Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

BADJI MOKHTAR UNIVERSITY-ANNABA -

UNIVERSITE BADJI MOKHTAR-ANNABA-

جامعة باجي مختار

- عنابة -

Faculté des Sciences de l'Ingénieur
Département d'Informatique

Année : 2013-2014

THESE

Présentée en vue de l'obtention du diplôme de ***DOCTORAT en Informatique***

Contribution à la classification non supervisée : application aux données environnementales.

Option

Informatique

Par

Soufiane Khedairia

Devant le Jury

Président:	Farah Nadir,	Pr,	Université de Annaba
Rapporteur:	Khadir Mohamed Tarek,	Pr,	Université de Annaba
Examinateurs:	Seridi Hamid,	Pr,	Université de Guelma
	Boukrouche Abdelhani,	Pr,	Université de Guelma
	Moussaoui Abdelouahab,	Pr,	Université de Sétif

Année Universitaire : 2013-2014

ملخص

ان الهدف من هذه الأطروحة هو اقتراح نهج تصنيفي غير خاضع للاشراف مبني على أساس خارطة التنظيم الذاتي kohonenوتقييم قابلية خارطة التنظيم الذاتي وبعض الأساليب الأخرى من الذكاء الحسابى في تحليل ونمذجة المشاكل المعلوماتية البيئية. ويستند النهج الأول على استخدام التنظيم الذاتي خريطة Kohonen (SOM) في المستوى الأول وطريقة التجميع (K-means) في المستوى الثاني. حيث يهدف هذا النهج الى تحديد أنواع الأرصاد الجوية اليومية لمنطقة عنابة اعتمادا على البيانات المتوفرة من طرف سماء صافية –عنابة- المجموعة خلال الفترة من عام 2003 إلى عام 2004. ويستند النهج الثاني للتجميع المقترح علىCHAOSOM من أجل تحليل وتحديد نوع التلوث الجوي اليومي لمنطقة عنابة. في المستوى الأول من النهج المقترح استخدم ثنائي الأبعادCHAOSOM لتدريب البيانات وفي المستوى الثاني استخدمنا CHAOSOM ذات بعد واحد. حيث أدخلنا معايير كمية باستخدام فئتين من مؤشرات النوعية للتحقق من صحة ومقارنة نتائج المجموعات. تجارب مختلفة باستخدام المناهج المقترحة ادت الى استخراج خمس فئات، والتي كانت مرتبطة بالأحوال الجوية النموذجية للمنطقة. كما تم استخدام المجموعات الناتجة لدراسة تأثير بارامترات الأرصاد الجوية (لكل مجموعة) على تلوث الهواء في هذا المنطقة، كما أجرينا عدة تحاليل خطية (المكونات الرئيسية للتحليل)، وغير خطية على أساس نموذج الشبكات العصبية الاصطناعية (متعددة الطبقات .(perceptron النتائج مرضية للغاية و عدة علاقات واستنتاجات تم استخلاصها.

كلمات مفتاحية: النهج التصنيفي الغير خاضع للاشراف، خارطة التنظيم الذاتي، تأثير بارامترات الأرصاد الجوية على تلوث الهواء.

