte « FNL » et d'une carte synoptique de la classification de Blanchet.

1.2.4 Eléments sur la classification obtenue

On remarque que le nombre de jours par classe est relativement similaire. Il est de l'ordre d'une dizaine, résultat lié à l'effet des différentes distances (Ward, Centroïd, Average Linkage) utilisables en classification (Kalstein et al 1997) *:* la distance du critère de Ward a tendance à créer des groupes de même taille mais dont le contenu est moins homogène qu'avec la distance du lien moyen. Ces différences de dimension de classe sont à rechercher au niveau de la définition des distances utilisées. Différents tests ont permis de choisir la distance de Ward : cette dernière fait une meilleure distinction entre les classes de situations météorologiques alors que la classification avec le lien moyen produirait une énorme classe contenant la moitié des individus de l'échantillon.

		Classes						
		1	2	3	4	5	6	7
Durée de la persistance (nb de jours)	1	10	9	10	5	9	1	5
	2	3	1	0	1	2	0	2
	3	0	1	0	0	0	0	0
	4	0	0	0	0	0	1	0

Tableau. 2.2.2 : Persistance à l'intérieur des classes

Afin d'évaluer la qualité de la classification, on évalue si plusieurs jours consécutifs (ayant les mêmes caractéristique) peuvent être classés de la même façon. Dans le tableau 2.2.2 on constate que des jours successifs sont classés dans la même classe. Ceci signifie que la classification arrive à détecter une similitude certaine; ces résultats convergent avec ceux de (Fernandez et al 1992) qui identifient ainsi des séquences de deux et trois jours dans leur classe. Cette durée pourrait correspondre à l'échelle de temps d'un passage de fronts des masses d'air en climat tempéré (durée de changement de temps météorologique des latitudes ~45°N).

Le recours à d'autres paramètres ou variables non utilisés pour effectuer la classification est absolument nécessaire pour la compréhension des classes obtenues. Ceci requiert l'acquisition de cartes synoptiques (figure 2.2.3). Et nécessite un niveau d'expertise relativement important.

Comparaison des classes avec ou sans les données d'altitude.

L'influence des variables en altitude par rapport aux variables au sol a été évaluée par comparaison avec une classification avec les données au sol. Le croisement entre ces 8 classes (au sol) et les 7 classes (échelle régionale) a permis d'identifier de fortes corrélations entre les deux classifications obtenues. (tableau 2.2.3*)* Ceci semble a priori cohérent car seules 4 variables sur les 11 sont liées à l'altitude.

		Classes obtenues (sol + alt)							
		1	2	3	4	5	6	7	total
Classes obtenues avec seulement les données au sol	1						10		10
	2							8	8
	3	3		2					5
	4	2	5	6					13
	5				5				5
	6		1	1		6			8
	7	12	2	1		1			16
	8	1	2	4					9
	total	18	10	14	5	7	12	8	74

Tableau 2.2.3: Tableau croisé des classification réalisées avec et sans les données d'altitude

Trois classes (de 10, 8 et 5 éléments) en gris clair constituent des noyaux forts (ayant les mêmes individus et le même nombre d'individus) qui sont identiques. Deux groupes (en gris intermédiaire) respectivement de 12 et 6 éléments forment eux aussi 2 sous-ensembles caractéristiques dans les 2 classifications.
On observe cependant que deux groupes de jours classés dans deux classes au sol différentes (6 éléments de la classe 4 et 4 de la classe 8, en gris foncé dans le tableau 2.2.3) sont classés dans la classe 3 de la classification à l'échelle régionale. Ceci nous amène à considérer que cette classe est assez inhomogène. La classification à l'échelle régionale ne prendrait pas en compte des distinctions mises en évidence par la classification au sol.
En revanche, l'apport des variables d'altitude permet de distinguer des situations météorologiques différentes pour des conditions au sol similaires. C'est le cas des jours des classes 3 et 4 au sol, classés dans des classes différentes pour l'échelle régionale. Par exemple, 5 jours de la classe 4, et 6 jours de la classe 4, se retrouvent respectivement dans la classe 2 et la classe 3 de l'autre classification.
L'utilisation des variables d'altitude permet d'affiner certaines classes mais peut aussi dans certains cas rendre inhomogènes des classes homogènes au sol.

1.3 Résultats

Une contribution importante de ces travaux, a été l'étude de la pollution sur Saint-Etienne, où peu d'études avaient été menées alors, hormis l'étude sur le Plan de Déplacement Urbain (*CERTU-ADEME, 2002*). Depuis, le réseau de mesure AMPASEL a pu bénéficier des travaux de modélisation effectués par le groupement GERSIA sur la région Rhône-alpes.

Ce travail s'inscrit dans l'élaboration de méthodes de diagnostic basées sur l'élaboration de scenarii de pollution pour des réseaux possédant peu de points de mesures, qui utilisent les concepts et outils développés en Reconnaissance de Formes, de la classification automatique et de l'analyse de données afin de déterminer des types de situations durant lesquelles la qualité de l'air est médiocre.

Les premiers résultats, établis pour une classification de type centres mobiles et hiérarchique ascendante ont montré l'importance de l'identification des variables pertinentes et du dilemme existant entre données /variables disponibles et variables pertinentes. Les variables locales sont des données locales telles que la température, l'humidité, la direction et la vitesse (station Météo France synoptique). L'intégration de données 3D (profils de radiosondage en température ou vent) montre l'instabilité de la classification dans certains cas mais aussi la possibilité d'intégrer des phénomènes synoptiques et locaux (Gaillard et Batton-Hubert 2003).

La simulation dynamique et spatiale de scénario a été initiée par la modélisation de scénario type sur des cas estivaux de pollution à l'ozone (été 2000) ; cette simulation montre la difficulté de valider les résultats de la simulation sur des secteurs peu instrumentés (3 stations Ozone, une station synoptique météorologique).

Ces résultats montrent que :
- la faible disponibilité des données doit être intégrée dans les traitements,
- le *multinesting* en modélisation est nécessaire,
- l'intégration des résultats dans un Système d'Information Géographique doit permettre le pré traitement des données en entrée et en sortie et le calcul des impacts sur la population.

Ces 3 points sont introduits dans les travaux suivants , paragraphe 2.

1.4 A retenir

Ces travaux s'inscrivent dans une problématique actuelle de l'évaluation des impacts sanitaires par les outils déterministes en terme de diagnostic (répartitions d'un champ de concentration en ozone) mais aussi pour le déclenchement des seuils d'alerte pour les modèles prédictifs.

Après avoir évalué si une ou des méthodes de classification pouvait donner un sens aux classes de jours obtenues, il demeure un certain nombre de questions en suspens:

- Les variables sélectionnées sont-elles les mêmes si l'on utilise d'autres méthodes de classification, sachant qu'une ACP préalable avait permis d'utiliser les coordonnées projetées de chaque jour ? en effet, l'ACP repose sur l'existence d'une corrélation linéaire qui existe effectivement d'un point de vue physique entre certaines variables mais la concentration en ozone n'évolue pas linéairement en fonction toutes les variables météorologiques.
- Un jour est représenté par des données moyennes ; que devient la similitude entre 2 jours sur des données non stationnaires (température, vent, ozone) ? puis *a posteriori* entre 2 - 5 jours ?
- Qu'apportent comme amélioration des techniques de classification de type K-moyennes, les réseaux de neurones et peut-on identifier des indices de confiance à chacune des variables explicatives de la classification ?
- Du nombre de classes choisi, dépend le nombre de simulations pour un jour donné ; mais que devient ce nombre si l'on doit introduire la séquence de 2-5 jours pour passer d'un scénario *jour* à un scénario *jours* non stationnaire ?

Des points de recherche à explorer demeurent :

- l'apport de la modélisation déterministe pour fournir une justification et une quantification de "l'écart type" d'un scénario météorologique,
- l'apport de l'approche bayésienne ou markovienne afin d'établir la succession possible de *n* jours pour un scénario météorologique,
- la classification en amont d'une modélisation déterministe de jours (ou scénario type) peut apporter les éléments nécessaires à la construction d'une modélisation simplifiée de type multi-modèles,
- le passage à la prévision peut être envisagé soit, en utilisant une méthode de classification sur des jours (ou scénario) simulés, soit en utilisant les multimodèles (cf. Volet 3),
- la fiabilité de la prévision de la classe ou d'un scénario type en fonction d'une donnée faiblement disponible et souvent imprécise.

2. Identification de scénario type d'apparition d'odeurs autour d'un centre de stockage de déchets

Les centres de stockage de déchets de type Ordures Ménagères et Déchets Industriels Banals assurent la valorisation du méthane et du biogaz vers une filière de production énergétique. Bien que soumis à la réglementation, des nuisances et/ou impacts peuvent apparaître de façon plus ou moins persistante sur l'environnement immédiat du centre de stockage et tout particulièrement des nuisances olfactives liées à la présence d'alvéole en exploitation et ponctuellement au réseau de captage du biogaz.
Les procédés techniques de réduction des odeurs au niveau d'une alvéole en exploitation posent des problèmes non encore résolus à ce jour (pulvérisation de molécules *masques*, dispositif de recouvrement partiel par de la terre et autres matériaux, arrosage du sol, dispositif de ventilation …).

Trois communes, particulièrement touchées durant l'été 2001 par des nuisances olfactives ont été le point initiateur d'un partenariat avec un industriel, la Satrod exploitant le site du Centre de Stockage de Borde-Matin sur la commune de Roche la Molière (42) et le centre SITE de l'école des mines de Saint-Étienne afin de comprendre et prévoir les conditions météorologiques et d'exploitation favorables à l'apparition de nuisances olfactives.

La formulation de ce projet s'inscrit dans le domaine de l'Ingénierie pour l'Environnement avec pour objectif de réduire et anticiper les nuisances et /ou l'émission des molécules odorantes (plus généralement de toute molécule pouvant être issue de ce type d'activité industrielle). Les premiers résultats ont fourni des recommandations de gestion à court terme de l'exploitation industrielle. Or cette recherche thématique s'inscrit dans le vaste domaine du diagnostic et de la prévision d'un processus physique naturel dans lequel intervient un processus anthropique (*ex.* enfouissement contrôlé des OM et DIB).

2.1 Analyse : construction d'une méthodologie

La dispersion et la diffusion d'une odeur *a priori* sont assimilables au transport de molécules spécifiques identifiables et quantifiables sous contraintes des éléments suivants:

- une campagne de mesure menée en partenariat avec le LACE (Université Lyon I) autour du site durant l'année 2002 a effectivement montré la présence d'une cinquantaine de molécules de type Composés Organiques Volatils.
- la perception d'une odeur dépend du type de composant présent, de la combinaison de plusieurs molécules et du seuil de perception, qui dépend de l'individu et notamment de son habitude à inhaler cette composition chimique. Il est difficile d'attribuer à une odeur trace « ordure ménagère », une composition chimique type.
- La physique du transfert de polluant impose de connaître soit la source, soit l'émission: or sur un site de stockage d'OM, on ne dispose pas d'un signal (composition, flux) pour l'émission de molécules gazeuses.
- la mesure de l'odeur (nez humain) ne fournit pas une information précise de l'immission[13] ; elle reste suggestive, est non systématique et non mesurable dans l'absolu (pas d'étalon physique).

[13] immission : en pollution de l'air, ce sont les mesures de la concentration de polluants en un point de l'espace : c'est une mesure résultante du mélange atmosphérique contrairement à l'émission.

Pour envisager des éléments réducteurs de la nuisance, il faut alors établir comment se forment et se déplacent les masses d'air entraînant les molécules odorantes, puis évaluer les moyens d'y remédier sur le site d'exploitation sachant que les moyens de type procédés sont limités, il est inutile dans un premier temps de chercher à asservir le dispositif.

En l'absence de l'émission et de l'immission (mesures fiables mais surtout effectives de flux d'émission), une modélisation déterministe du transfert et de dispersion d'un composant (représenté par un scalaire de type odeur) n'est pas probante dans l'immédiat pour obtenir un champs d'iso concentrations des molécules odorantes sur une période donnée.

Il s'agit alors de :

- décomposer les conditions météorologiques favorables à l'apparition de nuisances olfactives afin d'identifier des scenarii météorologiques caractéristiques selon la méthode proposée pour la pollution à l'ozone dans le § 1.2,
- prévoir un *traceur* (un estimateur *ŷ)* de présence ou non de molécules odorantes en certains points, comme alternative à la simulation de la dispersion[14] en identifiant les conditions météorologiques favorables à l'apparition de nuisances olfactives pour lesquelles des recommandations pour l'exploitation du site sont établies,
- réaliser une modélisation déterministe de la convection de l'atmosphère (dynamique de type passive, sans chimie). La mise ne place d'un modèle de connaissance (type CFD) doit permettre de valider les variables pertinentes de classes de jours type,
- établir les éléments permettant de passer d'un diagnostic à la prévision de la probabilité de présence ou non de molécules odorantes.

2.2 Classification et simulation déterministe de scenarii météorologiques de type jour

Cette première étape transpose la méthode développée pour la qualité de l'air au paragraphe précédent (Batton-Hubert 2000). L'absence de données sur les concentrations d'odeurs impose de se préoccuper uniquement du scénario météorologique - aérodynamique à l'origine du transfert d'odeurs même si d'autres paramètres sont concernés (composition chimique des émissions, composition moyenne de l'air de référence).

Un scénario météorologique est défini par un ensemble de variables pertinentes, décrivant au mieux une situation aérodynamique et son comportement pour laquelle une nuisance olfactive peut apparaître. Dans un premier temps le scénario concerne un jour météorologique, défini ponctuellement sur le territoire (en *(x,y,z*) en Lambert IIe - système de référence national IGN - NGF.

2.2.1 Données disponibles et choix de variables

Les données disponibles sont :

- des données météorologiques mesurées par la station météorologique du CSD (temps d'acquisition de 30 mn) : Température, Pression, Vent, pluie, Humidité
- recensement des plaintes de nuisances olfactives avec : lieu, heure, durée

[14] Ce qui constitue en aucun cas un équivalent à la simulation déterministe

- activités de l'exploitant : date des ouvertures de quai[15]

On cherche à représenter les jours *au mieux* (soient leurs caractéristiques physiques intrinsèques) par un minimum de variables permettant de comparer les jours (Riesenmey 2004).
La température fournit deux composantes, l'état de l'atmosphère et de la dynamique de l'air :

- la chaleur de la journée, par la moyenne des températures T_i sur la journée : $T_{moy}=\frac{1}{48}\sum_{i=1}^{48}T_i$.

- le gradient thermique diurne signe l'amplitude de la variation thermique du jour et le profil de la radiation solaire modifié par la couverture nuageuse ; il est donné par l'écart entre le minimum et le maximum mesurés sur la journée[16] : $T_{dif} = \max(T_i) - \min(T_i)$.

- la pression est représentée par la moyenne des pressions P_i sur la journée : $P_{moy}=\frac{1}{48}\sum_{i=1}^{48}P_i$

- l'humidité est représentée par la moyenne des valeurs d'humidité H_i sur la journée : $H_{moy}=\frac{1}{48}\sum_{i=1}^{48}H_i$.

La pluie caractérise un temps variable ou pluvieux :

- le cumul/ jour est : $Cumul\_pluie=\sum_{i=1}^{48}Pl_i$, avec Pl_i , quantité d'eau tombée entre i-1 et i

Le vent est la mesure directe de la dynamique de l'atmosphère ; quatre variables sont retenues :

- *Um, Vm :* coordonnées cartésiennes du vent moyen/ jour : $Um=\frac{1}{48}\sum_{i=1}^{48}U_i$ et $Vm=\frac{1}{48}\sum_{i=1}^{48}V_i$

- *Inst* et *DVi* décrivent l'instabilité du vent pendant la journée :
- *Inst* représente l'instabilité totale (direction et intensité) des vents par rapport au vent moyen (*Um*, *Vm*). *Inst*[17] correspond à l'inertie totale des points du nuage par rapport au centre de ce nuage, c'est à dire à la distance euclidienne moyenne entre les points et le centre :

$$Inst=\frac{1}{48}\sum_{i=1}^{48}(U_i-Um)^2+(V_i-Vm)^2$$

- *DVi*[18] permet de décrire le type d'instabilité (instabilité en direction ou en vitesse). *DiVi* représente l'allongement du nuage de points le long de la droite passant par l'origine et par le barycentre du nuage : $DVi=\frac{1}{48}\sum_{i=1}^{48}\frac{d_{(Ui;Vi)/droite}}{Inst}$ où $d_{(Ui;Vi)/droite}$ est la distance euclidienne entre le point de coordonnées (Ui,Vi) et la droite passant par l'origine et par (Um;Vm).

Compte tenu des données manquantes, l'échantillon retenu est de 881 jours (entre 01/01/2002 au 30/06/2004).

[15] L'ouverture de quai est un terrassement réalisé dans le stockage des alvéoles d'OM permettant la mise en exploitation d'un nouveau casier de stockage (~5000 m^2 × 5m épaisseur) où sont déversés les entrants du CSD. Ces manipulations durent quelques heures, génèrent des odeurs persistantes (de 1 à ~ 3 jours). La durée d'une alvéole sur ce site est ~1mois .
[16] exemple : une forte différence est synonyme de temps sans nuage, de vents locaux importants
[17] Plus *Inst* est fort, plus le vent est instable
[18] Si *DVi* tend vers 0, le vent est instable en vitesse (rafales) et sa direction reste globalement la même mais peut s'inverser. Si *DiVi* est proche de 1, le vent est très instable en direction mais assez stable en vitesse

2.2.2 Classification hiérarchique

Des méthodes classiques de classification non supervisée sont utilisées pour identifier au mieux le nombre de classes et l'association d'une classe à un état physique plausible de l'atmosphère. Le résultat est conditionné par le choix des variables pertinentes.
L'analyse des données a permis de sectionner 9 variables (§2.2.1) dont la non redondance entre variables doit être vérifiée. L'Analyse par Composantes Principales (ACP) permet de projeter les 881 jours (individus) représentés dans un espace à 9 dimensions dans un espace réduit dont les axes factoriels[19] représentent au mieux les principales caractéristiques du nuage initial. La projection du nuage des individus suivant les 2 premiers axes factoriels (représentant 26 % de l'information pour le 1er axe et 19 % de l'information pour le 2ème axe) corrobore la faible corrélation[20] (et anti-corrélation) observée entre les variables. Ceci justifie de conserver les 9 variables linéairement indépendantes pour la classification suivante. Cette non réduction ne signifie pas pour autant que les variables retenues sont significatives des classes constituées. Seule une étude de la variation du nombre de classes et de leur signification physique peut permettre définir un domaine de validité de la classification.

Identification du nombre de classes

La classification hiérarchique établit une arborescence (Annexe 2.4) dans laquelle les individus sont classés en fonction de leur distance dans l'espace des variables. L'agrégation se fait de proche en proche à partir des *m* individus (soit *m* classes initiales) jusqu'à une seule classe.
A chaque itération, la distance entre les éléments regroupés est évaluée par l'indice de niveau. La courbe des indices de niveau est croissante et comporte des sauts lorsque deux groupes d'individus assez éloignés sont agrégés (figure 2.3.1). Une partition correcte doit couper l'arborescence au niveau d'un saut.

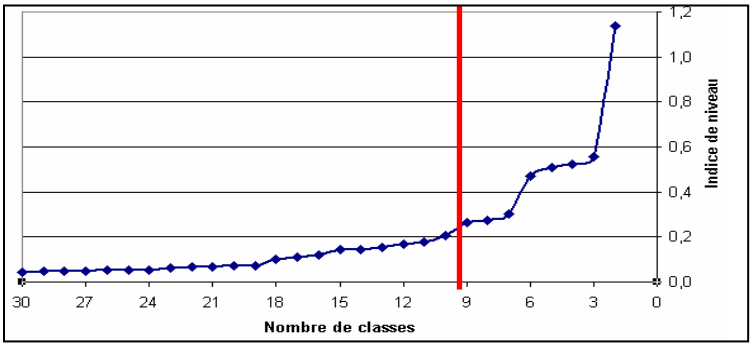


Figure 2.3.1. Courbe des indices de niveau et coupe au niveau du saut entre la 9ème et la 10ème classe

Agrégation autour des centres mobiles et partition des individus

La classification hiérarchique identifie le nombre optimal de classes (ici 6 ou 9 classes). L'agrégation autour des centres mobiles permet d'associer les individus à chacune des classes (annexes 2.3 et 2.4) et de définir pour chaque classe l'individu le plus représentatif (individu le plus proche du barycentre) et l'individu à contribution maximale (individu le plus éloigné

[19] Les axes factoriels retenus sont ceux qui minimiser les écarts entre les distances initiales des points et les distances de leurs projections.
[20] matrice des corrélations = matrice carré de dimension *n*, symétrique : $C = A^t A$ avec *n*, le nombre de variables

du barycentre). Les individus à contribution maximale sont les individus pour lesquels les valeurs des variables représentatives de la classe sont des valeurs extrêmes, situées aux limites du nuage de la classe.

2.2.3 Classes obtenues et interprétation

L'interprétation des classes obtenues utilise l'étude météorologique et micro-météorologique développée par (Batton-Hubert et Vaillant, 2003) et a permis de corréler les épisodes olfactifs aux conditions météorologiques. Neuf scenarii caractéristiques sur l'échantillon de données disponibles ont été définis. Chaque classe (scénario jour) possède une fiche d'identité qui comporte les valeurs moyennes de la classe (9 variables), l'écart à la moyenne de l'échantillon et une brève interprétation de la physique du phénomène ainsi que la statique concernant les plaintes observées (figure 2.3.2).

2.2.4 Passage d'un scénario jour moyen à 9 variables a une simulation par CFD

Les échelles de temps et d'espace fines dans la simulation numérique sont supportées par le logiciel ARPS (Advanced Regional Prediction System - version 5.1.5) (Xue et al, 2000, 2001), proposant un modèle météorologique compressible tridimensionnel non hydrostatique.
Les classes les plus productrices de nuisances olfactives ont été retenues en priorité pour effectuer les premières simulations numériques en l'occurrence, la classe 9, dont on a choisi de modéliser le jour le plus représentatif qui correspond au le barycentre de la classe (*individu du 17 août 2002*). Rappelons qu'ARPS comme tout code de modélisation déterministe a besoin de données de terrain réaliste (topographie, végétation et sol Annexe 2.2). Les conditions limites météorologiques aux bords du domaine utilisent les données du European Centre for Medium-Range Weather Forecasts (ECMWF) qui sont disponibles toutes les 6 heures à différentes altitudes et aux niveaux de pression compris entre 1000 hPa et 50 hPa.
Les résultats obtenus pour différentes configurations permettant de vérifier le basculement des vents locaux lors de ces jours d'été *classe 9* (Chemel et al 2005, Riesenmey et al. 2005, Riesenmey et al. 2006). Notons que les scenarii retenus sont les centres de chaque classe scénario, dont l'instance simulée est le jour de l'année (1999-2003) le plus proche de ce centre (Chemel et al 2005, Riesenmey et al. 2005).
Un exemple de résultat de simulation et une interprétation sont donnés en Annexe 2.5.

2.2.5 Résultats

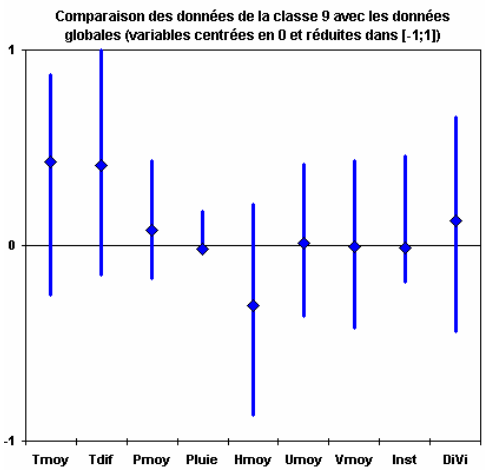
La prescription de recommandations durant la phase d'exploitation du site est effective pour le site de CSD, et constitue un des principaux résultats opérationnels.

Ce couplage entre la classification et l'identification de scenarii jour météorologique puis la modélisation déterministe sont développés dans les travaux de thèse en cours de C. Riesenmey, que j'encadre.

CLASSE 9 : beau temps, journée chaude et humidité faible , sans vent

Descriptif :
193 jours / 881 = 22 % des jours
20 plaintes sur 50 = 40 % des plaintes
Jour représentatif : 17 août 2002
Jour à contribution maximale : 15 juillet 2003

	Température moyenne (°C)	Différence min/max (°C)	Pression moyenne (hPa)	Pluie cumulée (mm)	Humidité moyenne (%)	U moyen (m/s)	V moyen (m/s)	Instabilité du vent	Type d' instabilité
Echantillon									
Moyenne	10,4	9,7	952	2,1	74	0,06	-0,11	1,44	0,57
Ecart-type	19,3	13,6	32	87,7	49	6,42	5,31	3,53	0,57
CLASSE 9									
Moyenne	18,7	15,3	954	0,5	59	0,13	-0,14	1,39	0,64
Ecart-type	13,2	8	11	16,9	28	2,61	2,33	1,67	0,32



Echantillon	RLM		LCF		Firminy	
50 plaintes	été	hiver	été	hiver	été	hiver
Matin	2 *	2 *	16 ****	2	3 *	2 *
Midi	4 *	-	2	1	-	1 *
Soir	-	-	2	3	2 *	8

CLASSE 9	RLM		LCF		Firminy	
20 plaintes	été	hiver	été	hiver	été	hiver
Matin	1 *	-	9	-	2 *	-
Midi	2 *	-	2	-	-	-
Soir	-	-	2	-	2 *	-

Recensement des plaintes pour la classe 9
* : ouverture de quai

Conditions météorologiques types de la classe 9 :

Cette classe comprend 22 % des jours et est caractérisée par une température élevée avec de forts écarts de température. Les vents sont tourbillonnants avec des vitesses faibles. La pression est élevée et l'humidité est très faible, avec des précipitations très faibles. Cette classe correspond aux journées d'été ensoleillées et sèches, sans vent dominant. Le régime de vent local joue un rôle important.
La formation des odeurs est liée à la température élevée avec de forts écarts entre le jour et la nuit, ainsi qu'à la faible humidité.
Les plaintes sur le SE le matin correspondent à des vents très faibles et très instables. Les plaintes au N sont dues, comme pour la classe 8, à des vents d'est ou de sud-est assez forts et très instables ou bien à des vents orientés sud-ouest assez faibles. Les plaintes sur SO sont liées à des vents faibles et instables orientés nord-est, ou alors dans le cas de la plainte en soirée à un vent sud-ouest instable dont la direction se serait inversée en soirée.

Figure 2.3.2 : Description d'une classe -scénario *jour*-

Les résultats obtenus permettent de corroborer les évènements de plaintes et les scenarii météorologiques sur le site du CSD et valident les paramètres retenus pour caractériser un *jour météorologique*.

La station météorologique du site d'exploitation a permis de valider les hypothèses sur la physique du phénomène et notamment la quantification de phénomènes locaux liés aux vents thermiques de pente lors de période de forte conditions anticycloniques, favorables aux nuisance olfactives.

L'intégration d'outils d'analyse spatiale permet un pré-traitement mais surtout un post traitement, qui permet de calculer l'exposition des populations aux nuisances olfactives en croisant les résultats de la simulation (iso concentration en toluène, premiers résultats d'un modèle de dispersion eulérien en cours de développement) et les données de population.

2.2.6 A retenir

L'ACP et la classification utilisées montrent une non corrélation linéaire forte et permet de justifier, soit de l'existence de non linéarité (connue pour la dynamique des équations de Navier et Stokes) soit d'une véritable indépendance entre variables au sens mathématique. La réduction du nombre de variables n'est pas justifiée mais la question subsidiaire est de savoir ce que devient la classification si l'on augmente le nombre de variables (robustesse et sensibilité de la classification).

L'introduction d'autres variables risque de modifier la classification obtenue en augmentant la dimension de l'espace et donc modifie la distance entre individus. En effet il s'agit de données scalaires.

Une approche par vecteurs d'état, où une variable est elle-même dynamique (fonction du temps) permettrait d'intégrer le passage de scénario moyen journalier à un scénario temporel sur 2 ou *n* jours.

2.2.7 Analyse critique des résultats

Les questions à explorer qui demeurent, sont les suivantes :

- que deviennent les méthodes de classification disponibles lorsque l'on intègre des données continues ? comment comparer deux objets dynamiques ?
- la modélisation déterministe d'un scénario météorologique *jour* restitue une information localisée et temporelle - température *T(x,y,z,t)*, pression *P(x,y,z,t), vent U(x,y,z,t)*... Sur un maillage en 2D et 3D : comment intégrer cette donnée simulée pour accroître la *Connaissance* nécessaire à la prévision lorsque les données mesurées n'existent pas ? Peut-on envisager alors une réduction de modèles, en transformant un modèle de type CFD à un modèle plus simple[21] ? Cette information peut-elle être intégrée à un prédicteur ? Des éléments de réponse et des arguments sont proposés au volet 3.

[21] modèle plus *simple* : on entend par là, par exemple, un modèle qui pourrait s'exprimer non pas par des équations aux dérivées partielles (EDP) mais par des équations différentielles ordinaires (EDO)

2.3 Vers une Probabilité de présence ou non d'odeur autour d'un site

2.3.1 Rappel du contexte

Les nuisances olfactives (recensement des plaintes sur [2001 ; 2002] autour du CSD d'étude, sont effectivement liées à des situations météorologiques spécifiques au lieu d'observation et à la période de l'année. Ces situations se déclinent en ~10 scenarii *jours* caractéristiques (§ 2.2). La mise en place de prescriptions et de recommandations destinées à l'exploitation des alvéoles (fiches associées à chaque scénario *jour*) sont mises en place afin de limiter le risque de nuisances olfactives en périphérie.

Ces fiches utilisent les observations *in situ* (croissance de la température ou de la pression, couverture nuageuse …) et fournissent les prescriptions et recommandations pour une journée donnée. Trois périodes d'observation (matin, midi et soir) de suivi identifient alors le type de scénario *a posteriori*, soit le soir : les prescriptions actuelles concernent le jour même.

Pour anticiper les actions à mettre en place pour le lendemain $j + 1$, il est nécessaire d'identifier la persistance du type de temps du jour j, en fournissant la prévision pour le lendemain du scénario.

Ces travaux posent les jalons de la faisabilité et le développement de méthodes d'analyse des tendances et de prévision, utilisant les observations de la veille j et des jours antérieurs ($j + 1$, $j + 2$, $j + 3$ …) avec l'ensemble [données météorologiques + classes des scenarii jour + plaintes] pour construire un « prédicteur » pour le jour $j + 1$.

2.3.2 Identification des frontières de décision

La faisabilité de méthodes d'analyse des tendances au jour j pour le jour $j + 1$, soit un prédicteur du type de temps, soit d'un estimateur d'une présence ou non d'odeur (ou plainte) est conditionnée par les variables explicatives et l'intervalle possible de leurs instances dans le système dynamique.

Une classification automatique des scenarii a permis de valider le nombre et les paramètres caractéristiques des scenarii (paragraphe 2.1) (Riesenmey 2004) : chacune des classes définies à cette étape correspond à un scénario *jour*. L'analyse spécifique des plaintes pour chaque scénario permet d'établir un ensemble de règles formelles[22], qui compte tenu de faits observés (faits et assertions données), identifie la cible potentielle (lieu) et l'heure ou une plainte peut être éventuelle voir son type d'odeur (en continu, par bouffée...).

Règles d'apparition

Des règles d'apparition pour chacun des scenarii de plainte ont été construites (Batton-Hubert et Vaillant 2003), pour les périodes les plus favorables aux nuisances olfactives hivernales ou estivales (avril-septembre) :

[22] ces règles ne sont pas implémentées dans un véritable système expert car elles demeurent statiques et sont proposées sous la forme de fiches descriptives à ce stade de l'étude

ℜ1 : *Odeurs vers SE ville L & lieu-dit (S-SE) par bouffée + ou – continue* : [
Le Matin, **avant 10h00** (heure locale) **SI** {vitesse faible (<1m/s), vent local nord, température nocturne minimale >10°C, forte pression atmosphérique sur plusieurs jours}
Entre 10h00 -13h00 (heure locale) **SI** {vent de 1-3 m/s et variation du vitesse du vent autour de Nord, température > 15 °C, hauteur de l'inversion thermique faible}
Après-midi SI {vent Sud-SE fort >4-5m/s avec passage localement au Nord, température après-midi à 15h00 >15°C}]

Pour chaque règle, le fait déduit concerne : la direction de la cible, la période de la journée, la dynamique d'odeur possible (continue, bouffée). Chaque règle utilise des seuils discriminants pour chaque variable, établis par l'analyse des résultats obtenus par la simulation numérique déterministe et retranscrit les basculements des circulations locales à l'origine des impacts olfactifs dans 3 directions différentes (N, SE, SO par rapport au site).

ℜ2 : Odeur vers le S-SE et vers S-O : [
Le Matin, Avant 10h00 (heure locale) **SI** {vitesse faible (2m/s), température ~10°C, minimum des températures nocturne ~5 °C}]

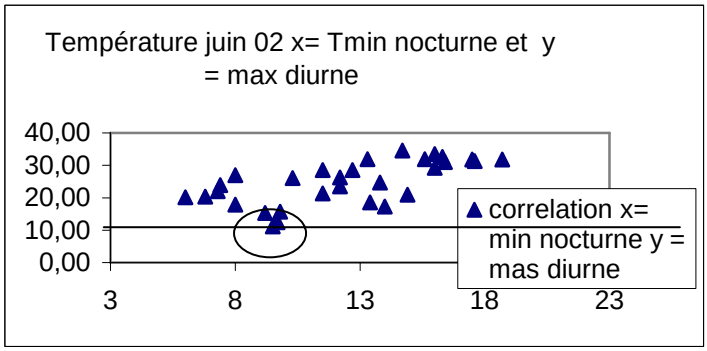
Dans certains cas, la période cible, (soit la potentialité à ressentir une odeur) se manifeste "*sur tout le jour* "(*règle ℜ3*). Au contraire, les périodes non indiquées où les cibles azimutales sont non spécifiées, ont une potentialité d'odeur nulle (pas de potentialité d'odeur lié au CSD pour la direction O et NO).

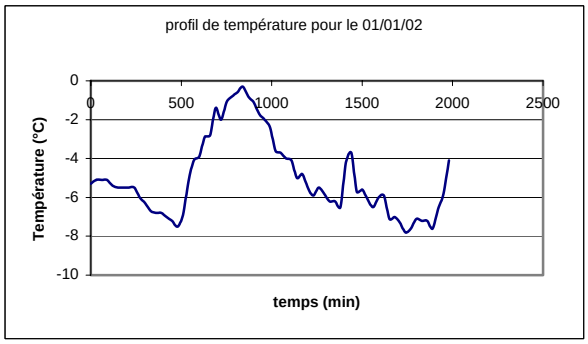
ℜ3: Odeurs vers le Nord : par bouffée + ou – continue : [
Nuit, et sur le jour SI {vent faible durant cette période de la journée S-SE, température > 15 °C}]

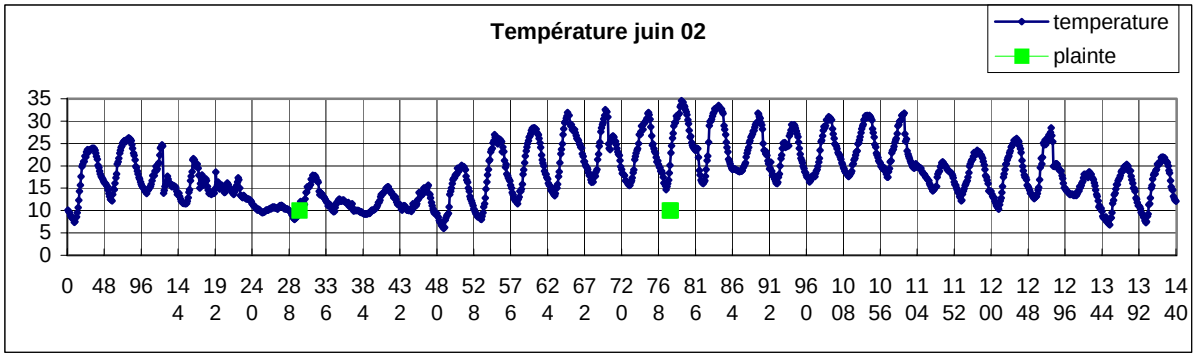
Sensibilité aux variables explicatives et aux frontières de décision

Une analyse statistique des données montre l'impact du choix des variables représentatives des jours par un scalaire moyen, variable discriminante du comportement thermique de l'air.

Considérons la courbe des températures (figures2.3.3) des relevés juin 2002 :
Pour les jours de beau temps, une sigmoïde régulière caractérise le jour avec un fort abaissement de la température nocturne (Dt_{noc} ~10°c) ; le jour, un fort ensoleillement entraîne un fort accroissement diurne de la température (Dt_{di} ~ 10°C). La température maximale, Tmax température est relevée aux heures les plus chaudes entre (12h00 et 16h00 en TU). La température nocturne minimale est le minimum entre 3h00 et 6h00 (TU). Classiquement on pourrait s'attendre à discriminer les jours de beau temps par l'assertion : si forte température de jour et faible température la nuit alors il s'agit d'un jour de beau temps.
Or la corrélation Tmax. en fonction de température Tmin nocturne montre que la règle [Tmax > 20 et Tmin. > 10 °C] n'est pas suffisamment sélective des jours de beau temps ; en effet les premiers jours du mois sont de type temps ensoleillé mais ayant des valeurs de température moyennes plus faibles, ils ne sont pas identifiés (idem en fin de mois).
Par compte les jours (6 et 7 juin) ayant un faible ensoleillement sont séparables des autres, de cette façon. Le jour du 5 juin se caractérise par une température minimale Tmin > 13 ° et une température Tmax maximale sur la journée de l'ordre de 17°C ; le comportement du jour n'est







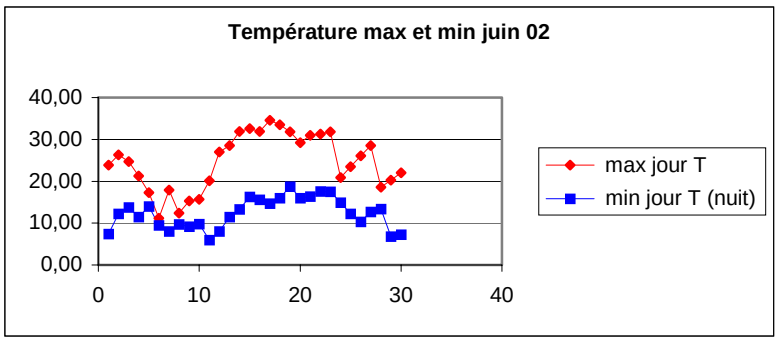
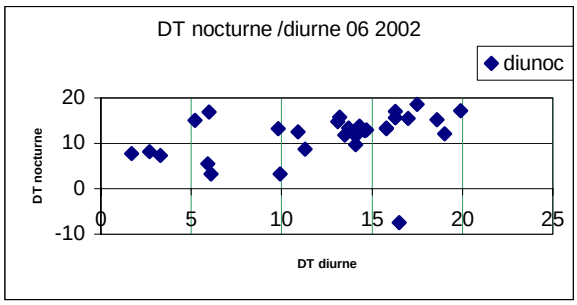
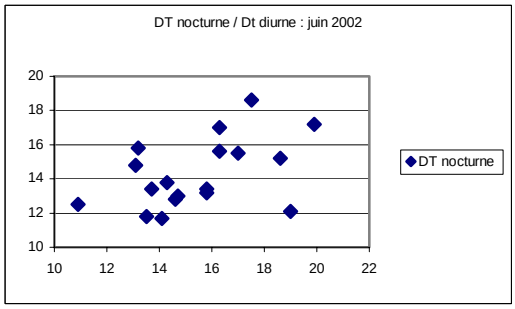


Figure 2.3.3. : Température de juin 2002 : Tmin. et Tmax.





Coefficient de corrélation : 0,815

Figure 2.3.4. Dt nocturne et Dt diurne sur le mois de juin 2002 (différence de températures entre 2 extremums) - filtre Dt_{noc} et $Dt_{diu} > 10$

pas semblable aux jours suivants: la variable *Tmin -Tmax* n'est pas suffisante pour le détecter. De même les seuils de Dt_{noc} >10°C et Dt_{di} > 10°C (figure 2.3.4) devraient discriminer les jours de forte signature de type beau temps anticyclonique mais ceci ne semble pas être suffisant. En effet la corrélation entre Dt_{noc} et Dt_{di} reste faible, de l'ordre de 0 ,4 pour l'ensemble des données du mois de juin. L'application d'un filtre aux Dt nocturnes et diurnes >10°C améliore la corrélation entre les 2 variables avec un coefficient de corrélation de l'ordre de 0 ,810. On peut alors identifier un fort abaissement de la température nocturne qui est souvent accompagné d'un fort accroissement de la température diurne pour le mois de juin. On identifie alors 60 % des jours du mois de juin mais une telle règle ne peut identifier les jours comme les 25-27 juin qui ont de plus faible température, une bonne signature de radiation solaire mais des Dt plus faibles.

A ce stade, les règles sont liées à des fiches de prescription utilisées de façon empirique sous la forme de recommandations :

[si, travaux prévus JOUR j **et si** type de temps identique à j - 1 **Et** beau temps alors prévoir travaux durant la période d'ensoleillement maximale (10 -16 h TU)].

Il apparaît le constat suivant :

- i) la classification sur des données moyennes scalaires ou vectorielles représentant un jour a permis de discriminer les groupes (clusters) de jours météorologiques et d'établir les scenarii jours météorologiques. Lors de la partition en classes, la similarité entre individus est calculée par une distance qui agrége différentes valeurs centrées réduites de l'individu selon un critère à minimiser (de variance intra classes minimale et une variance interclasse maximale).
- ii) Les règles expertes cherchent à réduire successivement l'espace de validité de l'individu sur des critères logiques de type exclusif où chaque variable est examinée successivement.

Par conséquent, les variables discriminantes obtenues à l'aide d'une méthode ne sont pas suffisantes ou adaptées pour une seconde méthode et mettent en évidence la nécessité d'intégrer ou modifier les variables explicatives afin de faire "cohabiter" deux types de raisonnement. Il semble incontestable d'exprimer le gradient de température, $\frac{dT}{dt}$ sous une forme explicite ou implicite.

L'expression sous forme de règles met en exergue des paramètres ou caractères propres à la dynamique de l'air sur plusieurs jours (tendance) ou selon une dimension verticale (profil 3D) non disponibles sur les mesures sur site.
Bien que l'utilisation de méthodes types séries chronologiques pour établir la tendance (MAV -Mobile Average) sur le température et la pression puisse apporter de l'information pertinente à l'évaluation « *forte pression atmosphérique sur plusieurs jours passés* » elle interroge sur la représentativité d'un scénario *jour* par des valeurs et non par des variables d'état (*x(t)*).

2.3.3 Du diagnostic à la prévision d'une odeur

La question est de savoir si l'on peut envisager un *traceur* (un estimateur $\hat{y}$) de présence ou non de molécules odorantes en certains points, comme alternative à la simulation de la dispersion ?

En absence de mesures d'odeurs (soit directes soit à l'aide d'un suivi par des nez humains) ou de nez électronique, il reste comme seules manifestations d'odeur, le relevé de plaintes. Cependant, cette donnée est qualitative, subjective et non systématique. Les relevés disponibles présentent toujours le même constat : lorsqu'une période météorologique présente des conditions propices à une plainte olfactive, il n'y a pas systématiquement plainte. Trois explications sont avancées : il n'y a pas de plainte parce qu'il n'y a pas de personne confrontée à cette gêne à ce moment de la journée, ou bien les personnes concernées ne formulent pas la plainte, ou bien il n'y a pas effectivement d'apparition de nuisance olfactive. Ces 3 hypothèses sont difficilement discernables.

L'enregistrement d'une plainte après avoir été validé comme possible (§2.2), est considéré comme une information *vraie c.a .d* : une odeur a été perçue effectivement à cette date, en ce lieu. Par conséquent les plaintes deviennent des données vraies ou fausses, mais non systématiquement observées. Ce qui entraîne un très faible échantillonnage. La probabilité explicite a priori de la plainte ne peut être approchée à partir des échantillons (loi des grands nombres non vérifiée).

Cette étude concerne la faisabilité d'une méthode robuste pour prévoir la persistance d'un type de temps (par exemple, persistance d'un temps ensoleillé de type anticyclonique) observé le jour courant pour le lendemain, et/ou prévoir s'il y a un possibilité de plainte ou non en aval, le lendemain, en fonction des données météorologiques de la veille.

Choix de 2 types de prédicteurs

Deux types de prédicteurs fondés sur le concept commun du scénario météorologique ont été retenus:
- un modèle d'analyse des tendances micro-météorologiques : estimateur du type de scénario météorologique du j+1 (au jour j)
- un prédicteur au jour j de la possibilité d'avoir une odeur observable en aval au jour j +1

Le premier, se base sur l'identification puis la prévision de la classe de scénario observé a partir de la classification établie Cette première étape doit transposer la méthode développée pour la qualité de l'air au paragraphe 1.(Batton-Hubert 2000). L'absence de données sur les concentrations d'odeurs impose de se préoccuper uniquement du scénario météorologique et aérodynamique à l'origine du transfert d'odeurs même si d'autres paramètres sont concernés (composition chimique des émissions, composition moyenne de l'air de référence). Cette analyse constitue un des éléments attendus de la thèse en cours de C. Riesenmey (§2.2).

La seconde étape consiste à développer une classification de type probabiliste (avec les deux phases d'apprentissage et de prédiction) à partir de l'identification des frontières de décision qui soient discriminantes. La faisabilité d'un modèle semi-predictif[23] a été établie sur la

[23] Cette notion de semi-predictif s'explique par le fait qu'il n'y a pas de définition d'une seule quantité $\hat{y}$ mais peut être de plusieurs variables à prédire $\hat{w}$ et que le nombre de classes peut être alors connu

température et semble possible (Projet PRIMO2, 2004). Le paragraphe suivant présente les perspectives de l'utilisation de la probabilité conditionnelle de réseaux bayésiens mais aussi les problèmes de résolutions numériques identifiés et non résolus.

Le choix d'un modèle de type réseau bayésien se justifie par les éléments suivants :

- la prévision d'un scénario météorologique, donne une information sur le type de temps au lendemain afin de décaler ou de modifier des tâches, si possible, lors de l'exploitation. Cette prédiction ne fournira pas directement une information sur la possibilité d'avoir des odeurs en aval donc, des plaintes,
- les données concernant les plaintes sont des données non exhaustives : la relation, *il n'y pas de plainte donc pas d'odeur* est fausse, *et il y a odeur et donc plainte* n'est pas vérifiée non plus. La relation conservée est alors : *s'il y a odeur (avec une possibilité) alors il peut avoir une plainte*, on pourra informer éventuellement la population d'une période à risque de nuisance olfactive,
- les données utilisées par ce type de modèle sont les mêmes que celles identifiées pour les scenarii.

L'intérêt du raisonnement bayésien, est d'exploiter l'expertise déjà acquise sous forme de règles (seuil de basculement de la circulation atmosphérique en fonction des paramètres météorologiques) et d'associer une probabilité à l'observation des états du système (exemple : *probabilité d'avoir un temps ensoleillé, sachant que la température est de 20°C*).

2.3.4 Prédicteur d'odeur jour j: possibilité d'avoir une odeur observable au jour j +1

Il s'agit d'établir le graphe du réseau bayésien permettant de déduire la possibilité (probabilité) d'une plainte sur une des 3 villes en aval du site.

Prédicteur de plainte : premier réseau bayésien

Les paramètres suivants ont été retenus :

- la variation de température sur 24 heures permet d'avoir une idée de la distribution de température au cours de la journée, renseigne sur le type de temps,
- la direction du vent,
- la vitesse du vent,
- l'humidité,
- le type de pression (haute pression, basse pression, pression variable).

Les intervalles de valeurs de chaque paramètre sont conditionnés par le fait que l'on cherche à limiter au maximum le nombre de valeurs prises par les différents nœuds du réseau. En effet, plus un nœud a de valeurs différentes, plus les paramètres à estimer (probabilités conditionnelles) sont nombreux et la convergence de l'algorithme difficile à atteindre.

Pour chaque paramètre (variable, nœud du graphe) les seuils suivants ont été choisis :
- pour la variation de température[24], ΔT, on fixe un seuil à 200 °C, donc 2 états :
  Inférieur à 200
  Supérieur à 200

[24] variation de température : $\Delta T = (T_i(12h00) - T_i(04h00))^2 + (T_i(12h00) - T_{i+1}(04h00))^2$, avec i représentant le jour. En effet, le plus important pour caractériser la variation de température est le différentiel entre la température maximale et la température minimale

- pour l'humidité, on fixe trois seuils : 25%, 50% et 75%, donc quatre états :
    De 0 à 25%
    De 25 à 50%
    De 50 à 75%
    De 75 à 100%

- pour la direction du vent, trois seuils de 90°, 180° et 270° fournissent quatre états :
    De 0 à 90°
    De 90 à 180°
    De 180 à 270°
    De 270 à 360°
- pour la vitesse du vent, trois seuils de 1m/s, 3m/s, 5m/s, fournissent quatre états :
    0 et 1 m/s
    2 et 3 m/s
    4 et 5 m/s
    6 m/s et plus

- pour la variation de pression, 2 variables[25] sont utilisées : $\Delta P_1 = |P_i(12h00) - P_{i-1}(12h00)|$ et $\Delta P_2 = |P_i(12h00) - P_{i+1}(12h00)|$. Un même seuil est fixé pour ΔP_1 et ΔP_2 : 10Mpa, et un seuil pour la pression actuelle de 950Mpa. Trois états permettent de caractériser la pression :
    Haute Pression : Pression supérieure à 950 Mpa , ΔP_1 et ΔP_2 inférieurs à 10Mpa
    Basse Pression : Pression inférieure à 950 Mpa, ΔP_1 et ΔP_2 inférieurs à 10Mpa
    Pression variable : ΔP_1 ou ΔP_2 supérieur à 10Mpa

Un premier réseau bayésien, a été construit pour évaluer le risque de plainte. Chacun des facteurs est indépendant. Les 5 facteurs (ou paramètres) conditionnent tous directement le risque de plaintes figure 2.3.5 :

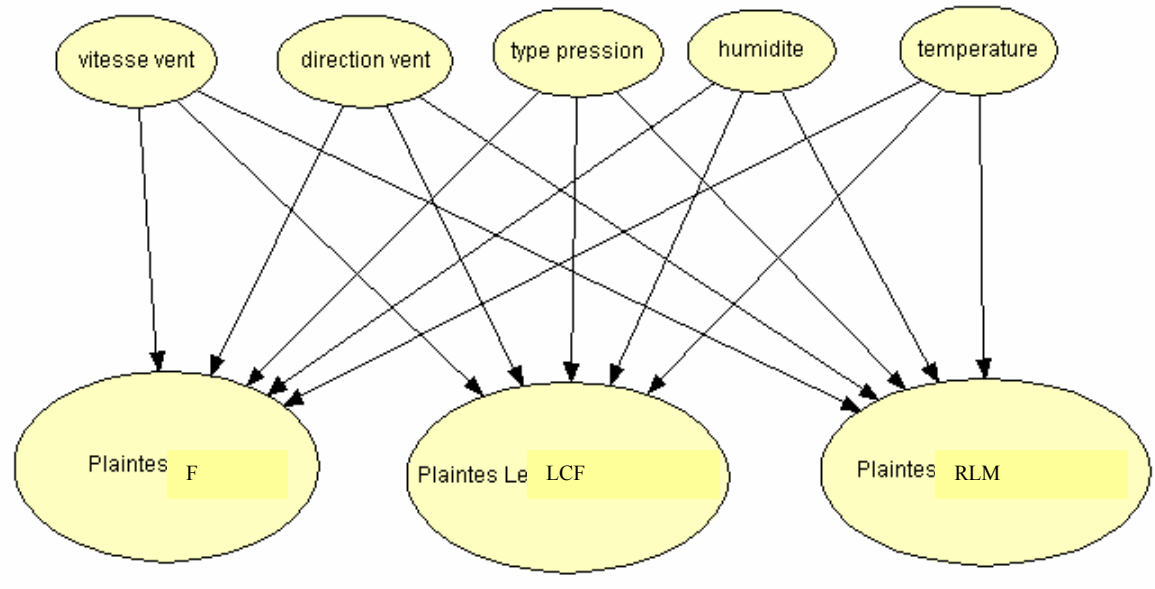


Figure 2.3.5: Premier réseau bayésien du risque de plainte en aval

[25] i représente le jour

Les états pour les nœuds « Plaintes au lieu F », « Plaintes au lieuLCF », « Plaintes au lieu RLM » sont des états binaires [0, 1]. Deux réseaux identiques sont construits, un pour l'hiver (tableau 2.3.1) et un pour l'été.

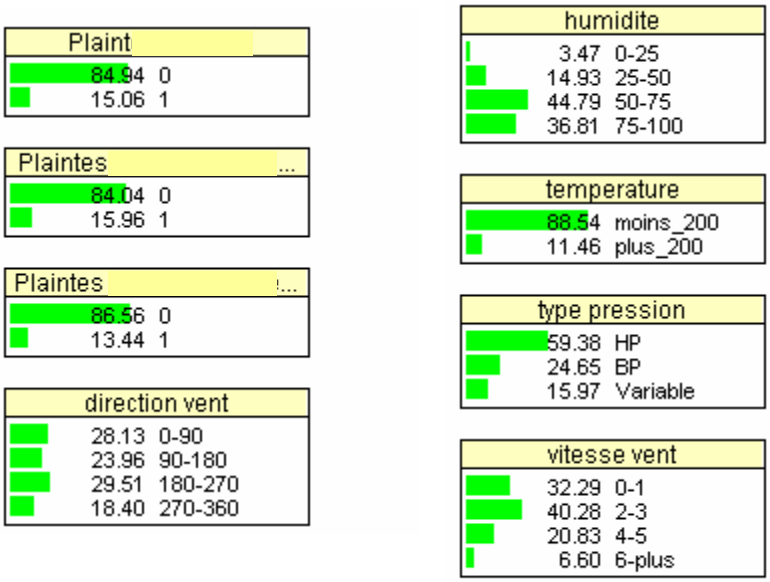


Tableau 2.3.1 Probabilités pour les différents nœuds pour l'hiver

Les résultats sont peu probants en effet on obtient une probabilité de plainte à RLM de 13,44% pour l'hiver alors qu'il n'y a pas eu de plainte à RLM pendant la période hivernale. L'algorithme semble fournir une équiprobabilité de 0,5 lorsqu'il ne peut calculer la probabilité conditionnelle. Cette non convergence de l'algorithme serait liée au fait que trop de paramètres présentent des combinaisons qui sont des situations non réalistes pour les données disponibles. Ainsi la situation où le différentiel de température est inférieur à 200 °C, l'humidité est inférieure à 25%, la pression est haute, la direction du vent est comprise entre 0 et 90°, et la vitesse du vent est nulle ou de 1 m/s ne doit pas être un cas possible dans le modèle ; c'est un cas contradictoire, une haute pression implique du beau temps alors qu'un différentiel de température faible indique un temps variable.

Prédicteur de plainte : second réseau bayésien

Un deuxième modèle, plus réaliste qui utilise les scenarii types (§2.) a été construit pour évaluer le risque de plainte. Les trois principaux facteurs influençant directement le risque de plainte sont la vitesse du vent, la direction du vent et le type de temps : beau temps, pluie, temps variables.
Le type de temps est directement influencé par le type de pression, le différentiel de température et l'humidité. Deux réseaux identiques ont été construits, un pour l'hiver et l'autre pour l'été, figure 2.3.6.

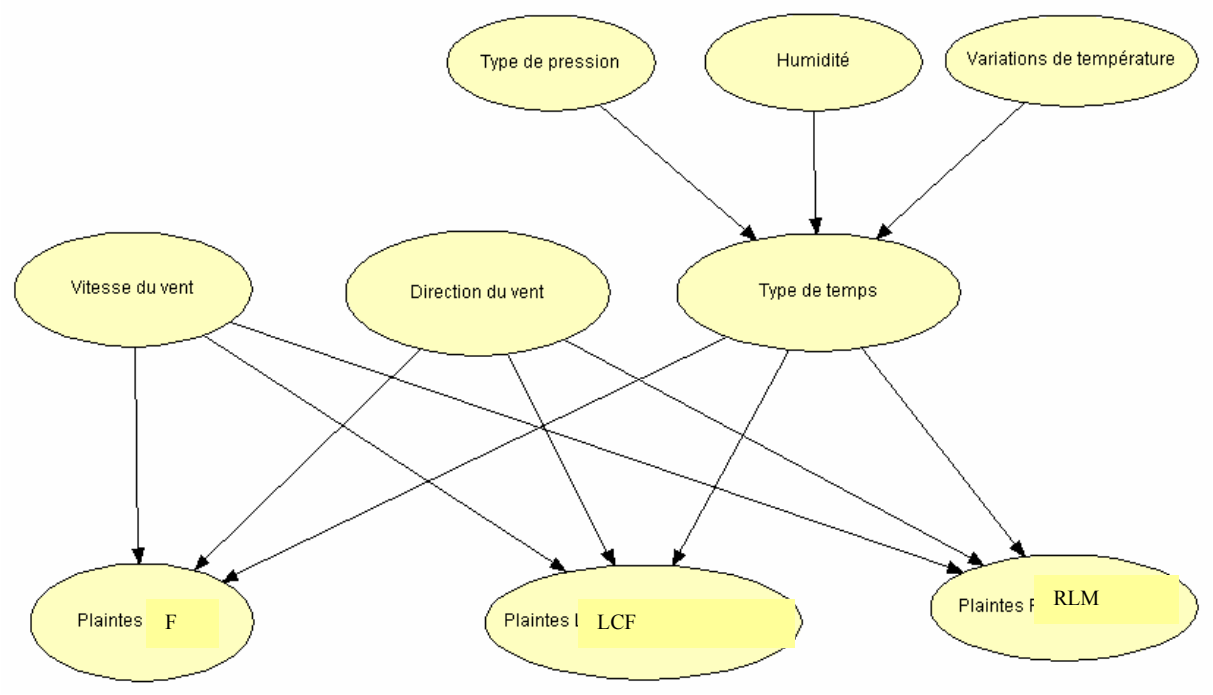


Figure 2.3.6 : Second modèle de réseau bayésien du risque de plainte en aval

Le type de temps : « Type de temps » est une variable cachée dans le modèle. On obtient les probabilités suivantes pour chacun des nœuds, pour l'été (tableau 2.3.2).

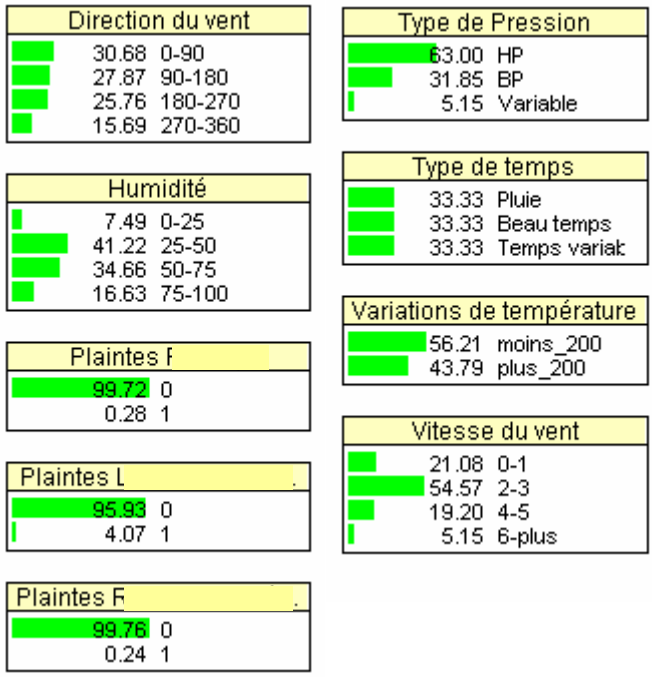


Tableau 2.3.2 Probabilités pour les nœuds : graphe établi pour l'été (2ème modèle)

Dans les tables de probabilité, il y très peu de situations où l'on ne possède pas « d'expérience » pour caractériser les probabilités conditionnelles. Le problème précédent du

trop grand nombre de paramètres à identifier pour les différents nœuds est résolu. Cependant les probabilités du nœud « Type de temps » sont des probabilités par défaut, l'algorithme utilisé[26] n'a pas réussi à identifier les paramètres du nœud « Type de temps » ni les probabilités conditionnelles des 3 cibles.

Problème de l'identifiabilité du réseau

La non convergence du modèle 2 est lié au à un problème d'identifiabilité :

Définition :
On dit d'un réseau qu'il est identifiable si pour deux valeurs différentes des paramètres du modèle (probabilités indépendantes pour les nœuds parents et probabilités conditionnelles) $\theta_m' \neq \theta_m$,s'il existe une observation y (ensemble des probabilités pour qu'une certaine situation soit réalisée) telle que $P(y|\ \theta_m) \neq P(y|\ \theta_m')$.

Dans notre cas, il existe plusieurs paramètres du réseau différents pour lesquels tout y et y' appartenant au domaine des observations, les probabilités soient égales.

On note 1, 2 et 3 les différents états de la variable « Type de temps », et T cette même variable.
On a :

$$P(y|\ \theta_m) = P(y|\ T=j,\ \theta_m)$$

Pour chaque nœud observé et directement en relation avec « Type de temps » on a donc (nombre d'état -1)*2 valeurs de probabilités conditionnelles à évaluer, soit 16 valeurs. A ces probabilités conditionnelles s'ajoutent deux probabilités P(T=1), P(T=2) et P(T=3) : soit 19 inconnues. Des situations observées, permettent d'écrire des équations pour évaluer ces inconnues. Cependant certaines équations sont liées.

La formule suivante permet d'évaluer si le même nombre d'équations est égal au nombre d'inconnues.

Soient : r_i le nombre d'états dans chaque nœud observé directement lié à « Type de Temps »
r_T le nombre d'états du nœud « Type de Temps »

Si :

$$\left(\prod_{i=1}^{n} r_i\right) - 1 < r_T\left(\sum_{i=1}^{n}(r_i - 1)\right) + r_T - 1$$

Alors le nombre d'équations indépendantes est insuffisant comparé au nombre de paramètres à estimer.

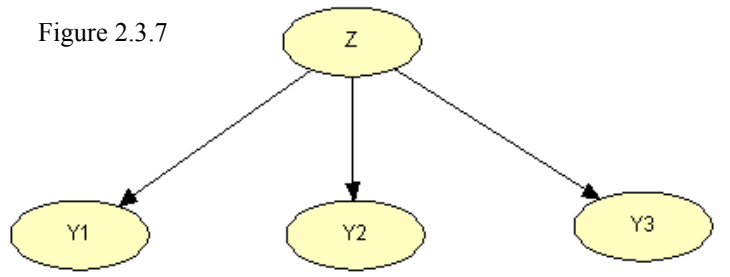
L'application de cette formule à la partie inférieure du graphe (figure 2.3.6) avec 3 nœuds cibles, conséquences possibles du type de temps, on obtient 7 en partie gauche[27] de la

[26] Tests ont été réalises logiciel Hugin, qui utilise l'algorithme EM pour un apprentissage des paramètres du réseau.
[27] (2*2*2) – 1 = 7

formule, et 11 pour la partie droite[28] de la formule: il manque des équations pour résoudre ce problème.

Ces différents problèmes d'identifiabilité du réseau conduisent à proposer plusieurs améliorations qui consistent à :
-fixer certains paramètres du nœud « Type de Temps » en fonction des paramètres météorologiques. Cela nécessiterait d'avoir des valeurs précises des probabilités conditionnelles,
- résoudre le problème du niveau inférieur du réseau identifié précédemment, en exploitant le cas suivant avec un nœud du type Z de la figure 2.3.7:

Figure 2.3.7



où Z est une variable binaire ainsi que Y1, Y2 et Y3 : le système est identifiable.

En effet : $\left(\prod_{i=1}^{n} r_i\right) - 1 = 7 = r_T\left(\sum_{i=1}^{n}(r_i - 1)\right) + r_T - 1$

On peut alors envisager de séparer « Type de temps » en trois nœuds binaires :
- Type de temps 1 comprenant les états « Beau temps » et « autre »
- Type de temps 2 comprenant les états « Pluie » et « autre »
- Type de temps 3 comprenant les états « Temps variable » et « autre »

De même, il serait intéressant de tester la réduction des états des nœuds « Type de pression » et « Humidité » pour que les variables deviennent binaires au détriment du modèle physique puisque la caractérisation de la pression s'effectue avec trois états au minimum.

Le modèle de réseau bayésien à ce jour qui semble probable est donné en annexe 2.7 mais un problème majeur subsiste, *vérifier que le système a une solution.*

2.3.5 Résultats - À retenir

Actuellement un prédicteur de type réseau bayésien semble tout à fait intéressant si le problème de la résolution du système d'équation constituant le noyau du modèle peut être levé. Ce mode de raisonnement est cohérent avec les règles d'apparition des conditions météorologiques mais impose d'avoir les intervalles de valeurs de chaque variable aléatoire et notamment de la variable *type de temps*.

La variable (ou paramètre) *type de temps* pour laquelle le graphe présente 3 arcs entrants et 3 arcs sortants n'est pas identifiable faute de probabilité a priori et d'instances dans la base d'apprentissage. La définition de ce paramètre et des seuils correspondants est délicate. En

[28] 3(1+1+1) + 3 –1 = 11

effet, les paramètres du réseau doivent avoir une certaine indépendance. Deux possibilités sont à envisager:
- soit, on utilise les classes de temps issues de la définition de scenarii météorologiques *jour* ; cette information calculée (§2.2.1) utilise les mêmes données brutes (mesures de direction du vent, de température..) mais pour des variables explicatives différentes de celles utilisées dans le réseau ; ce qui ne garantie pas l'indépendance,
- soit on utilise la variable *type de temps*, variable qualitative d'observation disponible dans les données de Météo France, décrivant plus exactement "l'aspect du ciel "(nuage, brouillard, pluie, soleil….). Les tests et la comparaison de résultats devront être réalisés pour finaliser ce modèle.

Le calage des probabilités a priori est délicat compte tenu du faible nombre d'observations de plaintes par an (~25 ± 10 plaintes/an). Pour les seuils des autres variables, les classes de scenarii[29] et leur modélisation déterministe devront à terme permettre de construire les tables de probabilité initiales nécessaires au modèle.

La construction du graphe du réseau bayésien représente une autre difficulté: remarquons que pour les arbres de décision construits dynamiquement, plusieurs critères d'optimalité fournissent l'arbre optimal pour une classification. Pour cela, on choisit les attributs (variables) ayant le plus d'influence ou qui introduisent le gain d'information le plus élevé (soit l'entropie la plus faible). Au contraire, le choix du graphe pour le réseau bayésien est de type expertise et conditionne le calcul de probabilités conditionnelles (direction des arcs du graphe + probabilités évaluées à partir de l'échantillon) et ne permet pas de garantir une identifiabilité du réseau. Il n'y a pas de critère disponible actuellement de construction du graphe qui pourrait assurer une optimalité du graphe.

Une approche intéressante serait l'utilisation d'un modèle de type régression logistique adapté à un signal de type booléen de type [*odeur, pas odeur*].

Compte tenu de ces résultats, la faisabilité d'un outil peut être établie pour un type de prédicteur d'odeur et de scénario permettant de passer d'un diagnostic à la prévision - probabilité de présence de molécules odorantes ou de plainte.

3. Synthèse du chapitre 2

Ces travaux s'inscrivent dans la préoccupation actuelle de l'évaluation des impacts sanitaires liés à la qualité de l'air (pollution et nuisances). Etant donnée la dualité entre l'approche déterministe usuelle en diagnostic (répartitions d'un champ de concentrations en ozone) et la modélisation dite de représentation pour le déclenchement des seuils d'alerte et la prévision à court terme, nous proposons la complémentarité au sens d'objet conceptuel par le biais d'une interopérabilité qui permettrait d'accroître la connaissance du processus physique concerné.

La simulation par CFD de scénario montre la difficulté de valider les résultats de la simulation sur des secteurs peu instrumentés et la nécessité d'intégrer la faible disponibilité des données, en pré traitement et en post traitement.

Du développement de méthodes de classification pour l'obtention de scénario, ressortent un point essentiel et les deux questions suivantes :

[29] La modélisation déterministe calcule l'état 3D d''un scénario en validant ou non ces valeurs de seuils et de l'écart type de la valeur

- l'identification des variables pertinentes compte-tenu du dilemme entre données /variables disponibles et variables pertinentes conditionne l'instabilité de la classification dans certains cas mais aussi la possibilité d'intégrer des phénomènes synoptiques et locaux, ceci constitue un élément intéressant.

Les deux questions sont :

- pour différentes techniques de classification un indice de confiance sur les variables explicatives de la classification peut il améliorer et assurer une sorte d'optimalité[30] de la classification obtenue?
- les méthodes de classification disponibles concernent des données moyennes; en intégrant des données continues que devient la similitude entre 2 jours sur des données non stationnaires (température, vent, ozone) *a posteriori* entre 2 - 5 jours ? Comment comparer deux objets dynamiques ?

Enfin, la modélisation déterministe d'un scénario météorologique *jour* restitue une information localisée et temporelle, sous-ensemble composé de {température *T(x,y,z,t)*, pression *P(x,y,z,t), vent U(x,y,z,t)...*} sur un maillage en 2D et 3D permettant de passer d'une information moyenne à une information dynamique en espace et en temps. D'un point de vue classification, la question est de savoir si cette information peut être intégrée dans un processus de classification et dans un modèle de prévision et si oui sous quelle forme ?

Enfin cette information simulée accroît la *Connaissance* nécessaire à la prévision lorsque les données mesurées n'existent pas. Elle pourrait mettre en évidence des comportements locaux qui permettraient d'envisager une modélisation dynamique localisée.

Applications possibles en Environnement

Ces travaux pourraient trouver une continuité à la suite de l'analyse des dysfonctionnements précédemment présentée sur les réseaux d'adduction en eau potable et qui pourraient aussi s'adresser aux réseaux d'assainissement pour lesquels de nombreuses approches ont été développées dont l'analyse de redondance entre un modèle et des données mesurées (Piatyszek 1998). L'expérience montre que les modèles physiques ne sont généralement pas opérationnels pour une exploitation en temps réel. Le calcul de probabilité d'appartenance à une classe de dysfonctionnement préalablement identifiée dans une typologie ou par un réseau bayésien représente une alternative pour ces applications ; ce qui montre la valeur générale de ce type de méthodes.

Travaux produits :

- DEA C Riesenmey - Diagnostic d'un CET source de nuisances olfactives à l'aide de modélisation déterministe et de campagne de COV, DEA "Sciences et Techniques du déchet", Ecole Nationale Supérieure des Mines de Saint-Étienne 2004
- 4 contrats de recherche industriels : 2003 - 2005
- 1 projet XIème contrat Plan Etat Région, programme fédérateur Environnement,2000
- 11 articles en colloque avec actes
- 3 articles en revues et ouvrages dont un chapitre d'ouvrage collectif
- une thèse en cours

[30] certains indicateurs comme les matrices de confusion peuvent valider ou non une classification mais seulement lorsque'il est possible d'instacier une classe à un objet de façon supervisée a posteriori

Bibliographie

U.S. Environmental Protection Agency (1999) Guideline for developping an ozone forecasting program. EPA-454/R-99-009 .

U.S. Environmental Protection Agency (1999) Draft guidance on the use of models and other analyses in attainment demonstrations for the 8-hour ozone NAAQS. EPA-454/R-99-004

Kuebler J. (2001). Integrated Assessment of photochemical air pollution control strategies : method development and application to the swiss plateau.

Cobourn W. & Hubbard M. (1999) An enhanced ozone forecasting model using air mass trajectory analysis. *Atmospheric Environment,* **33**, 4663-4674

NOAA (1998) HYbrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT4) http://www.arl.noaa.gov/ready/hysplit4.hml, NOAA Air Ressources Laboratory, Silver Spring, MD, USA.

Fernandez Mills G. & al. (1992) Une méthode de classification pour les types de temps à Barcelone. *Météorologie,* **43-44**, 43-51

Le Hégarat-Mascle (2005) Classification d'images de télédétection pour l'estimation et le suivi de paramètres géophysiques, Thèse d'Habilitation à Diriger des Recherches, Université de Versailles - Saint-Quentin en Yvelines et Institut Simon Laplace, 2005, 105 p.

Ludwig F.L., Jiang J. & Chen J. (1995) Classification of ozone and weather patterns associated with high ozone concentrations in the San Francisco and Monterey bay areas. *Atmospheric Environment,* **29**, 2915-2928

Kalstein L., Tan G. & Skindlov J. (1997) An evaluation of three clustering procedures for use in synopic climatological classification. *Journal of climate and applied meteorology,* **26**, 717-730

Lebart,L., Morineau,A. & Piron,M**.** (1997) Statistique exploratoire multidimensionnelle. Dunod, Paris.

Breiman,L., Friedman,J., Olshen,R. & Stone,C. (1998) Classification And Regression Trees. Chapman & Hall/CRC, Boca Raton.

Thompson M.L, Reynolds J. Cox L.H 2001 A review of statistical methods for the meteorological adjustment of troposphetic ozone, 2001, *Atmospheric Environment,* **35**, 617-630

Blanchet G. 1990, Régimes météorologiques et diversité climatique dans l'espace rhonalpin, Revue de géographie de Lyon, 65, N°2 106-117

Blanchet G. 1994, chroniques climatologiques : le temps dans la région Rhône Alpes en 1990, Revue de Géographie de Lyon, 69, N°1, 89-103

Obled C., Bontron G, Garçon R.,2002, Quantitative precipitation forecasts : a statistoical adaption of modles outpuits through an analogues sorting approach. Atmopsheric research, 63,303-324p.

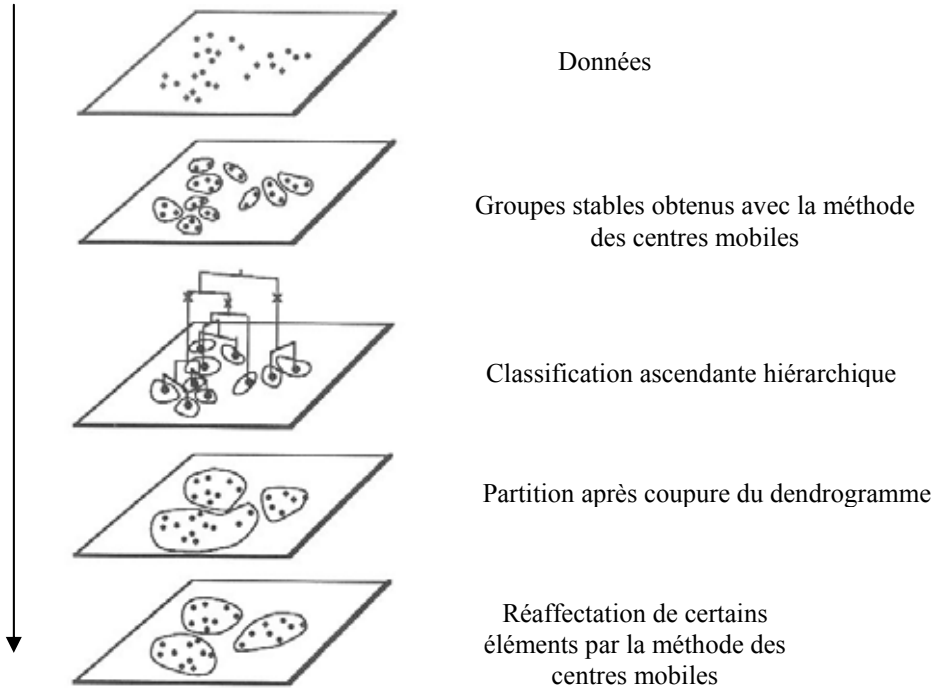
Piatyszek E. 1998, Détection de dysfonctionnements en système hydrographique. Application aux réseaux d'assainissement, Thèse de doctorat de l'Ecole des Mines de Saint-Etienne, juin 1998, N°ordre 187ID, 410 p.

Roth GD 2001, Guide de la météorologie , Lausane Delachaux et Niestlé cop.2001, Paris.

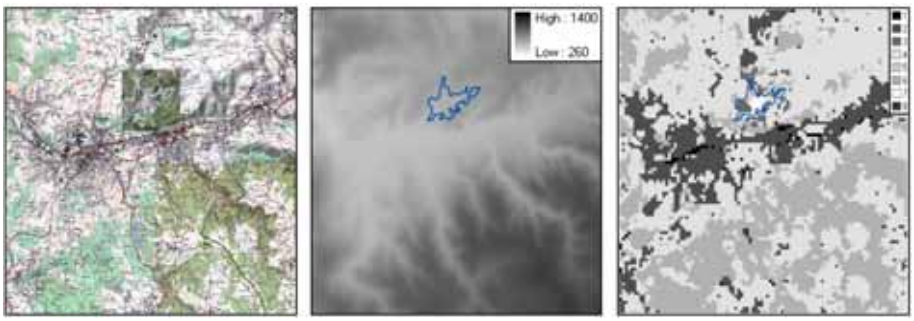
Annexe 2.1

Classification mixte d'après (Lebart et al . 1997):

- Méthode des centres mobiles :
- consolidation des centres et Obtention de groupes stables
- Utilisation de la classification hiérarchique ascendante sur les centres des groupes stables - selon le critère de Ward (1)
- Coupure de l'arbre
- Consolidation des classes en appliquant de nouveau la méthode des centres mobiles



Annexe 2.2 Orographie et occupation du sol utilisées par un code de CFD, ARPS autour d'un CSD



Annexe 2.3.
Analyse par Composantes Principales (ACP) : vérification du choix des variables

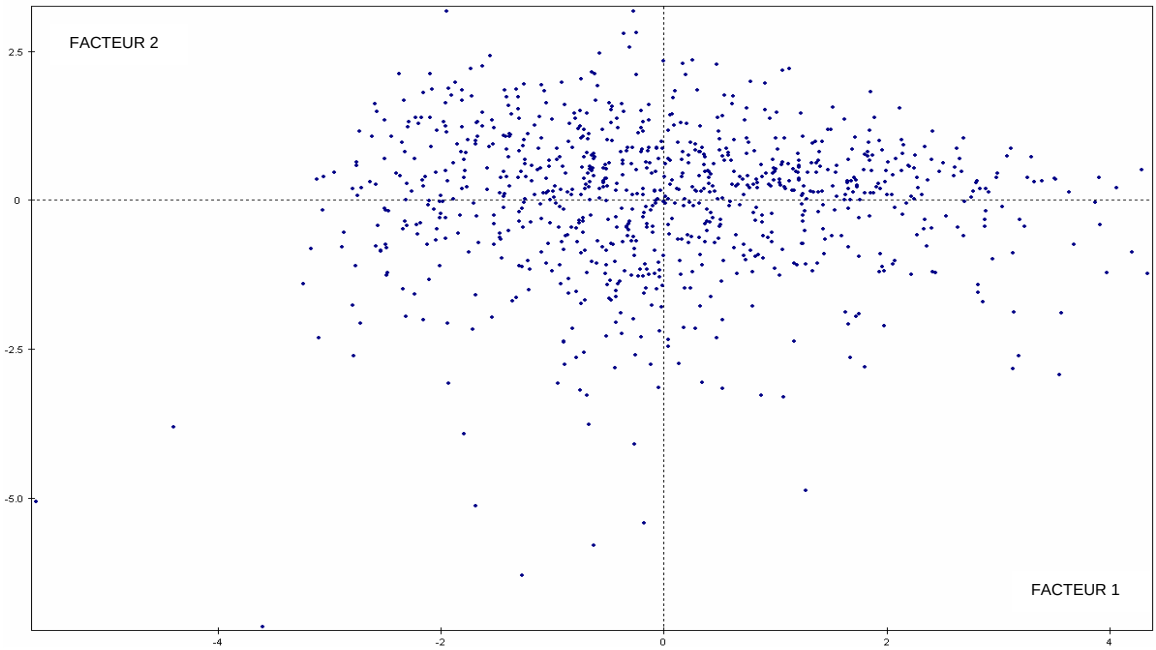


Figure 1. Nuage des individus suivant les axes factoriels 1 (26 % de l'information) et 2 (19 % de l'information)

	Température moyenne (°C)	Différence min/max (°C)	Pression moyenne (hPa)	Pluie cumulée (mm)	Humidité moyenne (%)	U moyen (m/s)	V moyen (m/s)	Instabilité du vent	Type d' instabilité
Température moyenne (°C)	1								
Différence min/max (°C)	0,49	1							
Pression moyenne (hPa)	0,02	0,26	1						
Pluie cumulée (mm)	-0,01	-0,22	-0,26	1					
Humidité moyenne (%)	-0,54	-0,65	-0,09	0,27	1				
U moyen (m/s)	0,09	0	-0,04	0,03	0,13	1			
V moyen (m/s)	0,03	0,05	-0,03	0	-0,02	0,07	1		
Instabilité du vent	0,13	0,1	-0,34	0,09	-0,33	-0,01	0	1	
Type d' instabilité	0,15	0,19	0,07	0,05	-0,05	0,01	0,01	-0,11	1

Tableau 1. Matrice des corrélations entre les variables
A noter : faible anti-corrélation de l'humidité à la température, et à la différence de température.

Annexe 2.4 Classification hiérarchique

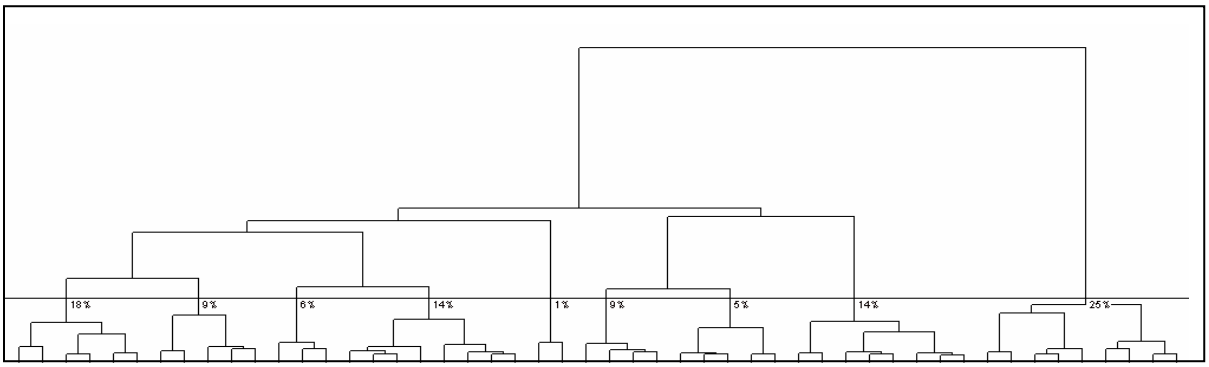


Figure 2. Arborescence de la classification hiérarchique obtenue pour les 9 classes des scénario

Agrégation autour des centres mobiles : répartition des individus par classes

Cette méthode itérative permet i) de trouver les barycentres optimaux de chacune des classes et ii) de regrouper les individus par classe en fonction de leur distance aux barycentres. On obtient la partition optimale des individus dans chaque classe (figure 2).

Principe de l'algorithme

[

Nombre de classes fixé ;

Le centre mobile de chaque classe est choisi aléatoirement dans l'espace des individus. ;

Itération sur :

- chaque individu est affecté à la classe dont le centre mobile est le plus proche ; les individus sont regroupés en classes en fonction de leurs distances aux centres mobiles
- recalcul des centres mobiles, et du barycentre de chaque nouvelle classe.

Jusqu'à ce que : deux itérations successives conduisent à la même partition.

]

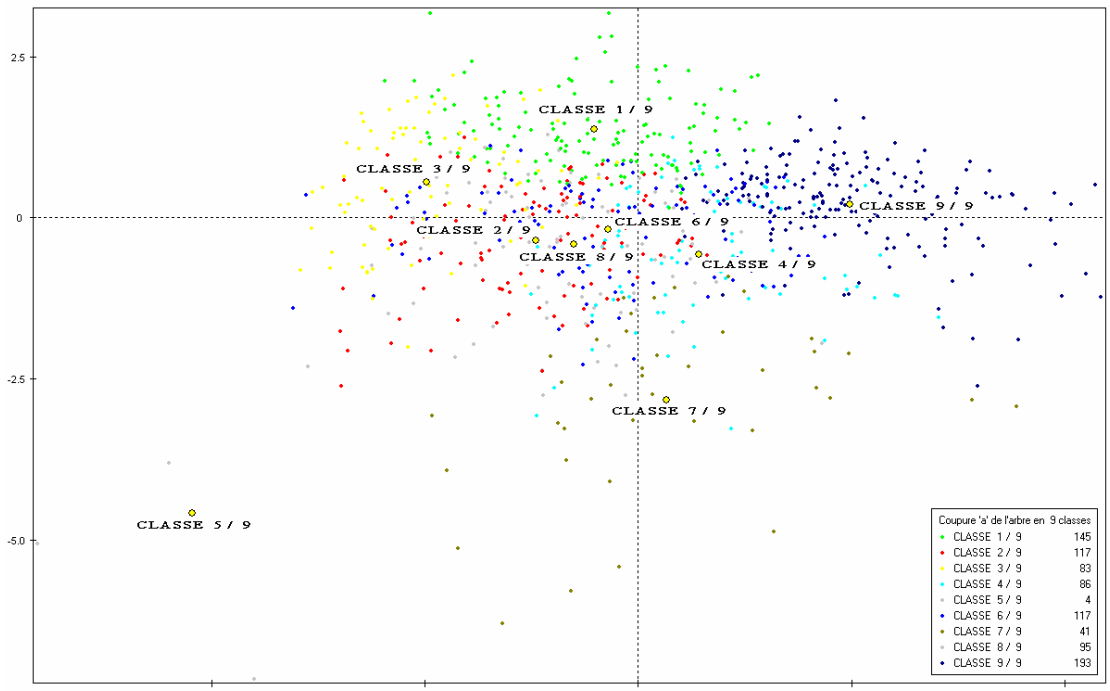
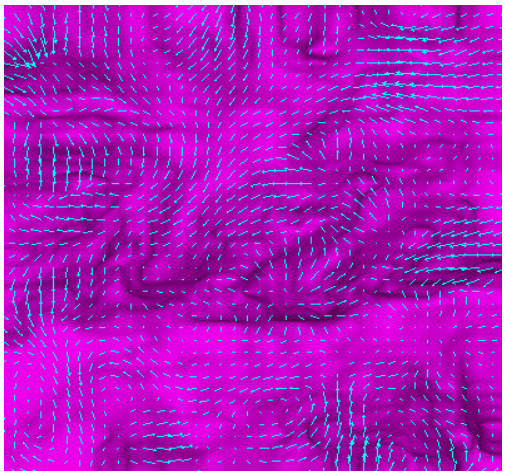


Figure 3. Barycentres des 9 classes et répartition des individus en 9 classes

Annexe 2.5 Résultats de la modélisation déterministe d'un scénario type : vent nul et profil adiabatique de 3,6°K/km



Champ des vents nocturnes pour un gradient de température potentielle de 3,6°K/km :
Résultats de la simulation à 6h- champ de vents 200 m au dessus du sol (Riesenmey 2004)

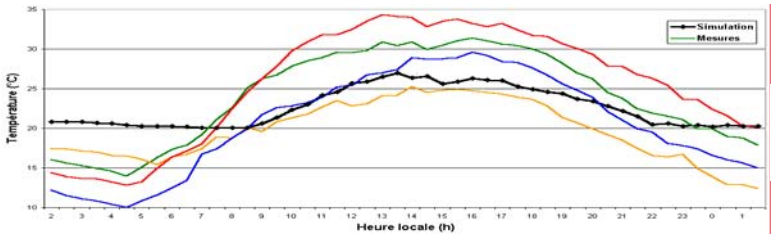


Figure 4 Comparaison entre les courbes de températures mesurées et calculées.
Les courbes lisses représentent les températures

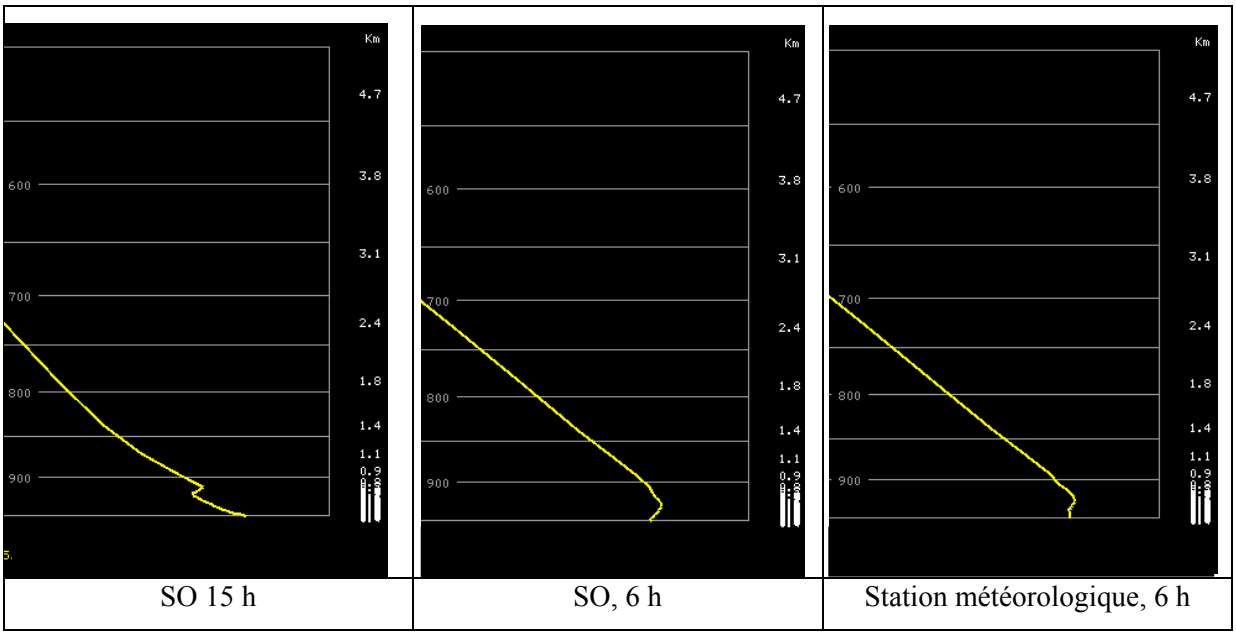
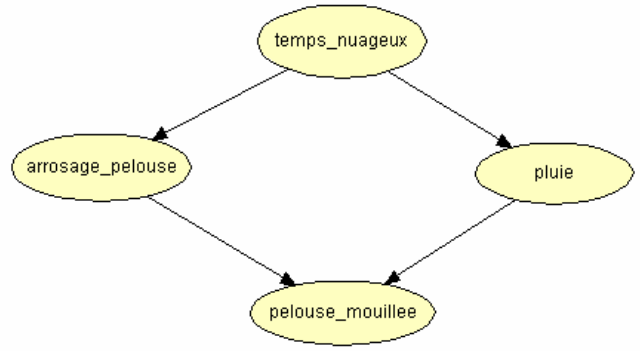


Figure 5. Inversions thermiques diurnes et nocturnes au SO et sur le site d'enfouissement

Annexe 2.6 : Quelques éléments sur les réseaux bayésiens

Un réseau bayésien est un modèle graphique dans lequel les connaissances sont représentées par une variable aléatoire. Chaque variable est un nœud du graphe et prend ses valeurs dans un ensemble discret ou continu. Le graphe est orienté et acyclique. Les arcs représentent un lien de dépendance directe (causalité). Un arc allant de A à B exprimera le fait que B dépend directement de A. L'absence d'arc ne renseigne que sur la non-existence d'une dépendance directe. Les paramètres expriment le poids donné à ces relations et sont les probabilités conditionnelles des variables connaissant la valeur des nœuds parents (p(B|A) ou les probabilités a priori si la variable n'a pas de parent.

L'exemple classique est : à partir d'un fait observé "*pelouse mouillée*", peut-on savoir s'il s'agit d'un *arrosage de la pelouse* ou s'*il a plu* :



Avec les tables de probabilité explicites suivantes:

Temps_nuageux	
Vrai	0.5
Faux	0.5

Pluie		
Temps_nuageux	Faux	Vrai
Faux	0.8	0.2
Vrai	0.2	0.8

Arrosage_pelouse		
Temps_nuageux	Faux	Vrai
Faux	0.5	0.9
Vrai	0.5	0.1

Pelouse_mouillée				
Arrosage_pelouse	Faux		Vrai	
Pluie	Faux	Vrai	Faux	Vrai
Faux	1	0.1	0.1	0.01
Vrai	0	0.9	0.9	0.99

Les probabilités des différents nœuds, suivent la forme générale théorème de Bayes :

P(arrosage_pelouse) = P(temps_nuageux) * P(arrosage_pelouse|temps_nuageux)
+ (1-P(temps_nuageux)) * P(arrosage_pelouse|pas temps_nuageux)

Le problème à résoudre est celui des inférences probabilistes : si l'on considère le nœud arrosage_pelouse et si l'on suppose que l'état observé de pelouse_mouillée est vrai ; il y a deux causes possibles à cela, il pleut ou on arrose la pelouse.

On peut ainsi définir la probabilité suivante :

$$P(ap = 1 | pm = 1) = \frac{P(ap = 1, pm = 1)}{P(pm = 1)}$$

où *ap représente l'événement arrosage_pelouse*
pm représente l'événement pelouse_mouillée.

Il est aisé de trouver P(pm=1) et P(ap=1,pm=1) en fonction des tables de probabilités déjà fournies.

Quelques précisions sur les modèles probabilistes :

Dans le cas d'un réseau bayésien, le raisonnement se fait sur des variables discrètes et non continues. Nous ne disposons pas de tables de probabilités mais de tables d'expériences où un certain nombre d'observations sont effectuées à partir d'expériences, par exemple :

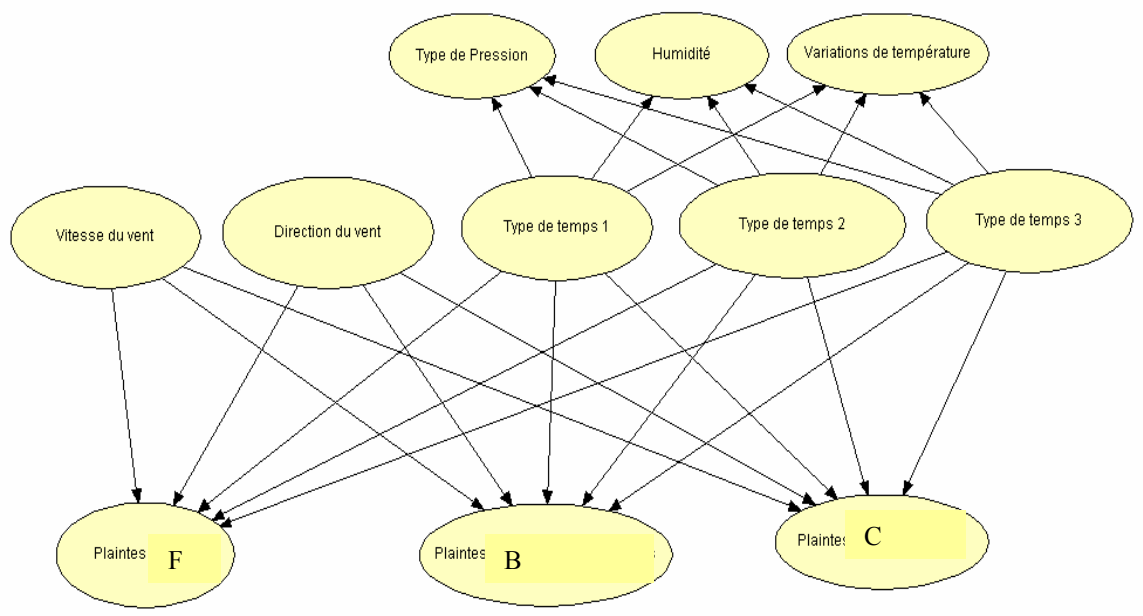
N° Expérience	**Temps nuageux**	**Arrosage pelouse**	**Pluie**	**Pelouse mouillée**
1	Vrai	Faux	Vrai	Vrai
2	Faux	Vrai	Faux	Vrai
3	Vrai	Vrai	Vrai	Vrai
4	Vrai	Faux	Faux	Faux
Etc…				

Il s'agit d'un problème d'apprentissage, comment définir les paramètres du réseau (probabilités conditionnelles et probabilité sur le temps nuageux) à partir des observations détenues ?
Cet apprentissage peut se faire par l'algorithme EM (*maximum de vraisemblance*) qui permet de calculer les paramètres du réseau bayésien à partir d'observations plus ou moins complètes avec des variables observées V et des variables inconnues H.
L'algorithme EM est de la forme :

[Initialisation des probabilités P(H|V)
Etape E : on utilise ces probabilités pour remplacer les données manquantes
Etape M : on utilise les variables connues et celles estimées pour calculer les nouveaux paramètres du réseau (P(H|V))
On itère jusqu'à convergence]

NB : Les tests ont été réalisés avec le logiciel Hugin, qui utilise l'algorithme EM pour un apprentissage des paramètres du réseau.

Annexe 2.7 Réseau bayésien identifiable du risque de plainte en aval



Réseau bayésien identifiable du risque de plainte en aval

Volet 3 : Approche Heuristique de la composante espace - temps de l'information environnementale : éléments - justifications - Recherches futures

Ce volet est consacré à la mise en perspectives des éléments d'analyse précédemment établis au volet 2, concernant la construction de Connaissance complémentaire et indirectement attribuable à un processus, et plus généralement à l'état d'un système. L'heuristique proposée concerne tout processus continu issu du domaine de l'Eau ou de l'Atmosphère.

L'objectif est de fournir les éléments d'analyse nécessaires aux développements des prochains travaux de recherche et de proposer explicitement des éléments pertinents et des pistes possibles de développement futur. Il s'appuie sur une bibliographie choisie en fonction des préoccupations abordées au travers de 4 points particuliers que sont :

1. la Reconnaissance de Forme dans la donnée environnementale et tout particulièrement les méthodes de clustering et classification (*data mining , pattern recognition*)
2. les Vecteurs d'état établis à partir du clustering et de la simulation déterministe (*système dynamique)*
3. les Multi-modèles et la réduction d'un modèle de type EDP à un système dynamique (EDO)
4. le Prédicteur d'état (*data mining , traitement de données*)

Cette partie du travail est plus qu'une perspective, elle constitue une prospective sans en être la réalisation. Pour cette raison, elle ne peut faire l'objet d'un inventaire détaillé et chronologique des thématiques et domaines d'application des dix prochaines années, tel un programme de recherche détaillé, que je pourrais explicitement décrire. De plus, les préoccupations environnementales évoluent rapidement en fonction des demandes des collectivités, de la réglementation ainsi que celles des industriels. L'essentiel est de proposer un programme méthodologique, ouvert, adaptatif, mais dont les questions de recherche resteront certainement ouvertes pour quelques années encore.

1. Introduction

Face aux problèmes, divers, nombreux et pluridisciplinaires de l'optimisation et de la prédiction des impacts environnementaux, rappelons d'abord, l'objet de notre préoccupation puis, l'idée majeure développée dans chacun des 2 volets précédents : (ii et iii)

i. Un phénomène physique évoluant dans un environnement naturel fortement "*anthropisé*" peut être à l'origine d'effets, d'impacts ou de risques éventuels. Il s'agit de fournir la meilleure information disponible pour l'aide à la décision quantifiable sous forme de valeurs scalaires, des cartes géoréférencées ou de toute autre donnée transcrite et adaptée à l'acteur concerné. Cette information doit caractériser un processus ou phénomène continu (mouvement atmosphérique, transport d'un polluant, écoulement hydrodynamique - hydraulique...).

ii. Le volet 1 s'inscrit dans le développement de la simulation déterministe intégrant différentes échelles de temps et d'espace pour l'évaluation de flux hydrauliques et hydrodynamiques autour de sites postindustriels. Il pose par le biais de la modélisation le problème de la juxtaposition de données de natures diverses, d'échelles et de processus physiques complémentaires mais qui ne sont pas systématiquement i*nter opérables*[31] dans les modèles. Les premiers éléments issus des dix années de recherche passées au centre Site, fournissent l'argumentaire, la méthode de structuration mais aussi la façon d'accroître la connaissance d'un processus. Reste à savoir quels sont les moyens mobilisables et pouvant être mis en œuvre ?

iii. Le volet 2 propose un schéma de raisonnement pour les classes de problème en Environnement (figure 2.0.1) : l'application "ingénierie" traite ce phénomène physique continu pour caractériser un état du système, prévoir ou diagnostiquer une (ou plusieurs) caractéristique(s) du processus sachant qu'il est délicat de l'obtenir par un schéma classique de modélisation déterministe. Deux postulats ont été établis :

- soit, la connaissance est insuffisante pour procéder à la modélisation :

{IL n'existe pas de modélisation fiable ET IL existe des Données/observations ayant certaines propriétés ALORS On peut générer de l'information par [analyse de la Donnée $\cup$ interpolation/estimation]}

- soit la modélisation est coûteuse : en données, en temps de calcul, en terme de représentation de phénomènes compliqués (cas de l'atmosphère), et génère un faible gain du rapport résultats/investissement :

{IL existe une modélisation déterministe ET IL existe dans les données une information qui caractérise l'Etat du système ALORS on peut accroître la connaissance de certains états du système par [simulation récurrente $\cup$ data mining] }

Les résultats obtenus et les perspectives de recherche dégagées dans ce volet permettent de conserver les hypothèses proposées et postulats établis dans l'analyse (préface du volet 2).

[31] Au sens interopérabilité

Pour chacun des volets, des applications en Ingénierie de l'Environnement ont été présentées : la maintenance de niveaux d'alimentation en eau de gravières, la détection des tronçons de réseau d'AEP vulnérables à la rupture, la mise en place de consignes d'exploitation d'un CSD pour limiter la pollution olfactive….

A chaque classe de problème, la démarche scientifique a été fondée sur :

- l'identification de la question thématique *e.g.* flux hydrauliques, détection de tronçons ayant un mauvais fonctionnement, à la détection des situations météorologiques à risque de nuisance olfactive…,
- le choix d'un cadre formel approprié à la modélisation des connaissances a priori, avec la formulation mathématique et la mise en œuvre d'algorithmes de résolution,
- l'analyse des résultats et opérationnalité auprès d'un utilisateur.

Chaque thème eau - air – déchets, se décline en classes de problèmes dont certains déjà traités dans les deux volets seront poursuivis dans un futur proche. On peut citer le dysfonctionnement et le vieillissement des réseaux urbains, la quantification des émissions pour une source surfacique industrielle, l'évaluation des expositions instantanées (seuil d'alerte à un polluant gazeux) ou a moyen terme (risque sanitaire)…

Ces travaux effectués se situent à la frontière de plusieurs disciplines : Geosciences[32] - Ingénierie de l'Environnement - Informatique - Mathématiques appliquées.

Cette dualité souvent exposée entre, le "tout par la Physique" établi par une série d'équations différentielles, et la conviction qu'il existe dans tout jeu de données l'expression d'un modèle intrinsèque, formalisé par modèle de représentation, m'ont convaincue d'articuler mes travaux de recherche autour de cette notion d'identification de connaissance complémentaire d'un processus physique, dynamique agissant en tout point de l'espace et évoluant dans le temps par lequel on cherche à extraire l'information pertinente non explicite à priori. Cette information concerne des propriétés intrinsèques au système, caractérisant son état et son comportement et qui sont nécessaires pour l'aide à la décision. Cette approche constitue une alternative à la modélisation physique lorsque les données d'entrée sont insuffisantes.

Je pourrais choisir de décliner en perspective d'autres problématiques environnementales (quelques unes seront citées au paragraphe 6). Ceci ne serait pas très judicieux car la liste risquerait d'être un peu longue, fastidieuse, et fournirait peu d'information sur la priorité de ces applications dans un futur à échéance de 10 - 15 ans, que ce soit pour un processus industriel ou pour un territoire.
J'aurais pu dire que l'on va poursuivre l'approche - simulation - diagnostic - prévision et identification mais cela aurait été une évidence puisque l'équipe de recherche à laquelle j'appartiens oeuvre pour cela… (en collaboration avec d'autres équipes de recherche !)

On peut admettre que :

Quelque soit le domaine {Eau - l'Air - les Déchets}, il existe une constance dans les données manipulées : elles sont instanciées dans le temps et dans l'espace, qualitatives ou quantitatives, souvent peu ou faiblement disponibles, *à trou* et erronées.

[32] au sens anglophone, terme qui regroupe l'ensemble des disciplines liés à la Terre comme la météorologie, l'hydrologie, l'hydrodynamique … alors qu'en France, il est restreint aux Sciences de la Terre et à la Géophysique.

Pour cette raison, j'ai choisi que ce rapport revête une dimension particulière en proposant une approche spécifique, une *heuristique* de la composante *espace-temps* de l'information environnementale. Pourquoi une heuristique[33] ? Parce que, la démarche proposée, intègre des démarches, des méthodes et des outils qui favorisent la Connaissance scientifique des phénomènes dans le domaine de l'Environnement.

Les 4 enjeux de cette approche[34] sont de :

- limiter la modélisation et la simulation numérique par CFD car coûteuse,
- favoriser ce type de simulation pour obtenir une donnée pertinente,
- organiser la donnée pour permettre l'interopérabilité entre outils différents,
- proposer une alternative à ce type d'outil numérique pour la prévision.

Cette heuristique s'articule autour de 4 points méthodologiques :

1. Reconnaissance de Forme dans la donnée environnementale et tout particulièrement les méthodes de clustering et de classification (*data mining , pattern recognition*)
2. Vecteur d'état à partir du clustering et de la simulation déterministe (*système dynamique)*
3. Multi-modèle et réduction d'un modèle type EDP à un système dynamique (EDO)
4. Prédicteur d'état (*data mining , traitement de données*)

Pour chaque point méthodologique, seront abordés successivement : le pourquoi de la question et sa formulation, l'existant méthodologique des domaines qui s'y rattachent, une proposition si possible illustrée, enfin les perspectives de développement et les résultats attendus.

L'objectif de ces points est à terme de proposer une chaîne de traitement destinée à réduire, transformer et traiter des données environnementales pour laquelle un modèle conceptuel devra être établi et qui permettrait de disposer de bases de données structurées. Enfin, les aspects fondamentaux de ces 4 points de recherche seront illustrés par différentes problématiques en Sciences de l'Environnement (§6.).

Ce volet dépasse le stade de l'intention et/ou des perspectives car un formalisme est proposé et analysé pour chacun des points; mais cela ne constitue pas encore une conclusion car tout reste à faire (développer, tester, valider et effectuer une rétroaction).

"Les limites n'existent que dans l'esprit"
Elsa Triolet

[33] Heuristique : terme méthodologique qui qualifie tous outils intellectuels, tous procédés, toutes démarches favorisant la découverte ; on s'intéresse plutôt ici à l'heuristique qui s'intéresse aux conditions de justification des connaissances. Il s'agit d'un approche qui intègre plusieurs composantes et non une théorie en particulier.
[34] ce qui constitue les 3 arguments initiaux de l'analyse du volet2

2. Reconnaissance de Forme de la composante *espace - temps* de la donnée environnementale : Méthodes de clustering et classification[35]

2.1 Scénario météorologique[36]

Chaque évènement de pollution atmosphérique ou plus simplement un certain état de l'atmosphère d'un point de vue température, humidité, vitesse du vent et concentration moyenne (CH4, ozone) ou nuisance olfactive est associé à des situations météorologiques spécifiques devant être identifiées avant d'être calculés par un CFD[37]. (Volet 2 chap. 2 .1.2)

L'identification de ces scenarii météorologiques se base sur l'hypothèse suivante: il existerait une relation univoque entre une situation météorologique sur un jour (a posteriori sur 3-5 jours) et les données disponibles, mesures de l'état de l'atmosphère et notamment de la concentration en ozone.

Par conséquent il serait possible de caractériser un jour ou une série de jours consécutifs par un ensemble de variables pertinentes qui permettrait de distinguer des familles de jours caractéristiques du système météorologique et de constituer alors des classes appelées "scénario météorologique". Dans un premier temps l'objet jour et les classes "scénario météorologique" sont décrits par des variables discrètes moyennes (volet 2.).

2.2 Quelques éléments sur la classification et la segmentation

La Classification

Soit la représentation d'un objet par un vecteur de caractéristiques $X=[x_1, x_2 \ldots x_d]^T$. Les k vecteurs qui représentent l'ensemble des objets peuvent être positionnés dans un espace Euclidien R^d, où ils correspondent chacun à un point. Ceux ci-peuvent être regroupés en amas, chacun de ces amas est associé à une classe particulière.
Le rôle d'un classificateur est de déterminer parmi un ensemble fini de classes, celle à laquelle appartient un objet donné. Il doit être capable de modéliser au mieux les frontières qui séparent les classes, par une fonction discriminante qui permet d'exprimer le critère de classification de la façon suivante : assigner à la classe ω_i l'objet X_i si et seulement si la valeur de la fonction discriminante de la classe ω_i est supérieure à celle de la fonction discriminante de n'importe qu'elle autre classe ω_j .

$$X \in \omega i \Leftrightarrow \Phi_i(X) \geq \Phi j(X) \forall\, j=1,2,3 \ldots C; j \neq j.$$

$\Phi_i(X)$ Fonction discriminante de la classe ω_i
C C nombre de classes total

Le classificateur optimal, bayésien est celui qui minimise la probabilité d'erreur et ainsi maximise la probabilité a posteriori $p(\omega_i|X)$; le critère optimal de la classification s'exprime par :

[35] l'application du data mining et pattern recognition à la données environnementale
[36] le scénario météorologique est un objet que la classification cherche à affecter à *1* ou *n* classes de scenarii
[37] CFD : Computational Fluid Dynamics : code de calcul numérique de résolution des équations de l'atmosphère par discrétisation par volume fini et/ou éléments finis

$$X \in \omega i \Leftrightarrow p\left(\omega i|X\right) > p\left(\omega j|X\right) \forall j=1,2,3...C; i \neq j.$$

avec
$p(X|\omega_i)$ la fonction de densité de probabilité d'observer X étant donnée la classe ω_i
$p(\omega_i)$ la probabilité à priori de la classe ω_I
$p(\omega_i|X)$ la probabilité à posteriori que la classe correcte soit ω_i lorsque l'on observe X

Le calcul exact des probabilités a posteriori est un problème complexe : des modèles sous optimaux de classificateurs ont été développés qui utilisent soit des fonctions discriminantes autres que la probabilité à posteriori, soit des fonctions qui bornent la probabilité d'erreur liée à la règle de décision du classificateur comme celui du Plus proche Voisin (Cover et al. 1967, Duda et al. 1973).

Plus généralement, il s'agit d'un modèle de prédiction à partir d'un échantillon d'apprentissage A={ $(\mathbf{x}_k, y_k)$, $k=1,..n$}, où la sortie de l'algorithme d'apprentissage est une fonction prédictive y = $F(\mathbf{x})$ qui minimise un critère d'erreur quadratique $E(F,A)$ entre la réponse y_k et la valeur $F(\mathbf{x})$:

$$E(F,A)=\sum_{k=1}^{n}\left(y_k - F(x_k)\right)^2$$

Eléments sur la segmentation[38] et l'apprentissage non supervisé[39]

On s'intéresse en priorité au découpage d'un ensemble d'objets : ici des jours météorologiques en plusieurs catégories de façon à maximiser la similarité intra-groupe et à minimiser la similarité entre groupes.

Des méthodes les plus usuelles comme la classification hiérarchique (ascendante ou descente), l'agrégation sur centres mobiles, aux plus récentes le clustering par Kmeans (Kmoyennes), puis le Fuzzy C-means (version flou de Kmeans), toutes ont en commun les phases suivantes :

- la construction des similarités ou dissimilarités entre l'espace des objets,
- le choix d'une fonction objectif (critère de clustering),
- la détermination de l'algorithme d'optimisation.

De nombreux travaux ont permis de relâcher la "rigidité" des méthodes d'assignation booléenne d'un objet à une classe par la notion de degré d'appartenance d'un objet à une classe par ensembles flous (Bezdek J.C 1974, 1981) et communément admise sous le nom de techniques *fuzzy C means*[40] pour lesquelles demeure une lacune de justification théorique selon (Coppi 2005). Les travaux récents cherchent à introduire un terme de régulation juxtaposé au critère de maximum d'homogénéité intraclasse (Miyamoto et Mukaidono 1997). Cette fonction de régulation cherche à mesurer le flou de la partition selon un schéma d'entropie floue (fuzzy entropy).

[38] on préférera par la suite adopter le mot anglophone de clustering pour désigner une méthode qui ne fixe pas a priori le nombre de classes contrairement aux méthodes de classification (*senso stricto),* ce qui permet de découvrir les groupes (clusters) d'individus similaires en se basant sur la similarité entre individus : car il s'agit a priori d'apprentissage non supervisé car on ne connaît pas a priori les classes, il s'agit de les discriminer

[39] il s'agit d'éléments de bibliographie : toutes ces techniques n'ont pas encore été développées dans nos travaux, car il s'agit d'une prospective avant leurs développements dans les travaux futurs

[40] version floue de l'algorithme des k-moyennes

2.3 Amélioration entre Scénario *jour* moyen et scénario *jour* temporel : utopie ou admissibilité ?

A partir de l'objet jour météorologique décrit par des variables utilisant les mesures de l'état de l'atmosphère, il s'agit d'obtenir une méthode de clustering qui permet dans un premier temps de proposer des classes (dénombrement et caractéristiques moyennes du barycentre et des écarts au barycentre) qui ont un sens physique en météorologie. Dans un second temps, il faut prévoir l'appartenance booléenne (ou un degré d'appartenance) d'un jour. Le mode prédictif est prématuré pour l'instant pour la raison suivante : nous cherchons à identifier au mieux des classes type, les scenarii météorologiques ; ils doivent permettre de choisir un jour dans l'échantillon disponible qui sera simulé et modélisé par un code météorologique à l'échelle locale (cf. volet 2 chapitre 2. §2.)

Or les résultats obtenus montrent que :

- l'introduction d'autres variables modifie la classification obtenue en augmentant la dimension de l'espace,
- le choix des variables explicatives modifie la classification.

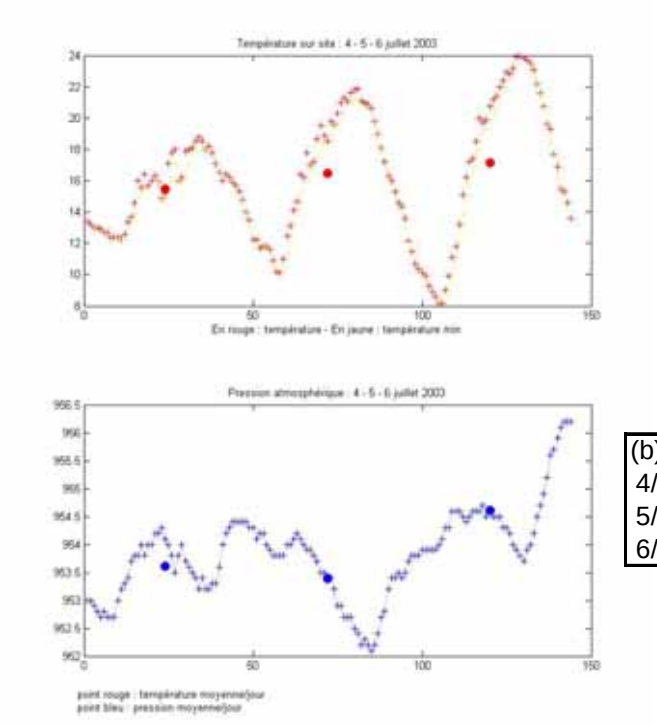
Ces deux constats sont classiques et peu de méthodes permettent de trancher. Améliore- t- on la classification en choisissant plutôt le gradient thermique diurne[41] ou l'abaissement de la température nocturne, Dt_{noc} [42] que l'accroissement diurne de la température Dt_{di} ?

Quelle que soit la méthode de clustering retenue, un élément prépondérant est la mesure de la dissimilarité entre deux individus (ici des jours) ; elle est évaluée par une distance euclidienne ou de Mahalanobis pour les *d* valeurs réelles des k vecteurs de caractéristiques X (échantillon de dimension k).
Cette distance dépend des *d* valeurs, dans R^d et dont a fortiori des *d* variables explicatives choisies précédents. Bien que l'ACP permet de limiter la redondance d'information en réduisant le nombre des variables corrélées linéairement, elle ne dit rien sur la pertinence de la *d* variable retenue. Or il ne s'agit pas réellement d'une donnée multivariée au sens ou l'on représente un jour météorologique par des valeurs scalaires moyennes or un jour est décrit par des mesures continues (T°, vent…) régulièrement échantillonnées. Enfin on assimile un scénario météorologique à un jour, même s'il a été montré qu'un type de temps est représenté par une séquence de jours (2 à 5 jours) (volet 2 chapitre 2.)

La question devient alors : comment passer d'une classe type scenarii jour à un scénario jour intégrant la dimension temps ?

Illustrons notre propos :
Prenons les relevés disponibles sur une station comme la température et la pression, on choisit dans un premier temps de représenter un jour par une valeur moyenne scalaire des températures, une pression moyenne … avec un vecteur $X=[x_1, x_2 \ldots x_d]^T$ et avec un espace Euclidien R^9, soit *k* vecteurs des objets (soient 365 objets /an).

[41] $T_{dif} = \max(T_i) - \min(T_i)$, écart entre le minimum et le maximum mesurés sur la journée
[42] Dt_{noc} = écart entre le Tmax (12 et 16 h00 en TU) la veille et la température nocturne minimale relevée entre 3h00 et 6h00 (TU)le jour même



(b)	Tmoy	Pression	T max	Tmin	Tdif
4/7	15,49	953,61	18,80	12,30	6,50
5/7	16,46	953,39	21,90	10,10	11,80
6/7	17,17	974,45	24,00	8,20	15,80

Figure 3.2.1 : (a) Données météorologiques : séries chronologiques (b) Données moyennes scalaires

La dimension 2 du vecteur *jour*, lisse les caractéristiques et les signatures temporelles de ces 3 objets[43] (figure 3.2.1.b). L'espace de représentation des objets est celui adopté par le clustering (figure 3.2.2).

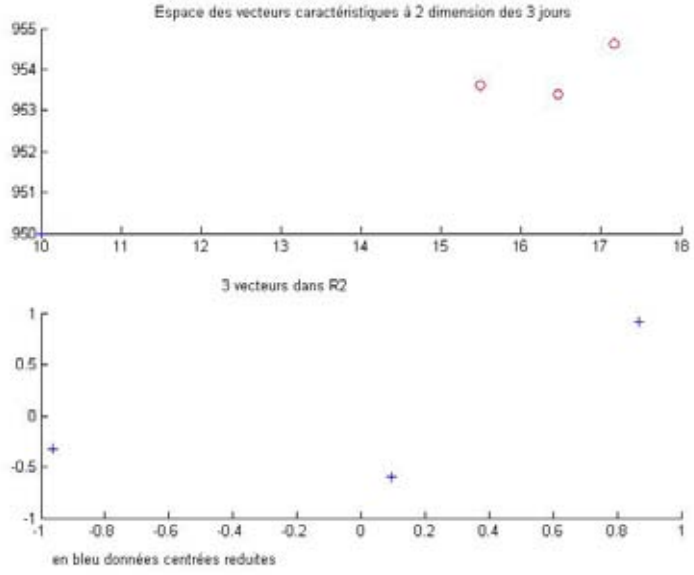


Figure 3.2.2 Espace des 3 vecteurs caractéristiques pour 3 jours dans leur espace caractéristique

[43] il s'agit d'un exemple 'trivial' pour illustrer le raisonnement et non d'un exemple représentatif d'un groupe important de jours (> 3*365 jours)

En réalité la mesure est une donnée continue dans le temps, échantillonnée à un pas de temps (figure 3.2.1 a. série chronologique). La question devient alors : quel mode de représentation de cette donnée est compatible avec le clustering, peut-on l'adopter ?

Deux approches sont possibles :

i) Le vecteur *jour* change de dimension mais reste constant pour tous les objets *i*

1ère solution :

On intègre une variable (indépendante ou non) supplémentaire à chaque pas horaire *i* d'un paramètre. Pour l'énergie T(i) on peut découper un attribut en attributs élémentaires comme le vecteur $X=[x_1, x_2 \ldots x_d]^T$ en x_i valeurs de la température ou valeur de pression à i …, puis on juxtapose les **p** paramètres, et on obtient alors $d = (24\times\mathbf{p})$ variables pour un pas de 1 heure et **p** paramètres.
Dans ce cas, on augmente considérablement la dimension de l'espace de représentation : une ACP permet d'identifier si une réduction est possible avant d'appliquer l'évaluation de la similitude entre les jours.

Points positifs[44] : les méthodes de clustering restent valables

Points négatifs[45] *:* il y a augmentation de l'espace de représentation ; la discrimination des objets n'est pas forcement meilleure; il reste l'impact du choix des variables sur les classes.

2ème solution : Une forme est représentée par une séquence temporelle de vecteurs caractéristiques.

Une représentation possible de la composante temporelle est la matrice à 3 dimensions[46] : *nombre d'objets × Nombre de variable quantitative × time*[47]
Le vecteur de cette donnée temporelle devient :

$$X \equiv \{ x_{ijt} : i = 1, N ; j = 1, M; t = 1, T\}$$

Dans cet espace, la normalisation appliquée sur les données doit prendre en compte l'intégration de la dimension temporelle sous peine de faire ressortir une distance fantaisiste. Le pré traitement proposé par (Coppi et al. 2005 et D'Urso 2004) propose de centrer, normaliser et standardiser les données en fonction du temps mais surtout de conserver un tableau de donnée temps **X**, dans un espace vectoriel de donnée $\Re^{M+1}$, appelé espace des unités, où chaque objet i est représenté par un vecteur "moyenne" à chaque temps t :

$$\mathbf{y}_{it} = (x_{i1t} \ldots x_{ijt} \ldots x_{iMt}, t)', \ i = 1, , N ; \ t = 1, T$$

L'ensemble des $N_T(i) \equiv \{\mathbf{y}_{it} : t = 1, T\}$ avec i=1, N représente l'ensemble des trajectoires temps des objets . Toutes ces trajectoires coupent le *T* hyper plan parallèlement à $\Re^M$ (D'Urso 2000).

[44] points méthodologiques résolus et intéressants dans ce type de modélisation de la donnée (le plus)
[45] points méthodologiques non résolus ou non adaptés de ce type de modélisation de la donnée (le moins)
[46] l'usage de ce type de données est courant dans des logiciels comme Matlab avec multidimensional arrays & multidimensional structure arrays.
[47] Exemple : *365 jours × 9 variables × 48 (échantillon régulier de 30 mn pour 24 heures)*

Conjointement, (Coppi et al. 2005) définit plusieurs types de dissimilarités :

- la distance instantanée[48] entre 2 objets *i* et *i'* à *t* donné
- la dissimilarité entre 2 trajectoires temporelles de *i* et *i'*(a chaque t = 1, T*), soit :*

$$ {}_1d^2_{ii'} = \sum_{t=1}^{T} d^2_{1w}(\mathbf{x}_{it},\mathbf{x}_{i't}) $$

$$ = \sum_{t=1}^{T} ({}_1w_t \left\| \mathbf{x}_{it} - \mathbf{x}_{i't} \right\|)^2. $$

- la dissimilarité "variationnelle" des trajectoires, assimilables à une mesure de vitesse de variation (gradient) .

$$ {}_2d^2_{ii'} = \sum_{t=2}^{T} d^2_{2w}(\mathbf{v}_{it},\mathbf{v}_{i't}) $$

$$ = \sum_{t=2}^{T} ({}_2w_t \left\| \mathbf{v}_{it} - \mathbf{v}_{i't} \right\|)^2 $$

$$ = \sum_{t=2}^{T} ({}_2w_t \left\| (\mathbf{x}_{it} - \mathbf{x}_{it-1}) - (\mathbf{x}_{i't} - \mathbf{x}_{i't-1}) \right\|)^2 $$

Ces 3 dissimilarités des trajectoires multivariées recherchent la proximité longitudinale (3ème dimension de la matrice) et une proximité instantanée avec les 2 premières dimensions. Ceci justifie le développement d'un modèle flou de clustering qui classifie des trajectoires temps avec un critère d'entropie où la fonction à minimiser intègre une minimisation de la distance (C-means flou) et maximise l'entropie (fonction de déviation intra classe) (Coppi 2005). Les résultats avancés par ces auteurs montrent que le suivi et le nombre de reproduction des formes temporelles permettraient d'identifier des groupes pour lesquels le degré d'appartenance d'un modèle de clustering flou reste faible.

Perspectives d'utilisation et valeur générale de la méthode proposée

Cette approche permettrait de calculer une nouvelle distance pouvant être intégrée dans diverses méthodes de clustering. Il s'agirait alors de voir si l'intégration d'une dimension temporelle des séries de mesure sur la température, le vent, la pression et l'humidité distinguerait plus ou moins de classes, ou des scenarii différents.
La seconde étape consisterait à développer une méthode de calcul de l'entropie floue pour les scenarii météorologiques ; il s'agira alors de pouvoir évaluer le gain obtenu par ces méthodes et le rapport entre robustesse et sensibilité du nombre de classes et de l'affectation des jours à ces différentes classes.

Points à résoudre

- l'utilisation de ce type de mesure de dissimilarité longitudinale (dans le temps) et transversale (instantanée entre 2 objets) pour des méthodes classiques de k-moyennes, classification hiérarchique, en adaptant le calcul des distances,
- la justification des classes obtenues par rapport aux situations météorologiques locales et leur validation,
- enfin, l'apport de la prise en compte du temps se justifie-t-elle par rapport à la complexité des prés traitement des données et l'adaptation des algorithmes au clustering ?

[48] Egale à la distance euclidienne

ii) le vecteur *jour* change de dimension qui n'est pas constante pour tous les objets *i*

Dans ce cas, le vecteur de caractéristiques de longueur T n'est plus constant pour l'ensemble des vecteurs, il est nécessaire de pourvoir tenir compte de cette variabilité (Gosselin 1995).
En quoi notre scénario jour est - il concerné: en réalité le critère permettant d'associer un scénario à un jour est calé sur la périodicité des températures. Cependant, la pression atmosphérique varie plus lentement et n'est pas caractérisée par l'oscillation locale + ou – journalière mais sur une tendance qui est en fait une onde de plus grande longueur (2-3 jour) qui correspond au passage de la frontogenèse. Dans ce cas, une séquence de 2-3 jours serait caractéristique d'une situation type intégrant le type de temps. La fenêtre d'observation du type de temps peut varier de 1 à 5 jours.
Soit la longueur de la séquence est supérieure à 1 jour et de longueur constante : on conserve la méthode du point i).
Soit la longueur de la séquence est supérieure à 1 jour et de longueur variable, on tente de comparer des séquences de temps variables. Les méthodes de comparaison dynamique (Dynamic Time Warping) permettent d'évaluer la dissemblance entre une séquence de longueur I et une séquence de référence de longueur R (avec $I \neq R$) en recherchant un alignement temporel linéaire ou non qui permet de minimiser la distorsion. Bien qu'intéressantes, l'applicabilité de ces méthodes ne doit pas aboutir à une interpolation ou à une extrapolation abusive de valeurs.

Le regroupement de courbes comme le propose l'analyse fonctionnelle introduite par (Ramsay et Silvermann 1997) admet de considérer les observations de la courbe comme une unité et non comme une analyse multivariée. On cherche alors à estimer la courbe par une fonction *f* à partir des points échantillonnés puis à calculer sa dérivée et son intégrale. Cette transformation de la courbe (par une fonction de base de type polynomial, en spline, en séries de Fourier..) s'effectue avant de calculer sa distance à d'autres courbes. Il reste alors à appliquer un classificateur de courbes par regroupement flou de type soustraction (Quach R.P 2002).

Des résultats obtenus au point i) ainsi que l'analyse des enchaînement des scenarii par d'autres méthodes (§4.) conditionneront l'exploration éventuelle de ce type de traitement.

2.4 Passage d'un scénario jour moyen à un scénario météorologique *espace*

2.4.1 Positionnement de la simulation déterministe

Partant du constat suivant : on décrit une situation météorologique par une classe représentant le scénario météorologique. Sa dimension temporelle représentative est un jour décrit par des variables moyennes et/ou caractéristiques ; ceci semble vraisemblable compte tenu des résultats précédemment obtenus. La question est classique : comment obtenir sur un plan 2D donné, la répartition des différents paramètres météorologiques climatiques (isothermes, isobares, iso concentrations…). Pour une situation moyenne du champ scalaire d'une variable et à condition de disposer de suffisamment de relevés, les méthodes d'interpolation et/ou d'estimateurs probabilistes sont probantes. Lorsque l'on décide d'obtenir un champs vectoriel (vent) en l'absence de données au sol suffisantes (au sens d'un semis de points relativement dense et homogène) et que l'on est capable de fournir des conditions aux limites cohérentes, la modélisation déterministe par EDP (Annexe 3.2) peut être fiable pour simuler un épisode sur 24 heures. En effet, le champ de la répartition zonale des gradients liée à un système d'EDP est sensible à la condition initiale *x0* qui fait diverger la solution au bout d'un temps *t*. Si on s'assure que la durée de simulation est inférieure au temps de divergence de la solution la simulation peut être considérée comme réaliste sinon il faut réinitialiser l'état du système par des méthodes d'assimilation de données. Les simulations sur 24 heures effectuées pour le CET montrent que l'échelle de temps est respectée.

Cette modélisation restitue assez fidèlement les mécanismes de transfert[49] et est mise en place sur quelques scenarii météorologiques caractéristiques identifiés par classification.

A partir d'un scénario jour décrit par des valeurs moyennes, la valeur de chaque variable température *T(x,y,z,t)*, pression *P(x,y,z,t), vent U(x,y,z,t)*... est obtenue en chaque maille d'un maillage 3D ; ceci pour chaque instant *t,* soit un volume de données produites très important.

Cette donnée simulée peut accroître la *Connaissance* nécessaire au scénario lorsque les données mesurées n'existent pas, selon deux postulats:

i) cette information peut améliorer le vecteur de caractéristiques du scénario *jour moyen* : on parlera de clustering *jour - espace*
ii) cette information permet d'identifier des zones de comportement similaires définissant les clusters spatiaux

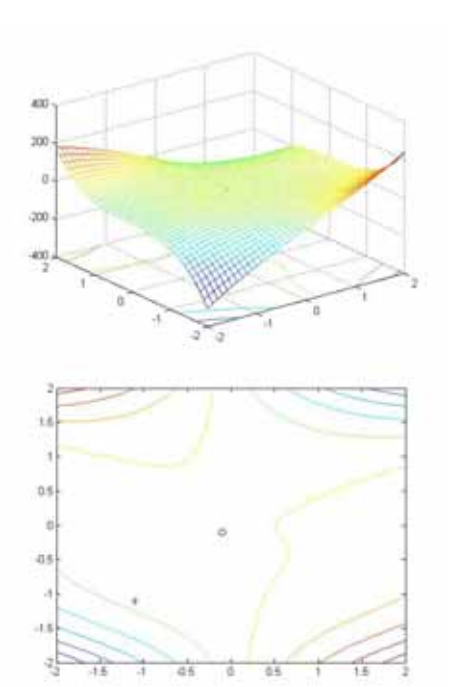
Dans les deux cas, on s'attache à considérer une collection de données référencées dans un repère cartésien pour un instant donné *t*. Il s'agit d'étendre la notion de scenarii jour ponctuel (sans dimension temps espace) à un scénario j*our espace*.

2.4.2 Identification de classe - 'scénario jour espace'

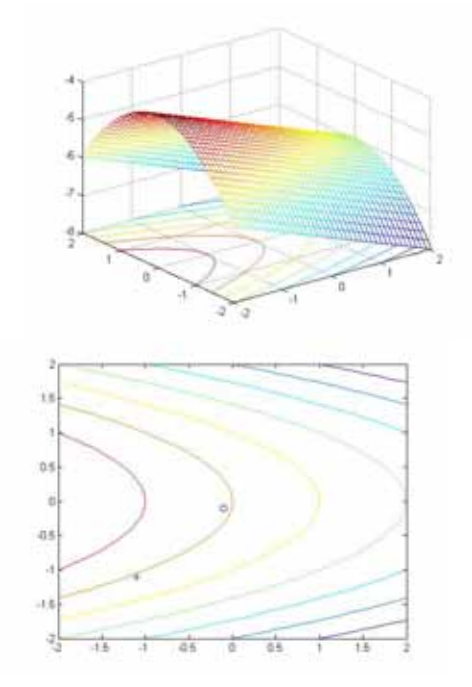
La simulation d'un jour type associé à un scénario est composée de la succession dans le temps de lots de données associées à chaque variable. En chaque maille, est évaluée la valeur d'une fonction *f(x,y,z,t)* : par exemple, $f_T(x,y,z, t_i)$ pour la température au temps t_i. La figure

[49] en annexe les équations de Navier Stokes - équations du mouvement de l'atmosphère

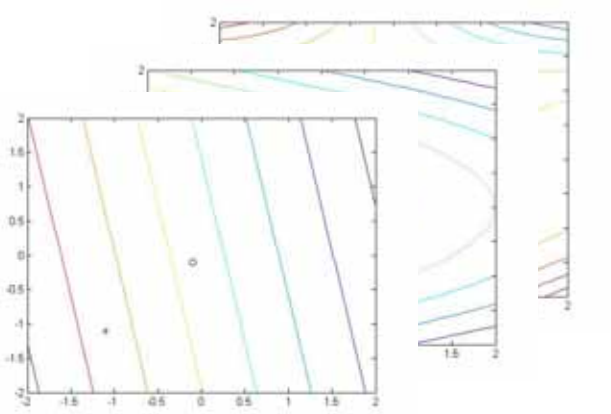
3.2.3 suivante[50] représente une fonction de type *f(x,y)* avec ses courbes d'iso valeurs *f(x,y)= k*. Les fonctions utilisées pour cet exemple sont des fonctions analytiques en 2D, réduisant le problème à un cas presque trivial. Chaque fonction représentée en figure 3.2.3, devient une surface équipotentielle d'une des fonctions (respectivement pour chaque variable, *T , P,* U..) approchée en simulation par CDF (ex. isothermes atmosphériques). Pour chaque variable, sont fournies des iso valeurs (figure 3.2.3.a et figure 3.2.3.b et annexe 3.1).



(a) isovaleurs d'une variable P



(b) isovaleurs d'une variable T



(c) 1 scénario : *n* variables données en (x,y) pour un temps t_i donné (12h00 TU)

Figure 3.2.3 Représentation d'un scénario espace

Un scénario *s* est décrit en 2D par *n* variables, donc par *n* matrices *Z (dim(NX*, N*Y))*, avec Z, la variable dépendante, et ceci pour chaque pas de temps donné t_i.
On cherche à comparer plusieurs scenarii instanciés *S* pour une date donnée t_i (12h TU par exemple) (figure 3.2.4).

[50] par commodité de représentation et clarté, nous avons choisi des exemples de fonction qui ont une projection en isocourbes de la forme *y = f(x,y)* d'une variable *y* . Ceci se généralise en 3D avec *y= f(x,y,z) ;* les surfaces de niveaux sont alors données par *k=f(x,y,z)* bien connues en atmosphère avec les surfaces isobares et isothermes : nous avons omis volontairement la variable indépendante, le temps.

Avant de définir la dissimilarité entre objets, il est nécessaire de définir leur représentation. Cela revient à établir les variables explicatives nécessaires au calcul des distances.

Segmentation spatiale des variables explicatives

Une manipulation sur un jeu de données (figure 3.2.4.) simplifié à l'extrême illustre le raisonnement. On schématise la restitution de la simulation par un ensemble de 6 scenarii pour lesquels a été construit un jeu de données théoriques pour 3 variables continues qui pourraient être la température, la pression et une concentration en CH_4. (Annexe 3.1)

Si l'on regarde les valeurs obtenues pour un point situé à proximité du centre du domaine pour chacun des scenarii (tableau 3.2.1), les valeurs ne sont pas discriminantes sur 1 ou 2 variables. Le choix du point sélectionné peut lisser les différences ou les accentuer.

Scenario	Var 1	Var 2	Var 3
1	8,98	0,74	-0,02
2	15,99	0,24	-0,02
3	5,01	0,18	-0,07
4	-5,01	-0,39	-0,09
5	-5,01	-0,20	-0,06
6	-5,01	-4,96	-0,06

Tableau 3.2.1

Proposons maintenant un découpage de type *quadtree* : le domaine est découpé en 4 : on calcule sur chaque pavé, les valeurs obtenues par les 3 variables ce qui conduit à (3×4) variables.

Scénario	D1	D2	D3	D4	P1	P2	P3	P4	S1	S2	S3	S4
1	6,13	6,26	6,26	6,39	1,01	1,87	-0,41	0,44	-2,06	-3,46	-2,20	-3,09
2	14,09	14,22	14,13	14,26	-0,24	1,37	-0,84	0,77	-0,55	-4,83	-0,83	-4,32
3	7,01	6,94	6,89	6,83	-0,87	0,74	-0,27	1,34	2,30	-23,65	1,03	-21,02
4	-7,52	-7,36	-7,46	-7,29	-1,29	0,31	-0,99	0,61	-55,03	28,22	12,05	-30,99
5	-5,72	-5,68	-5,69	-5,65	0,71	-0,89	0,41	-1,19	2,44	-18,62	2,53	-17,22
6	-5,72	-5,68	-5,69	-5,65	-5,19	-6,19	-5,13	-6,13	2,44	-18,62	2,53	-17,22

Tableau 3.2.2

Une classification hiérarchique couplée à une ACP donne un regroupement de classes suivantes :
[Le scénario 1 (indice 3,05) avec le scénario 2 (5,66)]
[Le scénario 4 (indice 16,39) avec le scénario 5(26,11)]
puis [[Le scénario 1 (indice 3,05) - scénario 2 (5,66)] et scénario 3 (indice 48,79)]
puis [[scénario 4 indice 16,39) avec le scénario 5(26,11)] et scénario 6 (indice)]
puis le regroupement final des 2 branches.

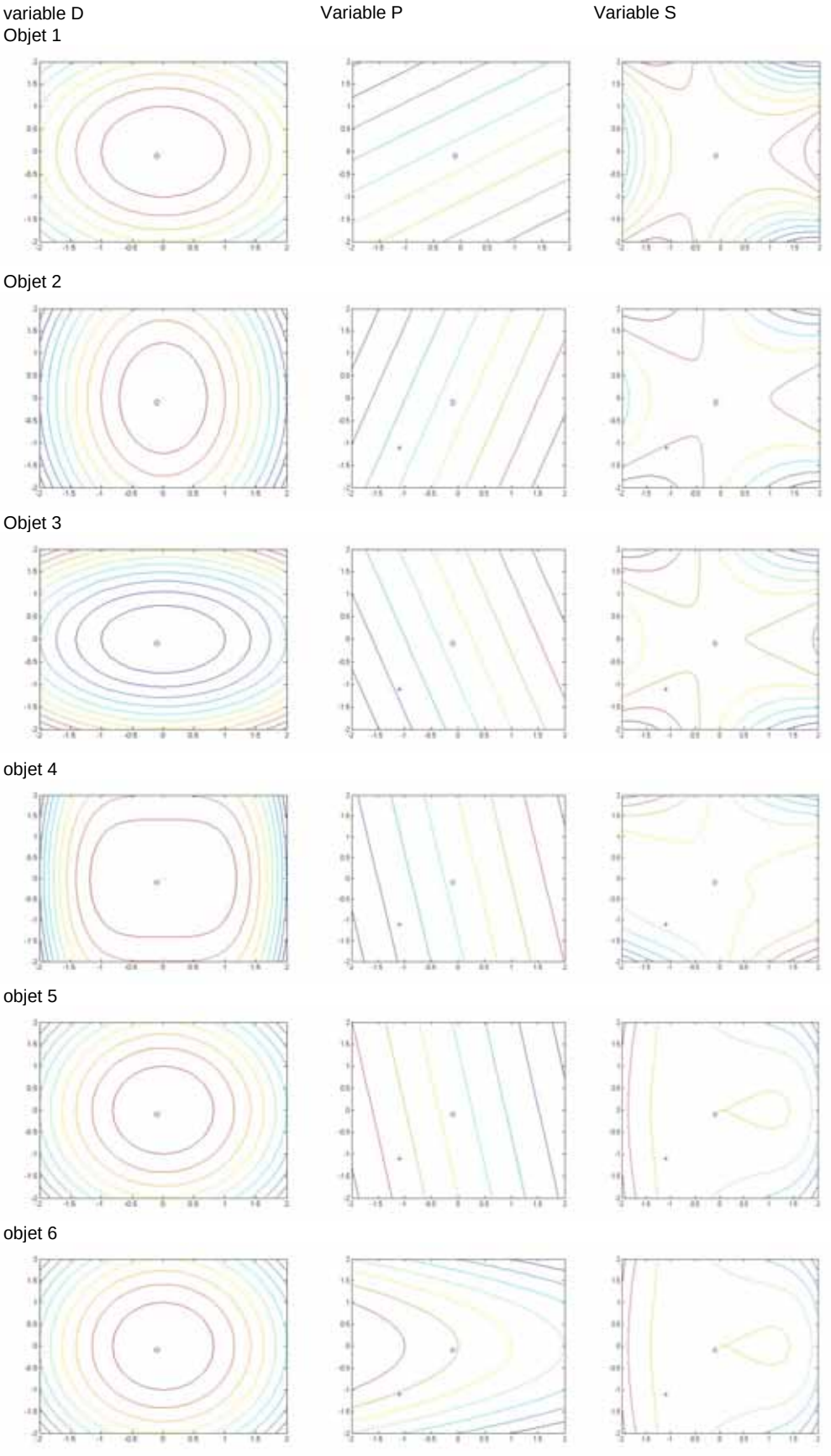


Figure 3.2.4 . 6 Scenarii construits à partir de 3 variables dont la solution est une fonction *z= f((x,y)*

Ce regroupement est différent de celui avec 3 variables[51]. La classification effectuée sur 3 variables à tendance à montrer 3 groupes, tandis que la prise en compte de 12 variables augmente la variance des classes. Ceci ne serait-il pas seulement la conséquence de l'existence d'une forte corrélation linéaire pour les variables *P ?* Cela reste- il vrai si on a des formes spatiales moins prévisibles ? Quels sont les points qui finalement ne sont pas des redondants d'information par rapport à ces variables moyennes ?

Autant de questions qui peuvent être énoncées de la façon suivante :
- soit les valeurs en tous points (X,Y) de l'espace plan constituant une variable indépendante de l'objet *s* (attributs de l'objet scénario), chaque *s* appartient à un espace Euclidien $\Re^d$ de dimension $d = n \times NX \times NY$: quelles méthodes de sélection ou de réduction des variables pertinentes sont envisageables ?

1ère solution) Elle est illustrée dans l'exemple précédent ; on décide de segmenter l'espace plan en zones régulières, sur chacune est calculée une valeur moyenne : on parle de *segmentation supervisée* car l'expert fixe les zones, le point représentatif et surtout l'obtention de la valeur : moyenne arithmétique, moyenne géométrique, moyenne quadratique,… On peut aussi envisager d'intégrer d'autres valeurs comme l'écart-type sur cette valeur moyenne.

Dans ce cas, on augmente considérablement la dimension de l'espace de représentation : une ACP permet d'identifier si une réduction est possible avant d'évaluer la similitude entre les jours.

Points positifs[52] : les méthodes de clustering restent valables pour une dimension *d* du vecteur de caractéristiques [53] : $d = (n \times 4^q)$ - le pré traitement de la donnée est sans difficulté

Points à résoudre :
- le choix du nombre de découpage, *q*
- le choix de la valeur représentative de la valeur centrée sur un cadran
- la disponibilité des données en chaque maille, issues de la simulation
- la mise en oeuvre de différentes méthodes de clustering

Perspectives et Résultats attendus

Pour pallier l'absence de données de simulation systématiques, d'autres types de données sont utilisables :
- 2 ou 3 stations au sol constituent un maillage privilégié sur lequel sont retenues les mêmes variables : un clustering sur des données localisées est implicitement faisable
- l'interpolation spatiale sur un réseau de mesures au sol suffisamment dense reste une piste, mais peu réaliste sur les jeux de données disponibles

Le résultat attendu doit permettre d'accroître la discrimination des objets, ce qui n'est pas acquis a priori.
Notons enfin, que finalement, un point moyen sélectionné pourrait être vu comme une pseudo-station.

[51] Le test de regroupement par classification hiérarchique des scenarii : 2 branches [1,2] et [4,5] puis, regroupement de [4,5,3] puis [4,5,3,6] enfin [1,2] avec [4,5,3,6]
[52] points méthodologiques résolus et intéressants dans ce type de modélisation de la donnée (le plus)
[53] q= 1 pour découpage en 4 puis on peut choisir de découper en quatre soit q=2 etc..

2ème solution) Le choix de sélection des variables pertinentes doit être non supervisé afin de ne pas biaiser l'information utilisée dans la classification. Il s'agit d'extraire ou de projeter un vecteur objet d'un espace dans $\Re^d$ où $d = n \times NX \times N Y$.
Le principe de l'ACP consiste à associer un vecteur caractéristique $t \in \Re^l$ *à* chaque vecteur de caractéristique *x,* qui optimise sa représentation au sens de la minimisation de l'erreur d'estimation de *x* ou de la maximisation de la variance de *t.* Les vecteurs *x* et *t* sont liés par une transformation linéaire du type $t = P^T x$ où la matrice de transformation $t \in \Re^{d \times l}$ vérifie la condition d'orthogonalité $P^T P = I_l$.
Cette transformation linéaire s'apparente à une projection de l'espace des données de dimension *d* vers un sous-espace orthogonal de dimension *l*, qui est optimale si l'erreur quadratique d'estimation des données *x* est minimale soit :

$$P_{opt} = arg \ _P \ min \ J_e(P)$$ avec J_e le critère d'erreur d'estimation[54]

La maximisation de la variance de projection des données s'apparente à la détermination des vecteurs propres de la matrice de covariance. Tout vecteur *x* peut être représenté par les *d* vecteurs propres de la matrice de covariance pondérés par les composantes principales $t_i = p_i^T x$. Le vecteur *x* de données est :

$$x = P\mathrm{t} = \Sigma t_i \ p_i \quad \text{pour i=1:}d$$

L'estimation consiste à retenir les *l* premières composantes qui présentent les plus fortes variances et $l < d$. De ce fait, l'ACP fournit une estimation de $\hat{x}$ et permet d'envisager une modélisation du système.

Pour le vecteur de donnée aléatoire scénario *espace*, un problème identifié rapidement sera que la matrice de corrélation est de grande dimension $(n \times NX \times NY) \times (n \times NX \times NY)$. La recherche d'une méthode d'agrégation ou de séparation en sous problèmes par exemple par ACP partielle pourrait être envisagée. Les travaux de (Oja 1989, Rubner et al. 1989, Kung et al. 1990) proposent des implémentations sous forme de réseaux neuronaux de l'ACP linéaire. Il s'agirait alors de tester ce type d'apprentissage sur chacun des 2 critères : la maximisation de variance et la minimisation de l'erreur quadratique de l'estimation des données de ce type.

Une restriction au développement de l'ACP, est liée au fait que seules les dépendances linéaires ou quasi linéaires seront révélées ; cependant, la physique de l'atmosphère montre une forte non linéarité entre certaines variables de température, de vent et de pression. Seule l'humidité a une corrélation linéaire avec la température.

L'extension de l'ACP pour des problèmes non linéaires est proposée par (Hastie et al 1989, Kramer et al. 1991) basée sur le principe des courbes principales. Une courbe principale est une courbe lisse minimisant la distance entre tous les points et leurs projections sur la courbe. Elle permet de calculer des composantes principales non linéaires unidimensionnelles. Les algorithmes d'apprentissage par réseaux de neurones cherchent à minimiser une fonction coût par des méthodes d'optimisation non linéaire de façon itérative en modifiant les poids en fonction du gradient de la fonction coût. L'approximation de la courbe doit permettre d'expliquer un pourcentage élevé de corrélations totales de variables (Harkat 2003). L'une des difficultés du développement de l'ACP non linéaire par approche neuronale est la difficulté de convergence de l'apprentissage du réseau (nombre de paramètres à optimiser du réseau important), l'initialisation difficile et des temps de calcul longs. (Harkat 2003) a proposé une combinaison des courbes principales avec un réseau de fonctions à base radiale (RBF) qui

[54] $J_e(P) = \mathrm{E}(\|x - \bar{x}\|^2\}$

permet également de déterminer le nombre de composantes principales. Une perspective d'application serait d'évaluer les différentes approches pour une ACP non linéaire qui permettrait d'estimer chaque scénario *espace* par une projection sur 1 ou l plusieurs composantes non linéaires, ce qui fournirait alors un modèle intrinsèque du système.

Points positifs[55] : des algorithmes pour des ACP linéaire de grande taille sont développés et de nombreux travaux sont proposés sur les méthodes d'ACP non linéaire.

Points à résoudre :

- la disponibilité des données en chaque maille, issues de la simulation
- le pré traitement pour extraire les matrices de la simulation par CDF
- la dimension acceptable et "gérable" de la dimension des données
- la mise en oeuvre d'une architecture pour évaluer les l courbes principales

Résultats attendus

L'ACP linéaire ou non linéaire représente déjà un modèle du système *scénario*, mais l'objet qui nous préoccupe a priori est la représentation du vecteur de caractéristiques x du scénario dans une dimension $l < d$ et un vecteur caractéristique $t \in \Re^{l.}$
Le résultat attendu est de savoir si la réduction de données pour ce vecteur scénario *espace* est raisonnable, permettrait- elle de représenter un objet de type forme 2D ?
On cherche à savoir si ces objets ont un sens en dimension d ou réduite l, et surtout s'ils apportent des éléments supplémentaires par rapport à une *segmentation supervisée* (1ère solution), ou à un scénario jour *moyen* ?
Enfin, en fonction du modèle transcrit par l'analyse en composantes principales et des résultats obtenus, il s'agirait de savoir si les méthodes de clustering sont utiles et comment elles peuvent être mises en place ?

La limite de ce type d'approche est liée à la disponibilité de données car les temps de simulation sont longs ; sur les projets de recherche en cours sur lesquels nous travaillons, l'échantillon disponible reste faible.
L'utilisation des résultats des simulations de code de dispersion de l'atmosphère comme le modèle de type Chimère, code de calcul à méso échelle de la qualité de l'air pour l'ozone représenterait une possibilité et une opportunité. En effet, ce type de code fonctionne en continu sur de longues périodes estivales, et les sorties de modèle sont peu exploitées. Ces sorties permettraient d'alimenter un échantillon important de résultats de simulation. Il s'agirait alors d'extraire une information sur le comportement des jours sur 6 mois ou un an de simulation comme la traçabilité des types de temps, leur évolution, la détection d'évènement moyen ou non ordinaire.

[55] points méthodologiques résolus et intéressants dans ce type de modélisation de la donnée (le plus)

2.4.3 Identification et reconnaissance de formes dans l'espace des vecteurs caractéristiques

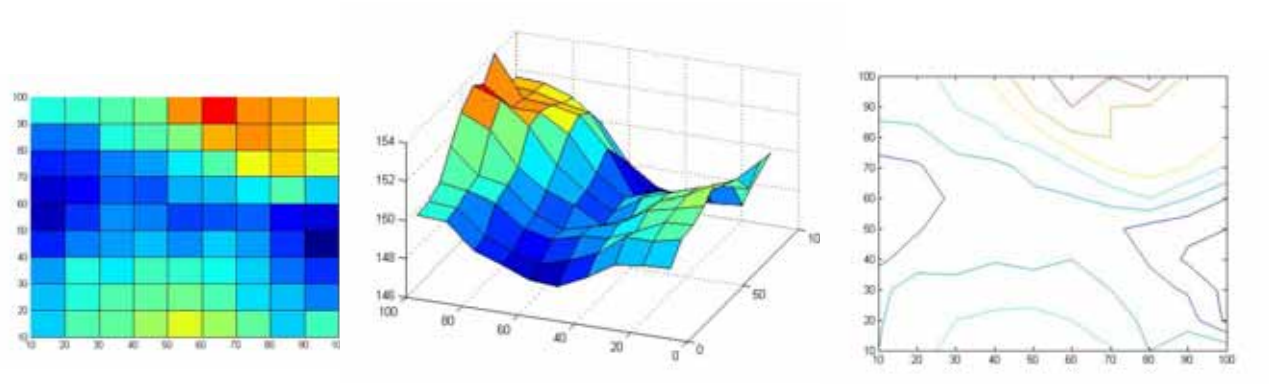
Rappelons que les sorties de simulation par un code de CFD permettent d'obtenir pour un scénario jour des valeurs en chaque maille (la température *T(x,y,z,t)*, pression *P(x,y,z,t), vent U(x,y,z,t).,.* d'un maillage 3D, et ceci pour chaque instant *t* . On cherche à exploiter cette information pour :

- i) accroître la description spatiale du scénario, en utilisant des clusters spatiaux
- ii) identifier des formes plus générales dans l'espace des variables.

Sur un empilement de maillages 2D (on réduit le maillage 3D à une donnée plan pour des altitudes caractéristiques) il s'agit identifier des formes comme si on disposait d'une image (mode pixels) pour chaque variable d'un processus dynamique. Chaque vecteur contient un élément du maillage, avec les données suivantes : les coordonnées en X et en Y, les valeurs des **p** paramètres (*T°, H ,P,* vent) expliqués par la simulation.

Classification spatiale

On cherche une partition des données en groupes où la dissimilarité doit intégrer les attributs spatialisés de l'espace cartésien. Cette classification est non supervisée car on ne peut assigner à un point, une appartenance à une forme lors de l'apprentissage ou designer le nombre de classes, comme on pourrait le faire pour identifier un objet sur une image comme le contour de forme, type de spécificité de la zone … telle l'identification de bordure, d'un caractère ou d'une bactérie (Ayala G. al 2006). L'analyse doit être de type multivariée. La figure 3.2.5 illustre ce propos : sur un maillage régulier pour une variable de type gradient, les pixels ayant une valeur proche doivent être regroupés en fonction de leur position . En effet, les 2 zones de valeur maximale sont proches si l'on n'intègre seulement la caractéristique température mais doivent être distinguées en intégrant la position dans le plan cartésien. Elles matérialisent deux zones de comportements différents soit deux sources distinctes (figure 3.2.5c).



a) b) c)

Figure 3.2.5 Clustering spatial avec l'identification de deux formes égales à 2 sources distinctes pour une variable de type gradient

La description de structures spatiales et le test de la présence d'auto corrélation spatiale utilisent le développement de 3 fonctions que sont le corrélogramme, le variograme et périodogramme. Elles permettent de quantifier la dépendance spatiale, la partition en fonction de la distance des classes de valeurs spatialisées et mettent en évidence le caractère isotropique ou non du phénomène observé. Ils peuvent qualifier une structure de type marche, gradient, homogène, ou sinusoïdale. Ils se généralisent à des données multivariées avec les travaux de (Sokal 1986) (Sokal et al 1997) en utilisant une statistique normalisée de Mantel.

(Legendre et al. 2004) montrent que le clustering spatial est nécessaire aux données multivariées spatialisées, ce qui corrobore (et le montre l'exemple précédent) cette idée de clusters sur un échantillon pris parmi des données simulées. Le maillage régulier fournit à chaque point une topologie explicite qui est assimilable à la contrainte de la continuité spatiale proposée par (Legendre 1978). Cette contrainte est fournie par une matrice de continuité entre deux points. Le produit de cette matrice avec la matrice de similarité (distance entre deux points projetés dans $\Re^d$) permet de modifier l'appartenance possible d'un point à un cluster. Dans (Legendre et al 2005), seules les méthodes de clustering de type hiérarchiques sont développées avec cette contrainte. Elles peuvent facilement être introduites dans la méthode des Kmeans. L'introduction d'un variogramme multivarié permet d'appliquer une fonction de pondération spatiale dans (Bourgault et al 1992). Plus récemment des travaux qui généralisent la proximité topologique comme contrainte supplémentaire pour un algorithme de type EM de plus proches voisins, sont proposés par (Ambroise C. et al 1998) et appliqués par (Hu et al 2006).

Points positifs : l'échantillon peut être de taille importante : (si nombre de mailles = 1000 alors on a (1000 × 1000) individus). Des travaux sont proposés sur les méthodes de clustering spatial avec des algorithmes disponibles.

Points à résoudre :

- la disponibilité des données en chaque maille, issues de la simulation,
- le pré traitement pour extraire les matrices de la simulation par CDF,

Perspectives et Résultats attendus

Il s'agirait d'obtenir sur un ensemble choisi de scenarii dont les matrices de simulation 3D sont disponibles, la mise en évidence de formes caractéristiques ; par exemple une source de chaleur (liée à une activité en sous-sol de décomposition des déchets identifiée, ou au contraire un puits, des zones de gradients moyens ou des points selles de la fonction température). Développer différents algorithmes de clustering qui introduisent des variables de localisation sous forme de contiguïté spatiale ou sous forme de variables explicatives supplémentaires (les variables X et Y sans topologie dans le clustering) est une perspective intéressante.
Rappelons qu'ici le vecteur de caractéristique concerne un scénario calculé pour un temps donné (ex.12h00 TU).

2.4.4 Généralisation : reconnaissance des formes dans l'espace des états

Un développement très attendu est d'utiliser ces méthodes de clustering spatial pour identifier des formes résultantes comme des entités homogènes d'un point de vue comportement (ex. pour la température : zone source ou puits). Les formes issues du clustering seraient projetées dans un espace des variables afin de distinguer la valeur moyenne, la variance et l'écart type du cluster. Afin d'identifier le domaine de variation de chaque variable de l'espace d'état du système (cf. paragraphe 3.2) dont l'espace géométrique. L'itération de ce type d'algorithme de clustering sur une séquence de 24h00 ou plusieurs jours (à heure fixe) sur des scenarii simulés permettrait de suivre l'évolution dans le temps.
Cette perspective dépend fortement des résultats obtenus au point précédent comme le clustering spatial et des formes qui peuvent être identifiées. Peu de choses sur cette approche ont été recensées en bibliographie, ce qui fait que ce projet de recherche a une échéance plus longue et pourrait être exploré de manière approfondie. Un cas d'école sur des fonctions classiques d'un transfert de chaleur par gradient thermique pour une source au sol constituerait un bon point de départ pour explorer les possibilités d'une telle approche.

Savoir si la prise en compte des formes par clustering spatial peut améliorer la connaissance d'un scénario jour espace dont le vecteur serait enrichi par des variables liées aux formes identifiées, représente un résultat attendu.

3. Approche couplée [vecteurs d'état - systèmes dynamiques - multi-modèles] : cohabitation entre classification et simulation déterministe

Ce paragraphe tente de transcrire deux réflexions à partir des éléments présentés aux §. 2.3 et 2.4.
Auparavant, revenons aux manipulations successives du vecteur de caractéristique (ou de données) $x(k)$ pour un scénario k :

- les premiers travaux considèrent ce scénario comme l'élément caractéristique d'une classe de jours, représentatifs de conditions météorologiques ; le temps t n'apparaît pas comme une variable explicite, les valeurs des variables explicatives sont de type valeurs moyennes ou agrégées.

- ensuite, le temps est introduit dans le §2.3 : soit en augmentant la dimension *d* du vecteur $X=[x_1, x_2 \ldots x_d]^T$, soit sous la forme matricielle $X \equiv \{ x_{ijt} : i = 1, N ; j = 1, D; t = 1, T\}$ avec *N* individus, *D* variables explicatives, *T* le temps.

- puis, est intégrée une dimension spatiale de la donnée, en introduisant une instance de chaque paramètre, soit *n* paramètres météorologiques ce qui conduit $\Re^d$ où $d = n \times NX \times N Y$; on parle de scénario *espace.*

- en réalité, le scénario *temps* et le scénario *espace* devraient ne représenter qu'un seul et unique vecteur de caractéristiques : sous forme matricielle : $X \equiv \{ x_{ijt} : i = 1, N ; j = 1, n \times NX \times N Y; t = 1, T \}$ ou encore $X_k(t) =[x(t)_1, x(t)_2 \ldots x(t)_d]^T$.

Dans la théorie des systèmes dynamiques, un système physique (comme l'atmosphère, l'hydrodynamique…) prend au fil du temps des configurations possibles définissant l'espace d'états ou espace de configurations. Les mesures associées[56] à un point de cet espace de configurations sont des valeurs qui vont caractériser les propriétés pertinentes du système. On peut alors assurer que le vecteur de forme 'scénario espace' représente pour chaque temps *t* un ensemble des mesures assimilables aux coordonnées généralisées du système.

On s'intéresse alors non plus à un vecteur de données mais à un vecteur d'état d'un système plus global (l'émission d'un CSD soumis à la dynamique atmosphérique locale).

3.1 Evolution vers un modèle de système par EDO

L'évolution de milieux continus comme l'atmosphère est exprimée par des systèmes d'équations différentielles aux dérivées partielles[57] ($\partial u / \partial t = F(u)$) ou $u = u(x,t), t$, t le temps et x le point ; pour définir complètement la solution il faut des conditions supplémentaires. La dynamique des équations aux dérivées partielles est difficile à obtenir car on en connaît peu de choses (Della Dora J. et al 1993).

L'étude de la dépendance de la solution d'une équation différentielle ordinaire non seulement par rapport au temps mais surtout en fonction des conditions initiales (formulée par H. Poincaré) conduit à décomposer une équation différentielle d'ordre *n,*

$$x^{(n)} = f(t, x, \dot{x}, \ldots, x^{(n-1)})$$

par un système du premier ordre :

$$x_0 = x \begin{cases} \dot{x}_0 = x_1 \\ \dot{x}_1 = x_2 \\ . \\ \dot{x}_{n-2} = x_{n-1} \\ \dot{x}_{n-1} = f(t, x_0, x_1, \ldots, x_{n-1}) \end{cases}$$

Un système dynamique général est alors :

[56] appelées coordonnées généralisées du système
[57] connues depuis Euler

$$(\delta) \quad \begin{cases} \frac{dx}{dt} = f(x(t),u(t)),\ x(t_0)=x_0 \\ y(t)=h(x(t)) \end{cases}$$

où $x \in R^n$ est le vecteur d'état, $u \in R^m$ le vecteur d'entrée, $y \in R^p$ le vecteur de sortie, x_0 la condition initiale au temps initial t_0, et $f : R^n \times R^m \rightarrow R^n$ et $h : R^n \rightarrow R^p$ ([58]).
qui devient pour les systèmes linéaires stationnaires :

$$(\delta_L) \quad \begin{cases} \frac{dx}{dt} = Ax(t)+Bu(t) \\ y(t)=Cx(t) \end{cases}$$

Dans le cas plus général d'un système non linéaire, la représentation de l'état est donnée par deux fonctions : la première représente l'équation d'état du système et la seconde, l'équation de sortie du système, soit :

$$(\delta) \quad \begin{cases} \dot{\mathbf{x}}(t) = \mathrm{f}(t,x(t),u(t)), \\ \mathbf{y}(t) = \mathrm{h}(t,x(t),u(t)) \end{cases}$$

Soit un système classique, un pendule libre dont l'équation différentielle est :

$$ml\ddot{\theta}(t) = -mg\sin\theta(t) - kl\dot{\theta}(t)$$

où θ est l'angle du pendule, m la masse du pendule, g accélération gravitationnelle, k coefficient de frottement, l le rayon du pendule.

Son système d'état devient :

$$\begin{cases} \dot{x}1(t) = x2(t) \\ \dot{x}2(t) = -\frac{g}{l}\sin x1(t) - \frac{K}{m}x2(t)) \end{cases}$$

où $x_1(t) = \theta(t)$ angle du pendule, $x_2(t)$ = vitesse angulaire et $\ddot{x}_1$ accélération angulaire du pendule.

La possibilité de remplacer le système d'équations différentielles partielles de l'atmosphère par un système d'équations différentielles ordinaires qui fournirait l'état du système à un temps donné, demeure imprécise, vaste et surtout floue. La première solution dépend des possibilités ou non de réduire ou simplifier les équations de l'atmosphère, ce qui semble un peu périlleux d'autant qu'une discipline entière de la Physique leurs est consacrée. La seconde solution consiste à évaluer si le comportement du système entre l'atmosphère et un site (comme un CSD) ne pourrait pas se ramener à quelques formes de base de la physique (système de ressort, de transfert de matière) qui ont déjà des solutions par système d'état.

[58] sous des propriétés de régularité des fonctions

3.2 Modélisation floue multimodèle

La modélisation floue introduite par (Zadeh 1965) via un système d'inférence floue, permet de trouver un ensemble de relations entre des entrées-sorties, qui sont des estimateurs du fonctionnement du processus dans une zone de fonctionnement donnée. Dans la base de règles associée, où la règle est "Si prémisses alors conclusion", la prémisse et la conclusion sont des conjonctions ou disjonctions de propositions floues de types *"x est A*", *x* variable floue et A un sous ensemble flou de type linguistique (grand, petit…). Les prémisses sont ici les entrées du modèle et les conséquences, la sortie prévue par le modèle. Le modèle Tagaki-Sugeno (Tagaki et Sugeno 1985) propose de décomposer l'espace des entrées en régions floues et d'approximer chaque région par un modèle linéaire ou non. Ce qui revient à considérer un système comme la combinaison de sous-systèmes. Chaque sous-système est une structure plus simple: le modèle global est alors obtenu par interpolation entre les différents modèles locaux, de type moyenne pondérée, d'ou l'approche multi-modèle, terme que l'on retiendra.

La structure d'un multi-modèle adopté est celle de (Ragot J. 2005) : soit un système physique ayant *r* entrées et une sortie *y* :

$$u = [u_1, u_2 \ldots u_r]^T \in U = U_1 \times U_2 \times \ldots \times U_r$$
$$y \in R$$

Le système est un ensemble de règles (modèles partiels) liant u_i (entrées) et *y (sortie)*. Un modèle partiel est représenté par des structures linéaires (généralisables au non linéaire) :

$$\begin{aligned} &règle \quad i : SI \ u_1(k) \ est \ M_{i1} \\ &\qquad\qquad ET \ u_2(k) \ est \ M_{i2} \\ &\qquad\qquad \ldots \\ &\qquad\qquad ET \ u_r(k) \ est \ M_{ir} \end{aligned}$$

$$ALORS \ \ y_i(k+1) = \sum_{j=1}^{q} a_{ij} y(k+1-j) + \sum_{j=1}^{p} b_{j1} u_1(k-j) + \ldots + \sum_{j=1}^{p} b_{jr} u_r(k-j)$$

où M_{ij}(j=1,…,r ; i=1…R) qualificatifs sont définis sur le domaine U_i

Une fonction μ(y(i)(k+1) traduit son influence limitée ; l'ensemble de ces fonctions représente les poids attribués aux différents modèles partiels[59]

Les modèles partiels peuvent être réunis sous forme barycentrique, fournissant la valeur prévue *y(k+1)* au temps k+1, soit :

$$y(k+1) = \frac{\sum_{i=1}^{R} \mu_i(k,\gamma) \left(\sum_{j=1}^{q} a_{ij} y(k+1-j) + \sum_{j=1}^{p} b_{j1} u_1(k-j) + \ldots + \sum_{j=1}^{p} b_{jr} u_r(k-j) \right)}{\sum_{i=1}^{R} \mu_i(k,\gamma)}$$

[59] importance = poids d'une variable ; différentes fonctions sont utilisables comme : $\mu_{ij}(u_i) = e^{-\left(\frac{u_i(k)-m_{ij}}{\sigma_{ij}}\right)^2}$; forme générale $\mu_i(k,\gamma)$ paramètres regroupant m et σ

avec : $\mu(y_i(k+1)=\prod_{j=1}^{r}\mu_{ij}(u_j(k))$

Les difficultés d'obtention du modèle sont liées à la définition du nombre de règles, la structure des prémisses, et les fonctions conséquences[60]. Résoudre simultanément l'identification de la partie prémisses et celle des conséquences est difficile ; une approche séquentielle est une alternative mais qui ne garantit pas l'optimalité de la solution. Des méthodes de partition par *grille* permettent d'obtenir la décomposition de l'espace des variables prémisses mais souffrent d'une rapide explosion combinatoire (cas de 4 variables à 3 modalités : soit $3^4 = 81$ règles). Une partition de l'espace des variables des prémisses basée sur la décomposition en hyper cubes flous permet d'identifier également l'ensemble flou de chaque variable. Ces méthodes sont proches des méthodes de segmentation par regroupement autour de centres pouvant avoir la propriété de noyau gaussien. Différentes méthodes peuvent se combiner pour affiner l'obtention du modèle (Quach R.P 2002).
L'application de ce type de multi-modèle à la prévision de la concentration d'ozone par (Mourot et al.1997) à partir d'un échantillon de données issues d'une station de mesures de la qualité de l'air montrent des résultats tout à fait intéressants pour un processus de formation de l'ozone troposphérique que l'on sait fortement non linéaire. Notons que le modèle utilise seulement un échantillon des 17 jours soient 1600 observations avec 11 jours utilisés pour l'identification du modèle.

La pertinence de la prévision par des modèles de représentation est évidente bien que dans le monde de l'atmosphère, la mécanique des fluides reste la référence en terme d'approche la modélisation d'un système et de sa prévision. La seconde idée qui émerge est de modéliser le système *scénario météorologique* ou un autre système d'échange atmosphère - infrastructure par ce type d'approche en exploitant simultanément les mesures et les données disponibles que sont les sorties de la simulation.

3.3 Eléments de construction des 2 propositions

A ce stade, il s'agit de formuler la démarche avec 3 types d'information :

- les données et les connaissances issues du paragraphe 2.
- la formulation en terme de multi-modèle
- l'approche d'un système d'état pour évaluer la production de polluant ou de production énergétique pour un site industriel (type CSD).

[60] en mode supervisé, il existerai un moyen d'avoir le nombre de règles, or dans un système quelconque, une identification du nombre de règles, des variables des prémisses, de leur modalités ,des fonctions de conséquences doit être évaluée de façon non supervisée

3.3.1 Information disponible

L'objet de notre préoccupation est le scénario jour météorologique : il est décrit dans le temps et l'espace par un vecteur caractéristique $X_k(t) = [x(t)_1, x(t)_2 \ldots x(t)_d]^T$.

Suite aux éléments et constructions précédemment proposés, nous admettons que les éléments suivants (figure 3.3.1) sont obtenus :

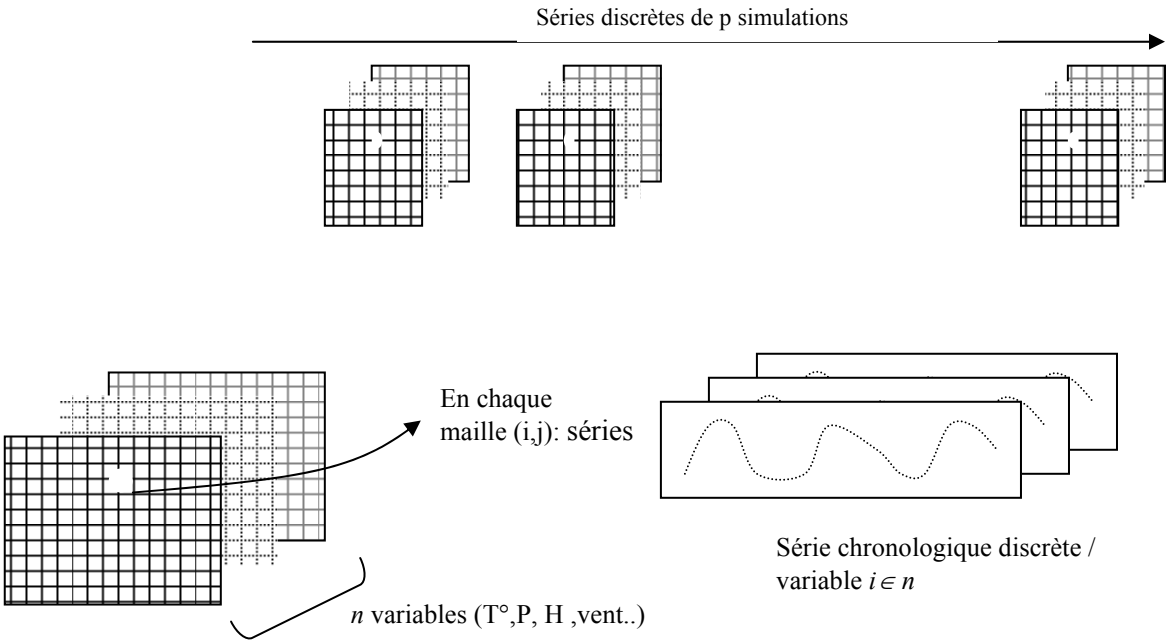


Figure 3.3.1 : représentation *temps* - *espace* d'un scénario jour

1) La simulation déterministe permet d'obtenir un ensemble de n matrices de dimension $(NX \times NY)$ pour chaque scénario s à chaque instant de simulation t choisi, avec t variant de [T1 ;T2]. Ce qui fournit N instances pour un pas de temps Δt donné.
A chaque maille (i,j) est associé un vecteur de type $X=[x(t)_1, x(t)_2 \ldots x(t)_n]^T$, qui est une suite discrète de valeurs liée au choix de ne retenir qu'un sous-ensemble N des simulations.
Chaque scénario dispose en théorie à t de $(NX \times NY)$ points (i,j) disposants eux-mêmes d'un vecteur $X=[x(t)_1, x(t)_2 \ldots x(t)_n]^T$; en réalité nous ne retiendrons dans un premier temps que certains points du maillage pour des raisons évidentes de taille de données.

Supposons également que l'on dispose des moyens qui permettent d'avoir (figure 3.3.2):

- une plainte qui peut être associée à un signal d'odeur de type échelon, en différents points (i,j) du maillage. Cette donnée n'est pas systématique mais est localisée pour certains scenarii simulés
- une concentration d'un polluant est instanciée en chaque maille (i,j) si le code de dispersion dispose de sources de polluant. On généralise alors le scénario météorologique comme un scénario de l'état dynamique et chimique de l'atmosphère.

A cette étape, les données décrivant le scénario espace et le scénario temps sont disponibles.

2) On serait capable d'identifier des formes (identifiables), clusters dans l'espace de dimension *n* avec *NX ×N Y* individus (§2.4.3.2).

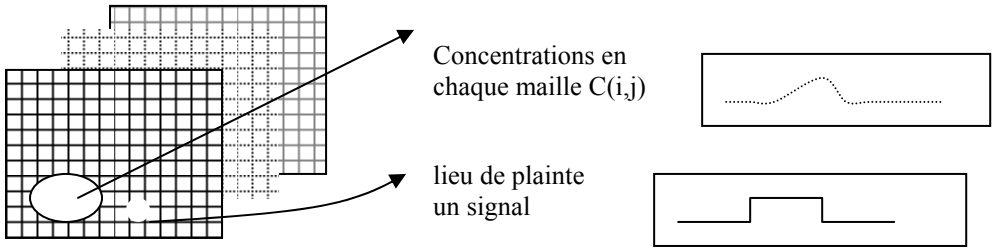


Figure 3.3.2 : Intégration d'une donnée de type odeur, concentration

3) Des stations de mesure au sol, permettent également de fournir des données, positionnables sur le maillage *NX ×N Y*.

4) La disponibilité des sorties de code CFD, n'étant pas systématique, sauf dans certains cas (utilisation des sorties de PREV'AIR), un ensemble de scénarios peut être construit à partir de cas analytiques (modèle gaussien pour la concentration, un champ de type gradient pour le vent, et la température).

3.3.2 Perspectives du développement d'une modélisation floue multi-modèle

On s'intéresse à un prédicteur localisé en une maille (ou une station de mesure) d'une variable *y(k)* un peu particulière, une odeur. Cette variable est assimilée à une fonction seuil dont on cherche à évaluer son apparition, faute de valeurs numériques de la variable à expliquer. La fonction à évaluer peut être assimilée à une forme récurrente dans l'échantillon pour des variables d'entrées de type continues. Une première étape consisterait à envisager un modèle de prévision de *y(k)* de type multi-modèle quasi linéaire de type Takagi Sugeno (Mourot et al 1997), fournissant ŷ(k) pour l'odeur.
Les difficultés attendues sont classiques bien que nombreuses : le choix des variables des prémisses, le nombre de modalités par variable, le nombre de règles, la sélection d'une structure de modèle, l'identification et la validation du modèle.
Bien évidemment, un capteur en continu d'une concentration d'un polluant comme le méthane aurait été plus adapté, mais n'est pas forcément disponible.
Itérer un tel multi-modèle sur des données simulées en différents points du maillage (sorte de station virtuelle), permettrait de savoir si les données de simulation peuvent être exploitées par des multi-modèles, d'évaluer si la même réponse dépend de la localisation du point du maillage, d'envisager si la prise en compte d'un ensemble de points du maillage pourrait améliorer la prévision.

La disponibilité des données *espace* pour une simulation introduit une dimension supplémentaire, la localisation géométrique : ainsi une variable de prémisse qui prend 3

modalités dans le temps, pourra avoir 3 modularités dans un certain espace géométrique et deux modularités (ou différentes) dans une autre partition de l'espace géométrique. La question est de savoir si cette contrainte déjà exprimée en Reconnaissance de formes dans l'espace des états (§2.4.3.2) pourrait être intégrée à un multi-modèle espace. S'agit-il alors d'avoir un vecteur d'entrée U_r avec $r = r + 2$, dont chaque entrée u_i a une dimension ($NY \times NX$) et deux entrées supplémentaires $u_{i'}$ = coordonnées en **x**, $u_{i''}$ = coordonnées en **y,** ou au contraire d'une juxtaposition de ($NY \times NX$) multi-modèles ?

3.3.3 Possibilité d'établir un système dynamique

Le postulat de base comporte plusieurs éléments disponibles :

- les équations différentielles partielles, qui régissent le mouvement de l'atmosphère,
- une reconnaissance de forme sur des données simulées, qui pourraient permettre de reconnaître des comportements homogènes dans certains sous-espaces des états,
- quelques points de mesure en continu des variables du système,
- des modèles de physique connus comme un ressort, l'équation de bilan, quelques lois (loi des gaz parfaits)…

Pour rester à l'échelle de ce qui est concevable, on proposerait de partir d'un cas d'étude, comme celui d'un centre industriel de type CSD qui produit des composés gazeux comme le méthane. Il est enclavé dans un substratum supposé inerte par rapport à cette activité ; ces émissions sont transmisses à l'atmosphère qui réagit et va absorber puis évacuer ces éléments vers l'extérieur du site. L'environnement est un milieu gazeux l'atmosphère, en mouvement.

On cherche à :

i) prévoir la possibilité d'une certaine concentration ou seuil (ex. avoir une odeur autour d'un site de CSD). On se reportera aux points envisagés au volet 2 et 3 sur le modèle de représentation - le clustering – la reconnaissance de formes - les réseaux bayésiens et la modélisation par CFD (cf. volet 2) ainsi qu'au paragraphe 4. suivant sur les prédicteurs d'état.

ii) quantifier la source émise par un site d'enfouissement en fonction des conditions météorologiques (la composition des déchets entrant est supposée ~constante, seul le fonctionnement biologique et les paramètres environnentaux perturbent le système). La production par un CSD d'une ou plusieurs familles de composants est issue de la réaction anaérobie ou aérobie de la décomposition des déchets organiques. Des modèles déterministes basés sur les lois d'écoulement dans le sol et la cinématique des gaz sont proposés mais ont du mal à être validés faute de connaître tous les paramètres intrinsèques au phénomène ; les systèmes sont donnés sous forme d'équations différentielles partielles. (Perera M.D.N et al 2002). Des modèles de type stochastique sont également développés pour évaluer le transport d'un élément trace dans un CSD (Zacharof A.I et al. 2001). Basés sur la production de composants, ils ne prennent pas en compte l'atmosphère comme un élément du système mais juste comme une condition limite de pression. Un couplage entre un CFD et un code d'écoulement hydrodynamique dans le sous-sol ne peut être envisagé faute de mesures et de la maîtrise de la physique complexe du système.

La proposition qui en découle, concerne le point ii) et serait la suivante : ne pourrait-on pas considérer le système[61] comme un ensemble de sous-systèmes dont chaque partie peut être assimilée à un système dynamique dont on pourrait connaître les variables, certaines sorties, les fonctions *f* et *h* . Un observateur local pourrait également être envisagé pour trouver les variables d'état sous réserve de l'observabilité de système.

Difficultés à surmonter sont les suivantes :

- la sélection et le passage de certaines équations EDP de l'atmosphère à des EDO
- l'identification des différents éléments du système et leurs caractéristiques : les fonctions de réponse de certaines variables comme la Température (soit en *T=f(k)*, soit en *T=z(x,y,t)),* pourraient être issues de la Reconnaissance de Forme (§2.4.3)
- la formulation d'un système d'état de chaque sous-système à partir de formes simples comme un amortissement, un ressort, dont le comportement est similaire
- la construction de la fonction de raccordement entre les différents sous-systèmes pour obtenir un prédicteur de l'ensemble, si la condition de superposition est admissible
- la faisabilité de l'algorithme et sa convergence possible d'un tel système.
- l'évaluation du réalisme de l'approche, et pour les résultats obtenus leur pertinence.

La liste des hypothèses et des difficultés est longue et la question est de savoir si une telle idée ne serait qu'utopie où contiendrait un zeste de réalisme, ce qui devra être évalué en priorité.
Cependant, si d'établir un tel système tient d'un idéal, une fois acquis, il pourrait être décliné pour apporter des éléments à:

- l'évaluation d'impacts,
- la quantification de la production d'un CSD en composant,
- l'évaluation de la fonction de production et du tarissement du gisement à échéance de 20-30 ans,
- la diminution de la contrainte de certains problèmes numériques liés aux codes de CFD.

3.4 Quelques éléments de conclusion

La faisabilité des approches proposées dans ce paragraphe 3 doit être approfondie. L'élément fédérateur et omniprésent est de savoir comment un processus de type écoulement peut être enrichi par la prise en compte d'éléments 3D obtenus soit par interpolation soit par modélisation des champs de variables du système.
Bien qu'il reste un certain paradoxe de Condorcet dans le raisonnement : on utilise la CDF pour fournir de la Connaissance, cette connaissance est injectée à une autre modélisation pour produire une prédiction et celle-ci se substituerait à la prévision idéale d'un modèle déterministe de type CFD. Question : a-t-on intérêt à utiliser tout ce cheminement ?
Que l'on utilise ou non les résultats de la simulation du scénario *espace*, l'approche du processus par un système d'états, ainsi que celle utilisant les multi-modèles restent valables sur des données mesurées par un réseau.

[61] Dans le cas d'un CSD, le système dynamique est composé : d'un site d'enfouissement, sorte de "bio réacteur" déchets, d'une situation météorologique locale ,d'une activité industrielle de production de biogaz liée à l'enfouissement des déchets. Un vecteur d'état associé à ce système pourrait avoir comme variables d'état x (t) [température des alvéoles, l'humidité, la pression en CH4, l'oxygène dissous, …]^{T} , les sorties y(x) seraient la concentration en CH4, l'émission de certaines molécules odorantes…les variables d'entrée u(t) [la radiation solaire, le vent synoptique (direction, vitesse), la pluie, la pression atmosphérique…] ^{T}. Une commande pourrait être associée comme la dépression dans l'alvéole, la couverture du site ..

4. Prédicteurs d'états

Cette étape est consacrée à l'évaluation de l'état futur d'un processus physique agissant sur un environnement *anthropisé*. Les deux cas d'étude concernent d'une part le processus qualité de l'air lié à l'activité d'un centre de stockage de déchets, et d'autre part l'analyse des dysfonctionnements d'un réseau d'AEP.
Bien que le cadre conceptuel ait été justifié au volet 2, il s'agit de formaliser le problème et d'envisager une solution de résolution. En réalité, ces approches se généralisent à d'autres applications en Sciences de l'Environnement présentées au paragraphe 5.

4.1 Prévision d'état pour un système échantillonné sous forme de classes

Après avoir identifié la forme, et le vecteur d'état d'un scénario jour *temps* et *espace*, la prévision au lendemain *j* ou *j* + *n*, comme la qualité de l'air du lendemain, la possibilité d'avoir une plainte ou une odeur le lendemain, le dépassement d'un seuil…sont les objectifs de la recherche.
Le scénario ne restitue localement que certaines variables d'état du système (§3.3.3) notamment pour la composante atmosphère sur un jour. On utilise des modèles explicatifs lorsque la modélisation déterministe par EDP qui ne peut prévoir à court terme une caractéristique, atteint ses limites. Les difficultés rencontrées sont celles liées aux données incomplètes et de type booléen de la plainte (indique ~ présence d'une odeur), à l'absence d'une mesure en continu d'un indicateur d'odeur (une concentration d'un composant particulier).
On peut schématiser l'approche retenue par le clustering (volet 2.) (figure 3.4.1) :

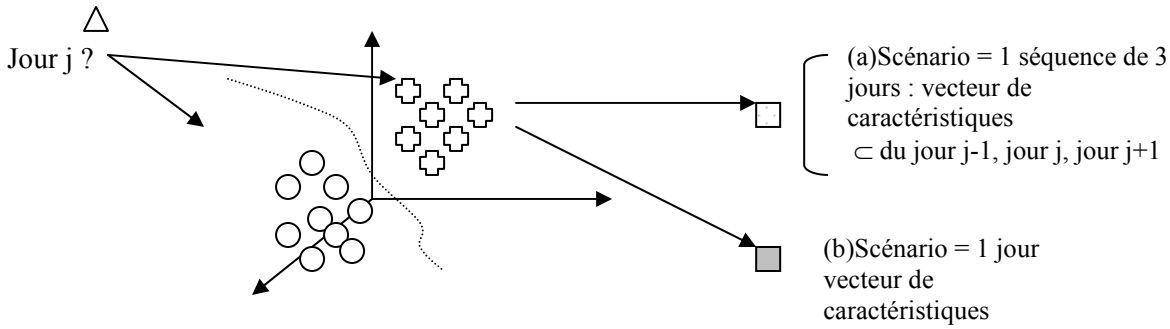


Figure 3.4.1. Discrimination du jour *j* à partir d'un ensemble de classes

Un jour est un objet d'une classe appelée scénario jour. La prévision consiste à instancier l'apparition d'un nouveau jour *j* à une classe. Une reconnaissance de forme classique (kmeans, kmeans floue) appliquée directement identifierait effectivement la classe la plus proche du jour mais n'assurera pas d'obtenir le jour suivant, sauf si l'apprentissage a été formulé pour identifier le lendemain j+1.

Si on dispose d'une grandeur du jour j+1, comme la température au lendemain (source Météo France) la fiabilité de la classification peut être accrue en utilisant la théorie des champs verticaux (Pearson D.W, Batton-Hubert 2005) pour prévoir la concentration en ozone.

Aucune grandeur ne peut a priori informer les variables d'état au lendemain : la reconnaissance doit intégrer la séquence des jours scenarii.

Deux cas sont envisageables (figure 3.4.1) :

i) dans un scénario j*our* moyen sont intégrés non plus un jour mais 3 jours (ou *n* jours) { j, j+1, j-1}(figure 3.4.1a). On peut espérer obtenir des classes qui intégreraient cette séquence. Lors du calcul de la similarité, les valeurs du noyau de la ou des classes (*i.e* une classification floue) concernées donnent une estimation du jour j +1 : on obtiendrait un prédicteur simple mais dont l'efficacité doit être vérifiée sur des outils classiques. Il semble pertinent de passer à un scénario temps, la classification doit s'attacher à classifier des courbes en considérant soit, la série de données comme une fonction par l'analyse fonctionnelle, soit en conservant la série par des méthodes multivariées adaptées (paragraphe 2.3) .

ii) le vecteur de caractéristiques est celui du jour j : le scénario jour du lendemain doit être prévu à partir de quelques observations (figure 3.4.1b).

Les réseaux bayésiens permettent de formaliser une probabilité conditionnelle d'apparition d'une odeur : les premiers tests ont mis en évidence la non résolution de l'identifiabilité du modèle avec l'état 'type de temps' pour lequel le réseau doit évaluer sa probabilité conditionnelle (cf. volet 2. chapitre 2. §2.3). Cet état étant déduit, il est nécessaire de reconsidérer le graphe associé afin de contourner cette non évaluation.

Le "type de temps" représente ce que toutes les observations, l'analyse multivariée et la classification cherchent à obtenir ou à déduire à partir des données disponibles. Ce *type de temps* caractérise effectivement l'état du système à l'origine du basculement des odeurs en fonction des vents. Il est conditionné par l'état du système du jour d'avant. Les chaînes de Markov à temps discret seraient effectivement capables de calculer la probabilité d'état dans lequel se trouve le système le lendemain, consécutivement à l'observation d'une séquence donnée. Un modèle rustique et classique serait celui où l'on peut définir N états avec : N ={ nuageux, ensoleillé, pluvieux}, et une matrice de transition de dimension 3 × 3. La probabilité des états peut être obtenue en utilisant comme loi de densité de probabilité, l'histogramme des fréquences des observations.
La principale difficulté est que l'on ne mesure pas un état ou un *type de temps*, mais des observations de l'état que l'on cherche à identifier. Les chaînes de Markov cachées sembleraient plus adaptées, les observations ne sont plus attachées à un état. A chacun des états possibles doit être associée une distribution de probabilité pour chaque observation envisagée. La difficulté réside dans la construction des paramètres du modèle : quels sont les états [62] possibles ? Les observations, ne seraient-elles pas plutôt de type continu ? Comment calculer la probabilité de transition entre les états et la matrice des probabilités associées à chaque observation dans chacun des états ? La mise en oeuvre pour une classification automatique devra résoudre l'obtention des probabilités a priori des observations, ou alors les transcrire en probabilité a posteriori et convertir la séquence d'observations en séquence d'états (Gosselin B, 1995).

Evaluer par une chaîne de Markov, la probabilité de l'état du système scénario est envisageable, certaines questions demeurent: quels seront les résultats ? peut-on adapter cette approche pour identifier un état "scénario mal odorant" ?

[62] Par exemple : les observations possibles {température, pluie, pression, vent} Etats du système : [anticyclonique, dépression, instabilité]

4.2 Estimation événementielle d'un processus mal échantillonné et mal fonctionnant

Rappelons le contexte : soit un système mal échantillonné, un réseau d'adduction en eau potable pour lequel il existe une information qualitative qui peut accroître la connaissance du système (volet 2. chapitre 1).

L'objectif consiste à identifier une donnée spécifique qui soit la manifestation d'un comportement défectueux du système [SI inondation ($h \rightarrow \infty$) ALORS $\exists$ rupture de canalisation ou d'ouvrage]. L'information indirecte des manifestations de dysfonctionnement sont les observations et les interventions sur le réseau (figure 2.1.1). Une analyse qualitative et quantitative des aléas survenus sur le réseau uniquement par le biais de leur manifestation (plaintes et interventions) a permis de proposer une analyse spatiale et temporelle des dysfonctionnements hydrauliques sur le réseau (volet 2. chapitre 1 §2.).

L'algorithme proposé est une réduction itérative d'une liste d'évènements répondant à des critères chronologiques puis topologiques et hydrauliques. Le choix de la fenêtre temporelle pour la chronologie et de la surface minimale de la corrélation spatiale et topologique doit assurer la non exclusion d'évènements liés. Une approche de type causalité avec des graphes de Pétri permettrait d'éviter d'obtenir une liste hétérogène d'évènements et permettre de détecter les causes possibles en mode diagnostic, ou les évènements possibles en mode prédictif.

Considérons une base d'évènements importante où un évènement est associé à un vecteur de caractéristiques, dont certaines variables sont la date d'apparition, la durée de l'évènement, la localisation sur le réseau.

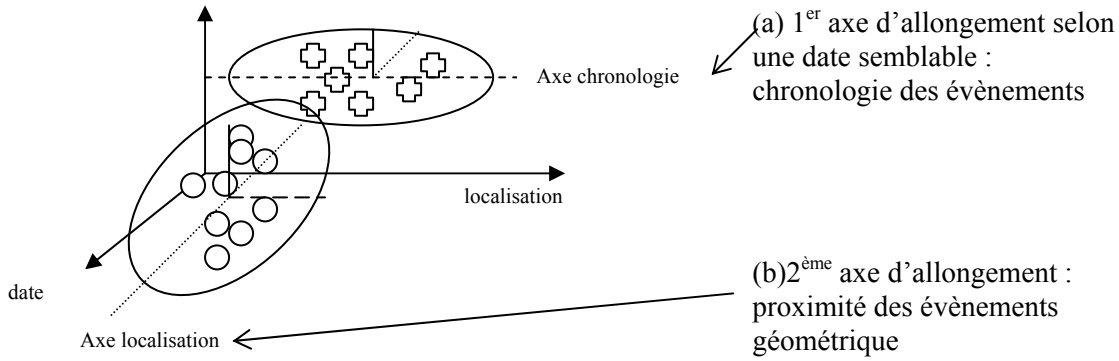


Figure 3.4.2. Analyse multivariée des évènements produits sur un réseau d'AEP

Une segmentation multivariée regrouperait des évènements selon une similarité ou distance. La synchronisation d'évènements se matérialise par un allongement préférentiel du nuage d'évènements selon un axe parallèle à celui de la localisation (figure 3.4.2a). Une proximité géométrique entraîne un allongement du nuage selon une direction parallèle à l'axe représentant la date (figure 3.4.2b). Une autre composante du nuage pourrait être liée au type d'évènement (trop d'eau, pas d'eau). Les axes de chaque nuage ainsi que la variance de la classe renseigneraient sur les évènements regroupés dans cette classe. L'hypothèse est alors la

suivante, compte tenu des classes d'évènements constituées, il serait envisageable d'instancier ces sous-ensembles respectifs aux nœuds d'un graphe de causalité puis de propager chacun des sous-ensembles. Cette phase de construction de classes constituerait alors un pré traitement pour trouver une corrélation entre chaque évènement. La difficulté est liée à la relation topologique du réseau : soit un évènement est positionné en amont ou en aval, soit une configuration de réseau maillé qui est contraint par une distance de voisinage. Il faudrait trouver un moyen d'exprimer dans une métrique les 3 variables caractéristiques cette composante. Ce qui semble un véritable défit puisque nous ne pouvons opposer topologie et métrique. Serait- il possible de trouver une transformation de l'espace qui respecte les propriétés du graphe du réseau ?
Ces difficultés font que le clustering peut se limiter à regrouper les évènements selon leurs caractéristiques temporelles ; cette segmentation non supervisée n'impose pas de fixer une taille de fenêtre d'exploration a priori.

5. Autres projets de recherche développés

Bien que la contribution principale de ce travail concerne les milieux continus (eau, air), je me suis également intéressée au problème des données discrètes, concernant l'optimisation des flux. Cette interopérabilité entre un SIG et un modèle de répartition optimale des flux, constitue un champ d'application de la Recherche Opérationnelle appliquées à l'Environnement.

· Plans départementaux d'élimination des déchets (PDED) et optimisation :

La loi de 1992 relative à la gestion des déchets édicte le principe de proximité : « réduire en quantité et en distance le transport des déchets » et confie aux Départements la mission d'organiser la gestion des déchets banals d'origine ménagère ou industrielle, au travers de Plans Départementaux promulgués par arrêté préfectoral. Un raccourci fréquemment emprunté conduit hâtivement à conclure qu'il convient par conséquent de traiter les déchets à l'intérieur strictement des limites géographiques de chaque département et à l'exclusion des productions de déchets issues du reste du territoire national et en particulier des départements limitrophes. Or la mise en œuvre effective de cette approche autarcique conduit ou conduirait, paradoxalement, à accroître significativement et pour une durée indéterminée les flux globaux de transports de déchets à l'échelle régionale. Il s'agit de réaliser une analyse objective des Plans Départementaux et d'évaluer l'impact des politiques libérales ou autarciques à partir de critères environnementaux (transports, rationalisation des sites de traitement, efficacité, économies d'échelles).
Les données de base sont la localisation cartographique des bassins de production régionaux (ville et syndicat) et des sites de traitement de déchets. L'approche choisie construit un ensemble de configurations actuelles et futures des schémas d'acheminements possibles des déchets entre la source - les grands bassins de production (OM et DIB) et les centres de traitement des déchets. L'échelle est celle d'une région (Région Rhône-Alpes et 8 départements). Il s'agit d'établir et d'optimiser un indicateur de proximité, un coût en kilomètres × tonnes. Ce problème d'optimisation peut se ramener à problème de type LP sous certaines conditions, qui peut évoluer vers un système non linéaire. La recherche d'un référent pour l'optimisation entre une demande et une offre, reste ouverte. L'introduction de contraintes supplémentaires et de variables possibles permet d'explorer un champ d'heuristiques comme les algorithmes génétiques. Un élargissement de cette problématique pour des échelles décisionnelles emboîtées, à d'autres indicateurs comme les émissions de gaz à effet de serre est envisagé : ces aspects d'optimisation multiobjectifs doivent être introduits dans un sujet de Master puis de thèse de doctorat, qui débutera en octobre 2006, sous ma direction.

Travaux réalisés : un rapport de recherche industriel (2004 - 2006)

Des travaux de projets de recherche industriels complètent cette expérience mais ne sont pas détaillés volontairement dans ce rapport :

- L'optimisation multiobjectif de type multicritères pour des données discrètes et spatialisées : Analyse du risque de transport de matières dangereuses sur le réseau routier de la Loire (1 contrat industriel et un Master en 2005).
- La gestion de l'information géographique et cartographique dans l'aide à la décision (2 articles en revue + 2 articles en colloque + 1 contribution à un ouvrage collectif)

6. Synthèse et applications en Sciences de l'Environnement

Les problèmes abordés précédemment relèvent de l'Ingénierie pour l'Environnement, auxquels on doit apporter des solutions quantitatives et effectives. Leur résolution fait émerger une nouvelle formulation liée à des disciplines actuelles en plein développement comme la Géomatique, le Data Mining et la Recherche Opérationnelle appliqués à l'Environnement. Ce domaine doit s'adapter aux nouvelles formes d'information mais aussi à l'information disponible malgré la puissance du stockage et de calcul des ordinateurs. Ceci ne doit pas nous faire oublier que cet usage est coûteux et doit être adapté au contexte et à l'échelle de l'utilisateur.
Pour cela 3 points de recherche sont à favoriser :

- conserver et exploiter les principes de la modélisation déterministe tant qu'elle ne se transforme pas en une « usine à gaz ». Il s'agit alors de fournir des indicateurs permettant de choisir le type de modélisation (analytique ou EDP) ou bien d'ajuster des lois pouvant conditionner au mieux les frontières du domaine, mais également d'intégrer des équations de transformation chimique dans les modèles eulériens ou lagrangiens,
- accroître l'opérabilité entre modélisations et outils informatiques: il est inutile de refaire un modèle ou de compliquer le système s'il existe déjà, en proposant des schémas conceptuels et fonctionnels adaptés,
- lorsque la modélisation déterministe n'est pas perfectible, il faut exploiter l'erreur ou l'imperfection dans les données pour créer de la connaissance à l'aide d'heuristiques établies sur de la donnée disponible et de méthodes dites de *data mining*.

Ce dernier point est fondamental : pour accroître la connaissance d'un système, une des solutions qui mérite d'être développée consiste à identifier l'état d'un système plutôt que le modèle déterministe et physique en lui-même. Cette approche, actuellement peu utilisée en Environnement constitue un véritable enjeu. Il s'agit alors de chercher un optimum robuste (plutôt une solution satisfaisante) pour l'identification de scénario semblable et vraisemblable.

Un effort tout particulier doit être conduit pour adapter la Reconnaissance de Formes aux données de l'espace 3D. Actuellement, la classification développée intègre des données ponctuelles localisées en un point (X,Y), à une altitude Z. Il s'agit notamment d'explorer le clustering en 3D et d'approfondir l'approche multimodèle et de systèmes dynamiques qui permettent d'intégrer la composante temporelle de phénomènes étudiés.

L'élaboration de ces méthodes et outils correspond à plusieurs préoccupations cruciales en Sciences de l'Environnement qui sont, pour les phénomènes continus eau - air :

- le choix optimal de l'emplacement de points de mesure en sachant que le coût d'acquisition de nouvelles données est élevé (stations météorologiques, station de mesure de polluant atmosphérique,
- la construction de stations virtuelles pour les secteurs faiblement échantillonnés,
- le choix de campagnes de mesure mobiles ciblées pour établir soit un poste de mesure fixe, soit une station virtuelle ,
- la corrélation entre des données de campagnes mobiles, ponctuelles et des données des réseaux d'acquisition fixes.

Les domaines d'application sont variés. On peut citer :

- la surveillance de dispositifs via des réseaux urbains linéaires ou ponctuels..

- la modification de pratiques industrielles comme la réduction des rejets, la construction d'ouvrages de protection (remblais ou levées lors d'enfouissement de déchets)…
- le renouvellement d'infrastructures comme les canalisations d'un réseau d'eau
- le calcul d'impacts sanitaires et environnentaux pour une exposition à plus ou moins long terme.
- l'aide à la décision pour l'aménagement du territoire comme le choix de développer ou non une infrastructure en terme d'impacts et de risques.

Pour un phénomène discret comme le transport des déchets[63], on citera le problème de la répartition optimale des sites de traitement et de la collecte des déchets. En effet, la question majeure qui reste ouverte est la suivante : vaut-il mieux disposer d'un centre de traitement important régional ou d'un semis d'installations sur un territoire ?

Abréviations

ACP : Analyse en Composantes Principales
CFD : Computational Fluid Dynamics
CSD : centre de stockage de déchets
EDP : équation aux dérivées partielles
EDP : équation différentielle ordinaire

[63] déchets de type ordures ménagères mais aussi déchets industriels banals

Bibliographie

Ambroise C., Govaert G., 1998, Convergence of an EM-type algorithm for spatial clustering. Pattern recognition Letter, 19(10), 919-927.

Ayala G. Epifanio I, Simo A. Zapater V. 2006 Clustering of spatial point patterns, Computational statistic and data analysis., n° 50, p. 1016 - 1032.

Bezdek J.C 1974 , Numerical taxonomy with fuzzy sets, J Math Biol ; 57-71

Bezdek J.C 1981, Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York

Bourgault G, Marcotte D., Legendre P. 1992, The multivariate covariogram as a spatial weighting function in classification methods, Math. Geol. 24: 463 - 478,

Coppi R. D'Urso P. 2005, Fuzzy unsupervised classification of multivariate time trajectories with Shannon entropy regularization, J. Computational Statistics and data Analysis, 1452-1477pp

Cover T.M, Hart P.E 1967, Nearest neighbor pattern classification, IEEE Transactions on information theory, Vol 13, N°1, janvier

Della Dora J., Gautier T. Révol Nathalie, Systèmes dynamiques, Support de cours de ENS, Grenoble, 1993).

Duda R.O, Hart P.E 1973, Pattern classification and scene analysis, John Wiley & Sons.

D'Urso 2000, Dissimilarity measures for time trajectories, J. Italian Statistics Soc.1-33, 1-31p.

D'Urso 2004, Fuzzy C-means clustering models for multivariate time varying data: different approaches Int. J. Uncertainty Fuzziness Knowledge Bases systems 12(3), 287-326p.

Harkat 2003 M.F., Détection et localisation de défauts par Analyse en Composantes Principales, Thèse de doctorat de l'Institut Nationale Polytechnique de Lorraine, 2003, 171p.

Hastie T. Stuetzle 1989, Principal curves. Journal of the American Statistical Association, vol.84, N° 406, p.502-516

Hu T., Sung S.Y. 2006, A hybrid EM approach to spatial clustering, Computational statistic and data analysis., n° 50, p. 1188 - 1205.

Kramer M.A, 1991 Nonlinear principal component analysis using auto associative neural networks , AIChE Journal, vol. 37, N°2, p.233-243.

Kung S.Y , Diamantaras K.I, 1990, A neural network learning algorithm for adaptative principal component extraction (APEX). Proceeding of the IEEE international conference on acoustics speech an signal processing, pp.861-864.

Legendre P., Legendre L. Numerical Ecology, 2nd English edition, Elsevier, Amsterdam, 2004

Legendre P. 1987 Constrained clustering, 289 -307, in Legendre P et Legendre L. Developments in numerical ecology, NATO ASI Series , Vol G-14 Springer Verlag Berlin

Miyamoto S, Mukaidono 199, Fuzzy c-means as a regularization and maximum entropy approach, in Proceeding of 7th international Fuzzy systems association world congress IFSA'97, II, pp 86-92.

Mourot G, Ragot J. 1997, Identification de modèles de Takagi-Sugeno : application à la modélisation de la concentration d'ozone. Journal européen des systèmes automatisés, vol. 31, n°9-10, pp.1587-1608.

Oja E. 1989, Neural networks principal components and subspaces. International Journal of Neural Systems, vol. 1, pp 61-68.

Perera M.D.N, Hettiaratchi J.P.A, Achari G., 2002, A mathematical modelling approach to improve the point estimation of landfill gas surface emissions using the flux chamber technique, J. Environn. Eng. Sci. 1, 2002, 451-463

Quach R.P, 2002 Identification d'un modèle flou appliquée à un problème de classification, Thèse de l'Université Jena Monnet, Saint-Étienne, 159 p.

Ragot J. 2005, Représentation de systèmes par multimodèles ou modèles flous, Notes de cours, INPL, 20 p.

Ramsay J.O, Silverman B.W, 1997, Functional data analysis , Springer.

Rubner J. Tavan P., 1989, A self organizing network for principal component analysis. Europhysics Lettre vol. 20, pp 693-698.

Sokal R.R. 1986, Spatial data analysis and historical processes, 29-43 : in: E. Diday et al. Data analysis and informatics, IV, North -Holland , Amsterdam,

Sokal R.R., Oden N.L, Thomson B.A 1997, A simulation study of micro evolutionary inferences by spatial autocorrelations analysis, Biol. J. Linn Soc. 60 : 73 - 93p.

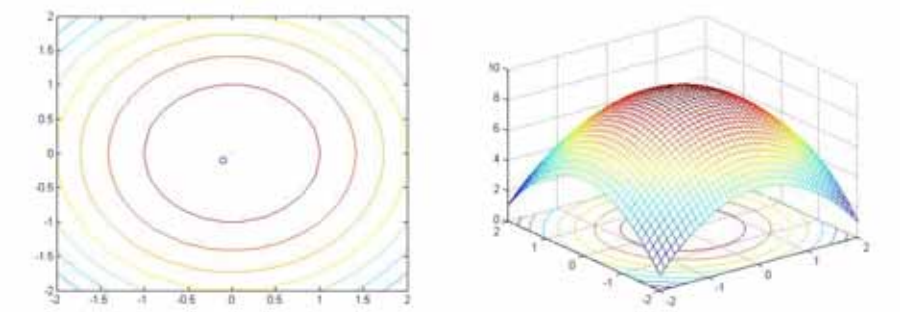
Takagi T., Sugeno M. 1985,Fuzzy identification of systems and its application to modelling and control, IEEE Trans. On systems Man and cybernetics, vol 15, p. 116-132, 1985.

Zacharof A.U, Butler A., 2001, Application of a stochastic flow and transport model for leachate production to tracer test data, in Proceedings of 5th International Waste management and landfill symposium of Sardinia,, Italy, 1-5 October 2001,

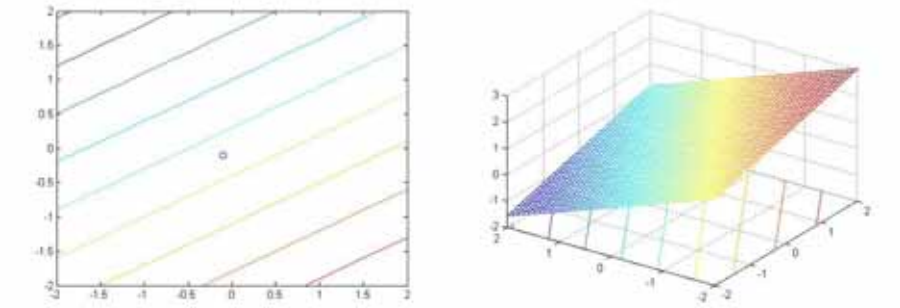
Zadeh L.A, 1965, Fuzzy sets , Information and control, 8 : 338-353 1965.

Annexe 3.1 : Jeu de 6 scénarii : 6 formes sont définies par 3 variables X1=T°, X2=Pression, X3=Concentration

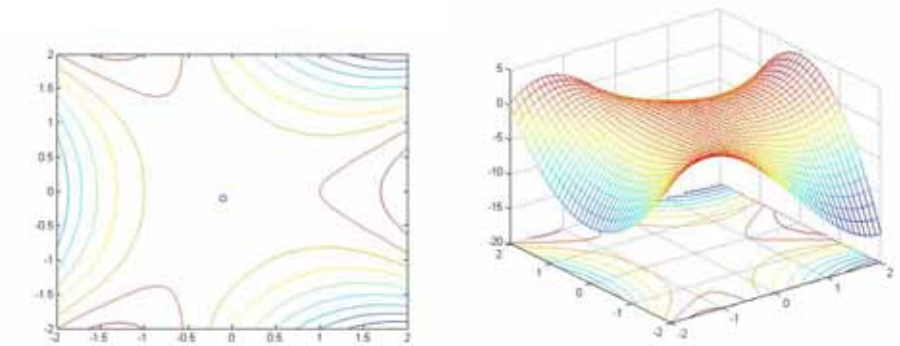
forme 1 Dome



X1

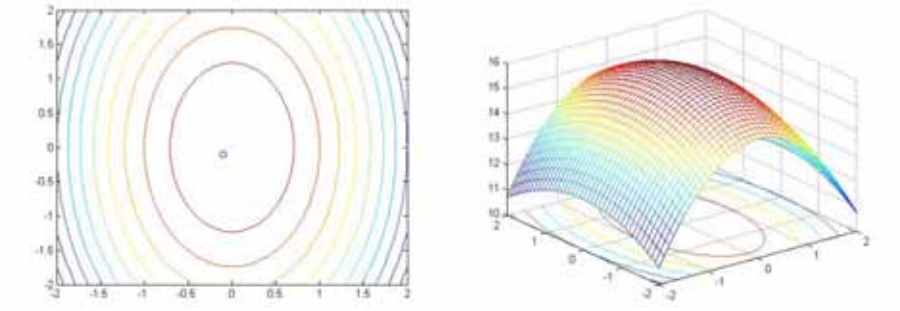


X2

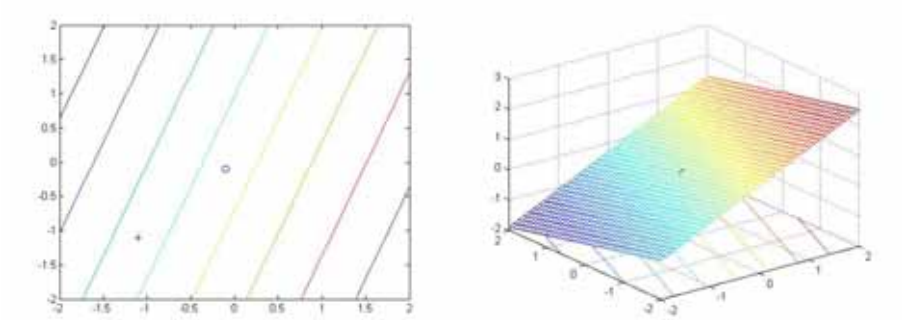


X3

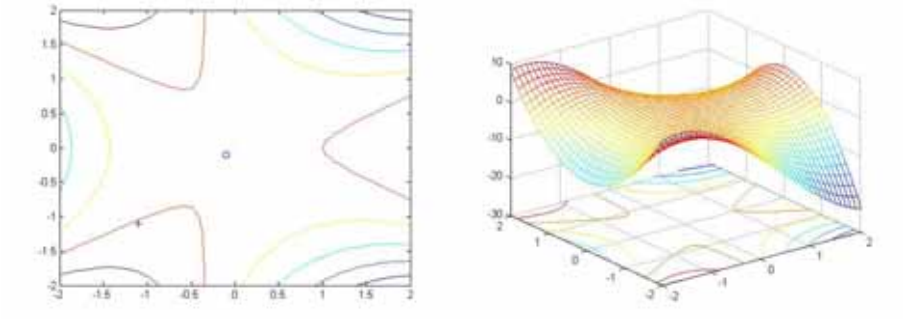
forme 2



X1

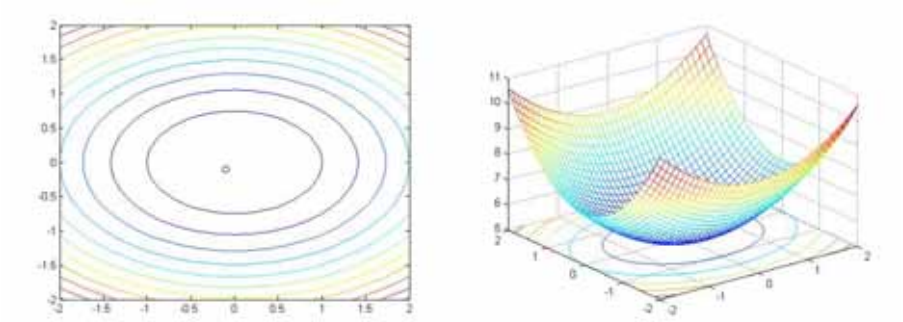


X2

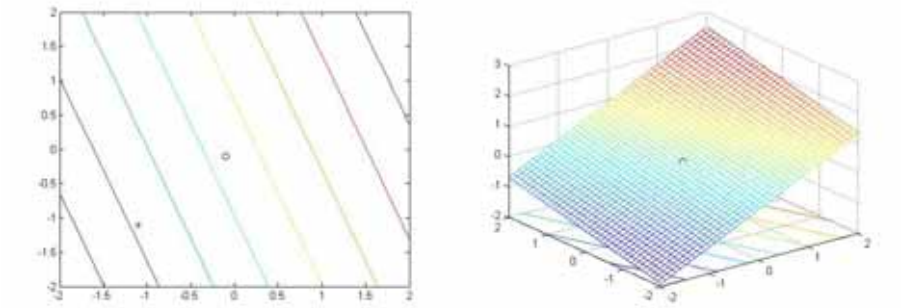


X3

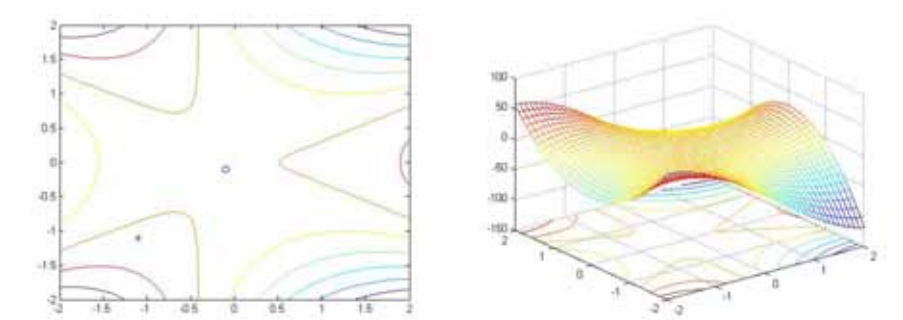
forme 3



X1

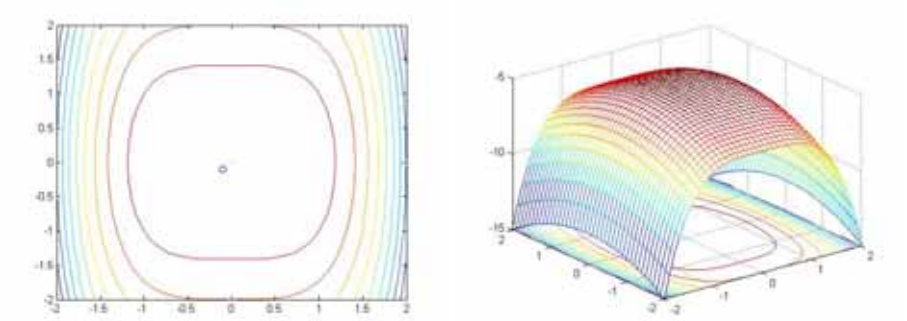


X2

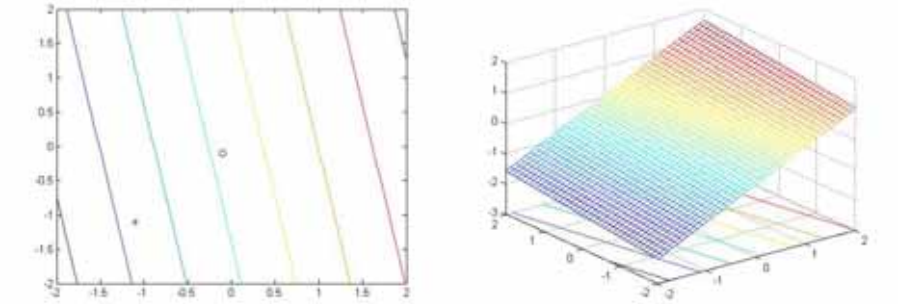


X3

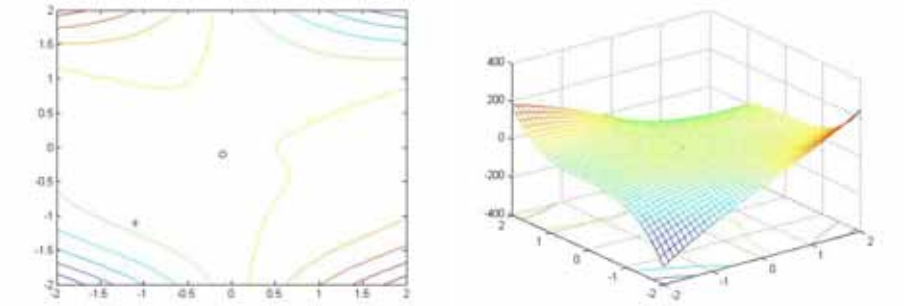
forme 4



X1

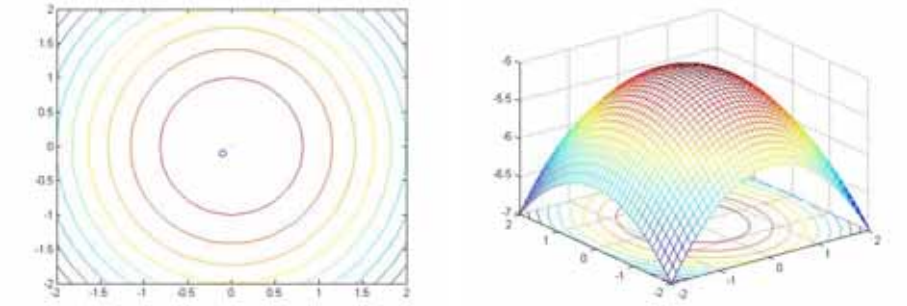


X2

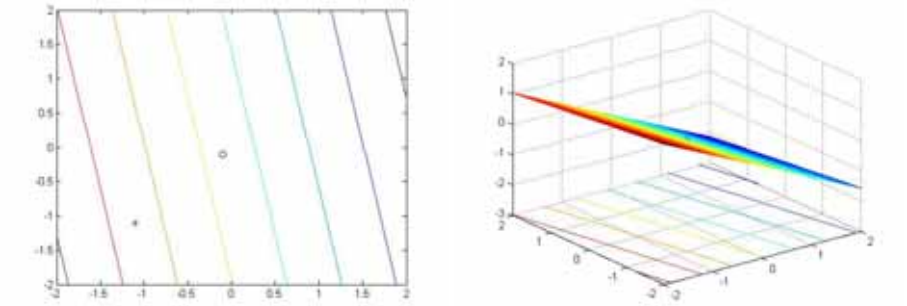


X3

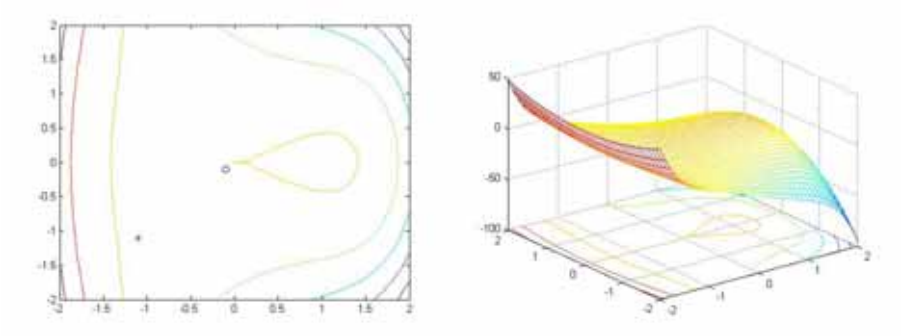
forme 5



X1

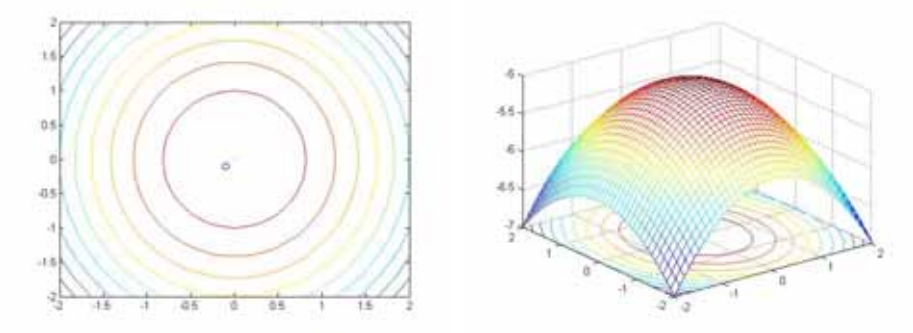


X2

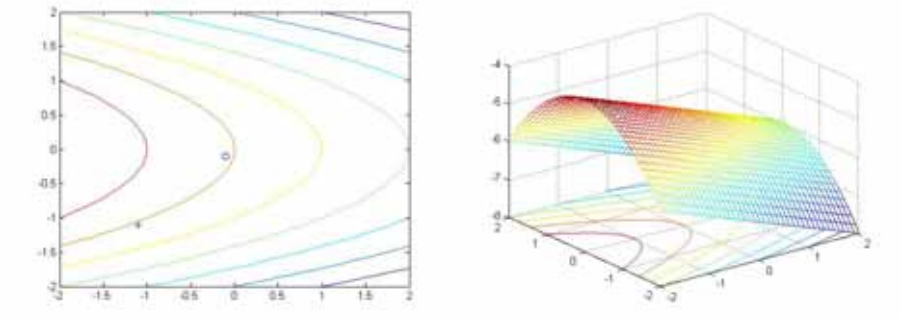


X3

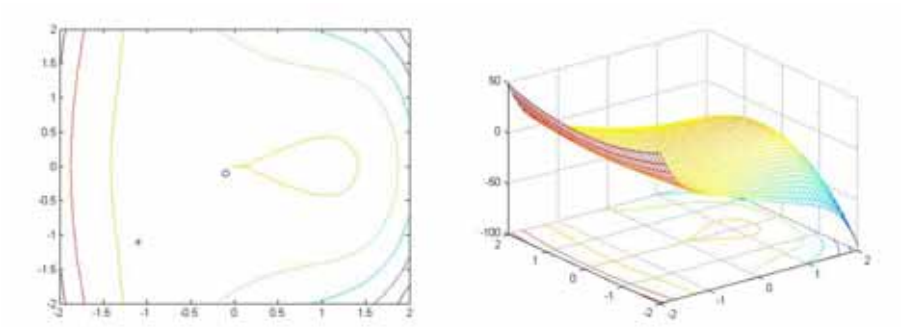
forme 6



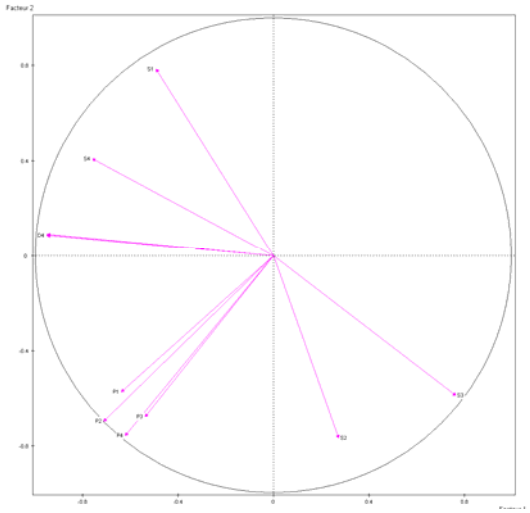
X1



X2



X3



- Résultats de la classification hiérarchique pour ces 6 scenarii par segmentation spatiale

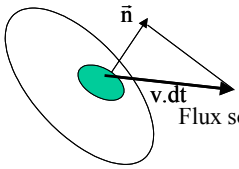
```
CLASSIFICATION HIERARCHIQUE (VOISINS RECIPROQUES)
SUR LES     6 PREMIERS AXES FACTORIELS
DESCRIPTION DES NOEUDS
 NUM. AINE  BENJ   EFF.    POIDS    INDICE   HISTOGRAMME DES INDICES DE NIVEAU
   7     2     1     2      2.00   0.36655   ******
   8     3     7     3      3.00   0.67878   **********
   9     4     5     2      2.00   1.96666   ***************************
  10     6     9     3      3.00   3.13318   ******************************************
  11    10     8     6      6.00   5.85482
*************************************************************
SOMME DES INDICES DE NIVEAU =   12.00000
DESCRIPTION DES NOEUDS DE LA HIERACHIE
(INDICES EN POURCENTAGE DE LA SOMME DES INDICES :  12.00000)
+-----------------+---------------+------------------+--------------------+
|      NOEUD      |  SUCCESSEURS  |                  |    COMPOSITION     |
| NUMERO   INDICE |  AINE   BENJ  | EFFECT.  POIDS   | PREMIER   DERNIER  |
+-----------------+---------------+------------------+--------------------+
|    7      3.05  |     2      1  |     2     2.00   |     1         2    |
|    8      5.66  |     3      7  |     3     3.00   |     1         3    |
|    9     16.39  |     5      4  |     2     2.00   |     4         5    |
|   10     26.11  |     6      9  |     3     3.00   |     4         6    |
|   11     48.79  |    10      8  |     6     6.00   |     1         6    |
+-----------------+---------------+------------------+--------------------+

DENDROGRAMME
RANG  IND. IDEN  DENDROGRAMME (INDICES EN POURCENTAGE, DE LA SOMME DES INDICES :  12.00000
MIN =  3.05% / MAX = 48.79%)
   1  3.05 1     --+
                   |
   2  5.66 2     --*-----+
                          |
   3 48.79 3     --------*-------------------------------------------------------------------------
----------------+
|
   4 16.39 5     ----------------------------------+
|
                                                   |
|
   5 26.11 4     ----------------------------------*----------------------+
|
                                                                          |
|
   6 ----- 6     ------------------------------------------------------------*------------------
----------------*
```

L'arbre regroupe successivement :
- Le groupe 1 et le groupe 2 puis le groupe 3
- Le groupe 4 avec le groupe 5 puis le groupe 6
- Enfin les 2 dernières classes constituées

Annexe 3.2 Principales équations EDP du système de Navier-Stokes

Conservation de masse : bilan de masse



Flux sortant = ρv.n dS

$$\frac{dm}{dt}=\frac{d}{dt}\iiint\rho dV=-\iint\rho v.ndS$$

...

$$\iiint(\frac{\partial\rho}{\partial t}+\mathrm{div}\rho v)\,dV=0$$

$$\frac{\partial\rho}{\partial t}+\mathrm{div}\rho v=0$$

$$(\frac{\partial\rho}{\partial t}+v.\mathrm{grad}\rho)+\rho\mathrm{div}v=0$$

$$\frac{d\rho}{dt}+\rho\,\mathrm{div}\,v=0$$

1) 1ère équation de masse : équation de continuité

Air : fluide newtonien

$$\sigma_{ij}=\underbrace{\sigma_{ij}'-p\delta_{ij}}$$

$$\sigma_{ij}'=\eta(2e_{ij}-\frac{2}{3}\delta_{ij}e_{ll})+\zeta(\delta_{ij}e_{ll})$$

Déformation = dilation + cisaillement

$$[\mathrm{div}\sigma']=\frac{\partial\sigma_{ij}'}{\partial x_j}$$

$$[\mathrm{div}\sigma']=\eta(\frac{\partial}{\partial x_j}(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}))+(\zeta+\frac{1}{3}\eta)\frac{\partial}{\partial x_j}(\frac{\partial v_l}{\partial x_l})$$

$$\mathrm{div}[\sigma']=\eta\,\Delta v+(\frac{\eta}{3}+\zeta)\mathrm{grad}(\mathrm{div}v)$$ Tenseur de force de surface

2)Fluide newtonien

Bilan de l'énergie :échelle micro-météorologique

Variation de l'énergie

$$\rho(\frac{\partial U}{\partial t}+u_j\frac{\partial U}{\partial x_j})=\kappa\frac{\partial^2 T}{\partial x_j\partial x_j}-p\frac{\partial u_j}{\partial x_i}+\Phi+Q$$

Source

Conductivité thermique : loi d Fourier

Travail lié aux Forces de pression

Chaleur dissipation visqueuse

3)Equation de l'énergie

Loi d'état du gaz F(P,T,ρ)

gaz idéal

$$P = \frac{\rho R T}{M_a}$$

Energie interne à un gaz

$$U = c_p \rho T$$

c_p *capacité thermique de l'air*

Champs de pesanteur

$$P = \rho g z$$

1ère loi de la thermodynamique

$$dP = c_p \rho \, dT \quad \left(\frac{dP}{dz} = -\rho g\right)$$

$$T = T_0 \left(\frac{P}{P_0}\right)^{R/c_p}$$

Relation (T, z)

$$T = T_0 + \frac{g}{c_p} \Delta z$$

Approximations de Boussinesq

$$\frac{\partial p_e}{\partial x_1} = \frac{\partial p_e}{\partial x_2} = 0 \qquad \frac{\partial p_e}{\partial x_3} = -\rho_e g$$

$$\frac{\partial^2 T_e}{\partial x_3^2} = 0$$

$$p_e = \frac{\rho_e R T_e}{M_a}$$

$$T_e = T_0 \left(1 - \frac{x_3}{H}\right)$$

4) Loi d'état : gaz parfait

Équation du mouvement = équation Navier-Stokes

$$\frac{d}{dt}\left[\iiint_V \rho v \, d\tau\right] = \iiint_V \rho f \, d\tau + \iint_S [\sigma].n \, dS$$

Forces appliquées au volume :

Force de pesanteur

force de Coriolis

Forces appliquées a la surface :

Tenseur de force de surface

Pression et viscosité

Équation du mouvement : eq. Navier- Stokes

$$\rho \frac{\partial v}{\partial x_j} + \rho (v.\text{grad}) v = \rho f - grad\ p + \eta \Delta v + \left(\frac{\eta}{3} + \zeta\right) \text{grad}(\text{div}\, v)$$

Force extérieure

force de pression

Variation explicite de la vitesse/temps

Forces de viscosité

Variation exploratoire de la vitesse de la particule /mouvement

5) Etablissement de l'équation de conservation de la quantité de mouvement

Équation du mvt. (échelles synoptiques)

- On s'intéresse à des mouvements atmosphériques à moyenne ou grande échelle, $L >> H \Rightarrow w << u$ et v, d'après l'équation de continuité
  - → Mouvements " quasi-horizontaux "
- Si, de plus, on considère $L << a$ (= rayon de la planète) $\Rightarrow \frac{dV}{dt} = \frac{du}{dt} i + \frac{dv}{dt} j + O\left(\frac{u^2}{a}, \frac{v^2}{a}, \frac{uv}{a}\right)$
- On définit le paramètre de Coriolis $f = 2\Omega \sin\theta$

$$\Rightarrow \boxed{\frac{dV}{dt} = fV \wedge k - \frac{1}{\rho}\nabla p + \frac{F}{\rho}} \quad (13)$$

- **Approximation géostrophique**
  - Mouvements à grande échelle, loin de la surface ($F \approx 0$), avec faible rayon de courbure → $dV/dt \approx 0$

$$\Rightarrow \boxed{fV \wedge k = \frac{1}{\rho}\nabla p} \quad (14a)$$

→ gradients horizontaux de pression compensés par les vents " géostrophiques " dus à la force de Coriolis

$$\boxed{fv = \frac{1}{\rho}\frac{\partial p}{\partial x}} \quad (14b)$$

$$\boxed{fu = -\frac{1}{\rho}\frac{\partial p}{\partial y}} \quad (14c)$$

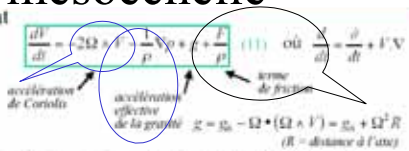
- Approximation valable sur Mars et la Terre à $z \geq 1$ km, $|\theta| \geq 10°$; sur les planètes géantes, $|\theta| \geq 5°$; pas valable sur Vénus ou Titan où $f = 0$

6) Equations de mouvement pour l'échelle synoptique

Équation mvt. échelle synoptique -mésoéchelle

Equations du mouvement

- Navier-Stokes :



Ω = vitesse angulaire de la planète

- Le terme de friction F/ρ est lié au cisaillement de l'écoulement ; il est parfois paramétrisé linéairement par $-V/\tau_F$ (friction " Rayleigh ")

- Continuité :

si fluide incompressible → ρ constant → $\nabla \cdot V = 0$

$$\boxed{\nabla(\rho V) + \frac{\partial \rho}{\partial t} = 0} \quad (12)$$

- En général, on suppose l'incompressibilité sauf, verticalement, pour tenir compte de la poussée d'Archimède → *approximation Boussinesq*

- Coordonnées sphériques

$x > 0$: est	vitesse : u (vent zonal)	
$y > 0$: nord	v	
$z > 0$: haut	w	

7) Equations de la dynamique de l'atmosphère échelle mésoéchelle

Équation de mouvement : approximation de Boussinesq

$$\rho\left(\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j}\right) = \frac{\partial}{\partial x_k}\left[\mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right) - \left(p + \frac{2}{3}\mu\frac{\partial u_j}{\partial x_j}\right)\delta_{ij}\right] - \rho g \delta_{3i}$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho_0}\frac{\partial \tilde{p}}{\partial x_i} + \frac{\mu}{\rho_0}\frac{\partial^2 u_i}{\partial x_j \partial x_j} + \frac{g\tilde{T}}{T_0}\delta_{i3}$$

Gradient de pression

Forces de viscosité

Force d'accélération /température de la particule mouvement ascendant (accélération verticale)

8) Approximation de Boussinesq

Bibliographie de l'auteur

Hubert M.1993 *Intégration d'une simulation spatio-temporelle à un modèle Topologique et Numérique de Terrain.* Thèse de Doctorat de l'Université, Pierre et Marie Curie, Paris 6, 781p.

Hubert M 1993, *Mise au point d'une organisation de données multi fonctionnelle pour la conception , la gestion et l'exploitation des réseau d'assainissement*, Rapport de DEA Méthodes de conception en bâtiment, Aménagement et Techniques urbaines, INSA Lyon, 71p.

Articles en ouvrages

Riesenmey C. , Chemel C., Vaillant H., Batton-Hubert M.	2006	Assessing population exposure to odorous pollution from a landfill site over complex terrain, In *The use of GIS in climatology & meteorology-COST-719,* Hermès Science London édition, à *paraître, fin 2006*
Batton-Hubert M., Despouy T., Vié J.	2006	Analyse du risque de transport de matières dangereuses (TMD) à l'échelle départementale - application au transport routier sur le département de la Loire, in *Aide à la décision pour l'aménagement du territoire*, ed Hemès IGAT, 2006
Batton-Hubert M., Déchomets R., Graillot D.	2006	Bases théoriques des méthodes et outils d'aide à la décision territoriale - conditions d'utilisation pratiques, in *Aide à la décision pour l'aménagement du territoire,* ed Hemès IGAT, 2006
Pearson D.W, Batton-Hubert M.	2005	Improved clustering by rotation of clusters centres, in *Adaptative an Natural Computing Algorithms*, B Ribeiro et al(eds), (ed) SpringerComputerScience, 2005, p.136-139.
Pearson, D.W. Batton-Hubert, M. Dray, G.	2003	Vertical Vector Fields and Neural Networks: an Application in Atmospheric Pollution Forecasting, in *Artificial Neural Nets and Genetic algorithms*, (ed) SpringerComputerScience, 2003, p.91-94
Pearson, D.W. Batton-Hubert M., Dray G.	2003	The use of vertical fields and neural networks in Atmospheric Pollution Forecasting . A paraître in *Systems Analysis - Modelling - Simulation*, (2003).
Mounier G. Couach O. Batton-Hubert M. Clappier O.	2002	Eulerian photochemical modeling - multinesting approach -The Rhône-Alpes case study, Air pollution Modelling and simulation, Special issue, B. Sportisse (ed), Springer - Verlag publ., 2002, p.42-49

Articles en revue

Pearson D.W Batton-Hubert M.	2005	Atmospheric pollution forecasting : Applications of Horizontal and Vertical Vector Fields. *Revue des systèmes - Journal Européen des Systèmes Automatisés* - volume 39 n°4/2005, Hermès Sciences - Lavoisier
Déchomets R. Batton-Hubert M. Graillot D.	2005	Inondations : les plans communaux d'interventions graduées ou l'importance des démarches locales, *Revue Préventique* - n° 82, juillet-août 2005, 41-48p.
Batton-Hubert M. Mimoun D	2000	Modélisation « emboîtée » et analyse spatiale du comportement hydrodynamique d'un système alluvial. Cas des épisodes de crue de la Loire. *Revue Internationale de Géomatique, Volume 10 - n°3-4/2000* , 313-333p.
Roche V. Batton-Hubert M	2000	Gestion de l'ambiguïté et de l'incertitude : ses implications théoriques dans le développement d'un système d'information à référence spatiale. *Terminal : technologie de l'information, culture et société* , 2000, n° 82.
Batton-Hubert M. Pauze A.	1998	Apports des Systèmes d'Information géographiques pour la protection des ressources en eaux. Evaluation simultanée des impacts de pollution et du dimensionnement des zones de protection autour des captages d'adduction en eau potable. *Revue Internationale de Géomatique,* volume 8 - n°3/1998, 13 - 32 p.
Roche V. Batton-Hubert M.	1998	Pratiques des SIG en aménagements du territoire : les nouveaux usages de l'information géographique, *Revue Internationale de Géomatique,* volume 8 - n°1-2/1998, 9 - 25 p.
Roussel I. Rouchouse S. Batton-Hubert M.	1997	Pollution atmosphérique et climat local ; l'exemple de l'agglomération stéphanoise en 1997, *Revue de Géographie de Lyon* , Vol 724/97, 315-321p.

Colloque à comité de lecture

Tacnet J.M, Escande JM., Batton-Hubert M.	2006	Methodology of snow avalanches post-event field investigations : tools, difficulties and perspectives, in *Proc. International Snow Science Workshop* 2006, Telluride - Colorado (USA)
Riesenmey C. , Chemel C, Vaillant H., Batton-Hubert M.	2006	Assessing population exposure to odorous pollution from a landfill site over complex terrain, In *Proc. of the COST-719 Conference 'The use of GIS in climatology & meteorology*, Grenoble, France
Graillot D., Batton-Hubert M.	2006	GIS and geodatabases for groundwater flow modelling, in *Proc. 5th International Conference on Analytic Element Method ICAEM*,Manhattan, Kansas State, USA
Riesenmey C. Chemel C. Batton-Hubert M. Chollet J.P	2005	Mixed method to assess odour impact using data classification and high-resolution numerical simulations. In *Proc. of the 5th International Conference on Urban Air Quality*, Valencia, Spain
Chemel, C. C. Riesenmey J.-P. Chollet M. Batton-Hubert	2005	High-resolution large-eddy simulations for odour-impact assessement. In *Proc. of the 2nd General Assembly of the European Geosciences Union*, Vienna, Austria.
Chemel, C. Riesenmey C. Chollet J.P Batton-Hubert M.	2005	Characterization of odour emissions from a landfill through numerical simulations. In *Proc. of the 5th Annual Meeting of the European Meteorological Society*, Utrecht, The Netherlands.
Riesenmey C. Chemel C. Batton-Hubert M. Chollet J.P	2005	Influence of the topography on odour dispersion from a landfill located in complex terrain. In *Proc. of the 5th Annual Meeting of the European Meteorological Society*, Utrecht, The Netherlands.
Mimoun D. Batton-Hubert M.	2004	Analyse multicritère spatiale pour l'évaluation du potentiel écologique des carrières en eau , *CASSINI '04* 7ème conférence du GDR SIGMA, Grenoble , 2-4 juin 2004
Devigne V. Batton-Hubert M. Clopeau T. Graillot D	2003	Simulating groundwater and vortical flows using a non-conforming approximation of Darcy'law and Stokes problem, 4th International Conference on the Analytic Element Method for the modeling of groundwater flow and application s in environmental sciences, November 20-21 2003

Batton-Hubert M. Vaillant H.	2003	Scénario climatologique et météorologique pour la semi prédiction de nuisances olfactives : application à un centre de stockage d'ordures ménagères et de déchets industriels banals, Actes du 3ème Colloque STIC et Environnement, INSA , 19-20 juin 2003, Rouen
Pearson D.W Batton-Hubert M. Dray G.	2003	Champs de vecteurs verticaux appliqués à la fiabilité de la prévision de la pollution atmosphérique par classifieur neuronal, Actes du 3ème Colloque STIC et Environnement, INSA , 19-20 juin 2003, Rouen
Gaillard E. Batton-Hubert M.	2003	Classification de données météorologiques pour l'étude de la concentration en ozone, Actes du 3ème Colloque STIC et Environnement, INSA , 19-20 juin 2003, Rouen
Batton-Hubert M. Vaillant H.	2003	Identification de scénarii « type » d'apparition de nuisances olfactives pour la prescription de recommandations lors de l'exploitation d'un centre de stockage d'ordures ménagères et de déchets industriels banals : réduction de la production d'odeurs sur une alvéole en exploitation, Actes du 10ème Colloque EURODEUR, 25-26 Juin 2003, Evreux
Pearson D.W. Batton-Hubert M. Herrera Garcia G	2002	Predicting Ozone Peaks : A combined CBR and cell mapping approach, In Proceedings of *Meeting of the International Environmental Modelling and Software Society*, Lugano, Switzerland, (2002).
Batton-Hubert M. Mimoun D.	2001	Mesures en gravières par analyse de scénarios hydrologiques : évaluation et suivi des impacts hydrauliques sur un site post industriel, Congrès annuel de Société de l'Industrie minérale, Clermont-Ferrand , 9 -12 octobre 2001
Mounier G. Couach O. Batton-Hubert M. Clappier A.	2001	Photochemical eulerian modelling using multinesting methodology, application to Rhône-Alpes district*, in proceedings the 2nd International Conference on Air Pollution Modelling & Simulation* , INRIA-ENPC, Champs sur Marne, France, April 9-12 2001.
Roche V. Batton-Hubert M. Déchomets R.	2000	4th International Symposium on Spatial Accuracy Assessment in Natural Ressources and Environmental Sciences, Amsterdam (Hollande), 12-14 juillet 2000Ambiguity and uncertainty in GIS design.

Batton-Hubert M.
Mimoun D.
2000 Modélisation « emboîtée » et analyse spatiale du comportement hydrodynamique d'un système alluvial. Cas des épisodes de crue de la Loire. Journées de la recherche CASSINI, La Rochelle, 7-9 septembre 2000

Batton-Hubert M.
Roelens M.
Piatyszek E.
2000 *Evaluation d'un modèle d'émissions liées au trafic automobile en milieu urbain par analyse spatio-temporelle du comportement d'un station de proximité*. Journées thématiques « Automatique et Environnement », Nancy, 9-10 mars 2000

Batton-Hubert M.
Roelens M.
2000 *Modelization of the thermal and convective behaviour of the urban island setting with radiant energy transfer simulation using the ray tracing –application to Saint-Etienne city*, 2nd International Conference decision making in urban and civil engineering, Lyon, 20-22 novembre 2000, 199-213 p , vol 1.

Mounier G.
Couach O.
Batton-Hubert M.
Chappaz C.
Clappier A.
2000 Analyse spatiale de la qualité de l'air à méso-échelle à l'aide d'un modèle eulérien photochimique . Application à la région Rhône-Alpes par multinesting, Actes du 9ème colloque « Transports et Pollution de l'air », Avignon Juin 2000, Actes INRETSn°70, 375 –380 p. vol2.

Batton-Hubert M.
Madaleno E.
1999 A tool to interpret the urban atmospheric pollutant measurements based on emissions analysis and modelling. Proceedings of the sixth international conference on the harmonisation within atmospheric dispersion modelling for regulatory purposes, session 6 october 11-14 1999, CORIA-UMR 6614, INSA-Rouen, p.50.

Batton-Hubert M.
Pauze A.
1997 Evaluation et propagation du facteur d'atténuation d'une pollution sur un Modèle Numérique de Terrain. Etablissement de zones de protection autour de captages d'adduction en eau potable sur le Parc Naturel et Régional du Pilat , 4ème rencontre Hydrologique Franco-Roumaine, 2-4 Septembre 1997 -Suceava, Roumanie.

Batton-Hubert M.
1995 Valuation of risks erosion and hill slumping using a DTTM/DTM environment. "International trade fair and congress for the Geosciences and Geotechnology" GEOTECHNICA'95 (2-5 May 1995), Cologne, Allemagne, p.55-57.

Batton-Hubert M.
1994 Automated searching of crests and talwegs in DTTM/DTM GIS basis for environmental surveying. Actes du 6ème rendez-vous européen des acteurs de l'information géographique numérique EGIS/MARI'94 (29 Mars - 01 Avril 1994), Paris, Vol 1, p.371 - 381

Batton-Hubert M. 1994 Automated searching of 3D cartographic shape submitted to an expansion basis on DTTM/DTM environment. Application to GIS and environmental surveying systems. Proceedings of "Europe in transition : the context of GIS" Conference (28 - 31 Août 1994), Brno, R,publique Tchèque, p.41 - 51.

Batton-Hubert M. 1994 Simulation of 4D digital and topological terrain model through natural phenomena modeling. The erosion and impacts on the environmental surveying. Geo-information systems for environnent, Conférence du 125ème anniversaire de l'Institut Géologique de Hongrie (14-15 Septembre 1994), Budapest, Hongrie, p.110 - 120.

Batton-Hubert M. 1994 Automated searching of shared properties during the juxtaposition of adjacent maps. General model for management of maps continuity into cartographic databases and GIS. Proceedings of EUROCARTO XII (10 - 12 Octobre 1994), Copenhague, Danemark, p.21 - 31.

Batton-Hubert M. 1993 Expert System managing a large set of discrete processes for 3D Cartography and mapping. Proceedings of EUROCARTO XI (8-11 December 1993), Kiruna, Suède, p.67-77.

Hubert M 1991 4D Topological terrain modeling with discrete simulation applied to Environmental Impact Surveying, Proceedings of GIS International Conference (22-24 April 1991), Brno, Tchécoslovaquie, p.130 - 143.

Rapports d'étude

Batton-Hubert M. ,Vaillant H. Projet PDED, Plan d'élimination départementaux des déchets : développement durable et principe de proximité, Rapport final de contrat pour ONYX-ARA. Contrat Armines - ONYX - N°31064, Ecole nationale supérieure de mines de Saint-Étienne, Juin 2006, confidentiel.

Batton-Hubert M. ,Vaillant H. Projet PRIMO 2 (Plan de réduction intensive et de maîtrise des odeurs) - Analyse des tendances de variables météorologiques pour la détection des scénarii météorologiques favorisant l'apparition de nuisances olfactives sur un Centre de stockage de déchets. Rapport final de contrat pour la SATROD (groupe SUEZ). Contrat Armines - SATROD (groupe SUEZ) N° 30175, Ecole nationale supérieure de mines de Saint-Étienne, Juin 2004, confidentiel.

Batton-Hubert M. ,Vaillant H. Projet HIVERNUS - Etude du contexte de plaintes et des conditions d'apparition de nuisances olfactives liées à l'exploitation d'une plate-forme de compostage : site de Saint-Just Saint-Rambert, rapport final de contrat pour O*NYX ARA*. Contrat Armines - ONYX ARA, Ecole nationale supérieure de mines de Saint-Étienne, Juin 2004, confidentiel.

Batton-Hubert M. ,Vaillant H. Projet PRIMO 1 (Plan de réduction intensive et de maîtrise des odeurs) - Définition des conditions d'apparition de nuisances olfactives. Etude météorologique et micro-climatique pour l'identification des épisodes de nuisances olfactives. Rapport final de contrat pour la SATROD (groupe SUEZ). Contrat Armines - SATROD (groupe SUEZ) N° 11175 & 20663, Ecole nationale supérieure de mines de Saint-Etienne, Février 2003, confidentiel.

Batton-Hubert M. , Analyse spatiale et temporelle de la qualité de l'air en milieu urbain : application à l'agglomération stéphanoise, Rapport de bilan de recherche , XI ème contrat Plan Etat région programme fédérateur environnement, 2ème tranche thème 6 *Pollutions atmosphériques en site urbain*, janvier 2000, 70 p.

Batton-Hubert M. : Opération « SIG à Chisinau »- Synthèse technique : axes de développement prévisionnels - Outil d'aide à la gestion des infrastructures et services urbains - SIG pour la distribution en eau potable, rapport ENSM.SE , mai 1998.

Bacle. PY, Batton-Hubert M., Tardy A., Tidière A. : Opération « SIG à Chisinau » , Compte-rendu de voyage d'étude du 7 au 17 juillet 1997.

Batton-Hubert M., Tardy A. : Elaboration d'un Guide méthodologique de Gestion des Ressources en eau du Parc Pilat - 2ème Phase : Protection des ressources en eau souterraine : délimitation des périmètres de protection des captages et autres adductions en eau potable & Qualité des eaux de surface, rapport ENSM.SE -PNRP en collaboration avec le Ministère de l'Environnement et la Région Rhône-Alpes, février1997, 63p.

Rapports de DEA /Master

Mimoun Djamel (1999) Etude des échanges Loire/nappe alluviale à différentes échelles à l'aide d'un SIG et de modèles hydrauliques et hydrodynamiques, Rapport de DEA "Sciences et Techniques du déchet" , Ecole Nationale Supérieure des Mines de Saint-Étienne.

Blindu Igor (1999) Aide au diagnostic du réseau d'AEP pour la ville de Chisinau par analyse spatiale et temporelle des dysfonctionnements. Rapport de DEA "Sciences et Techniques du déchet " , Ecole Nationale Supérieure des Mines de Saint-Étienne.

Devigne Vincent (2002), Une condition aux limites particulière : la loi de Beavers & Joseph : illustration et application au Gourd de Villeneuve. Rapport de DESS. S.I.T.N Université Claude Bernard Lyon1.

Riesenmey Caroline (2004), Diagnostic d'un C.E.T source de nuisances olfactives, à l'aide d'une modélisation déterministe et de campagnes de mesure de C.O.V. Rapport de DEA "Sciences et Techniques du déchet", Ecole Nationale Supérieure des Mines de Saint-Étienne.

Thèses de Doctorat

P. Le Grand (2003). Formes curvilinéaires avancées pour la modélisation centrée objet des écoulements souterrains par méthode des éléments analytiques, Thèse de Doctorat de l'Ecole nationale supérieure des Mines de Saint-Etienne et de l'Université Jean Monnet, avril 2003, N°Ordre : 310I, 120p.

D. Mimoun (2004). Spatialisation de l'information : une aide à l'analyse hydraulique et paysagère développée lors de la réhabilitation des sites post-industriels - cas de réaménagements des gravières en milieu alluvionnaire, Thèse de Doctorat de l'Ecole nationale supérieure des Mines de Saint-Étienne et de l'Université Jean Monnet), février 2004, 361p.

I. Blîndu (2004). Outil d'aide au diagnostic du réseau d'eau potable pour la ville de Chisinau par analyse spatiale et temporelle des dysfonctionnements hydrauliques ,Thèse de Doctorat de l'Ecole nationale supérieure des Mines de Saint-Étienne et de l'Université Jean Monnet), mai 2004, N°ordre 336ID, 304p.

V. Devigne (2006). Ecoulements et conditions aux limites particulières appliquées en Hydrogéologie et théorie mathématique de processus de dissolution/précipitation en milieux poreux, Thèse de Doctorat de l'Ecole nationale supérieure des Mines de Saint-Étienne et de l'Université Jean Monnet), mars 2004, N°ordre 401SGE, 224p.