Résumé

Cette thèse se place dans le cadre de l'apprentissage non supervisée dont l'objectif est de proposer une approche de classification basée sur les cartes auto-organisatrices de kohonen, communément désignées par SOM (pour Self Organising Maps), et d'évaluer l'utilisabilité des SOMs et d'autres méthodes de l'intelligence computationnelle dans l'analyse et la modélisation des problèmes de l'informatique environnementale. La première approche à deux niveaux de classification consiste à utiliser les cartes auto-organisatrices de Kohonen (SOM) pour le premier niveau et un algorithme de classification par partition (K-means) pour le deuxième niveau. Cette approche a été également utilisée pour l'identification des types de jours météorologiques pour la région d'Annaba à partir des données météorologiques captées par la station Samasafia durant la période 2003 à 2004. La deuxième approche de classification à deux niveaux consiste à utiliser une carte CHAOSOM à deux dimensions dans le premier niveau pour l'apprentissage des données et dans le deuxième niveau, une carte CHAOSOM unidimensionnel a été utilisée pour regrouper les vecteurs prototypes générés par le premier niveau de classification. Une analyse quantitative basée sur deux catégories de critères (internes et externes) et qualitative ont été utilisées pour valider et interpréter les résultats obtenus de la classification. La première approche a permis d'extraire cinq classes qui sont liés directement aux conditions météorologiques de la région. Les clusters météorologiques obtenus ont été utilisés pour étudier l'influence des paramètres météorologiques (par cluster) sur la pollution atmosphérique dans cette région. Une analyse non linéaire basée sur un modèle neuronal (perceptron multicouches) a été effectuées. Les résultats obtenus sont très satisfaisants et plusieurs relations et conclusions ont été tirées.

Mots-clés : classification non supervisée, cartes auto-organisatrices de kohonen, K-means, influence des paramètres météorologiques sur la pollution atmosphérique.

Abstract

The aim of this thesis is to propose an unsupervised classification approach based on Kohonen self-organizing maps and evaluating the usability of self-organizing maps and some other methods of computational intelligence in analyzing and modeling problems of environmental informatics. The first approach is based on using the self-organizing Kohonen map (SOM) in the first level and the partition clustering method (K-means) in the second level. This approach was also used to identify meteorological day types of Annaba region from the captured Samasafia data during the period 2003 to 2004. The second clustering approach is based on CHAO-SOM in order to analyze and identify the air pollution day type for the region of Annaba. In the first level of the proposed approach a two-dimensional CHAO-SOM is used to train data and the second level is to cluster data based on one-dimensional CHAOSOM. Quantitative (using two categories of validity indices) and qualitative criteria were introduced to verify the correctness and compare the clustering results. The different experiments developed extracted five classes, which were related to the typical meteorological conditions in the area. The obtained meteorological clusters were used to study the impact of the meteorological parameters (by cluster) on air pollution in this region. A nonlinear analysis based on neural network model (multi-layer perceptron) was carried out. The results obtained are very satisfactory where several relations and conclusions have been extracted.

Key-words: unsupervised classification, clustering, Self-Organizing Maps (SOM), K-means, impact of the meteorological parameters on air pollution.

Dédicaces

A mes parents

à mon épouse et mon fils : Mohamed Iyed

et à vous

Remerciements

Je tiens à remercier vivement Monsieur le Professeur Mohamed Tarek KHADIR mon directeur de thèse, pour sa confiance, son aide précieuse et surtout la disponibilité qu'il m'a accordée.

Je tiens à exprimer ma profonde reconnaissance au Professeur Nadir FARAH pour avoir bien voulu me faire l'honneur de présider mon jury de thèse et pour ses conseils précieux et son aide varié.

Je remercie également Messieurs les professeurs Hamid SERIDI, Abdelhani BOUKROUCHE et Abdelouahab MOUSSAOUI qui ont accepté d'examiner ce travail et de faire partie de mon jury de thèse.

Enfin, je remercie tout le personnel technique et administratif du département Informatique et mes collègues du laboratoire LABGED de l'université d'Annaba. Je remercie également tous ceux qui de près ou de loin ont contribuée à l'aboutissement de ce travail.

Pour leur patience, leur soutien permanent et leur sacrifice, je remercie fortement mes parents et mon épouse.

Que toutes ces personnes trouvent ici ma profonde gratitude pour m'avoir fait bénéficier de leur confiance scientifique et technique.

Liste des figures

Liste des tableaux

Table des Matières

Introduction générale

Introduction générale

Depuis l'apparition de l'informatique, l'ensemble de données stockées sous forme numérique ne cesse de croitre de plus en plus rapidement partout dans le monde. Les individus mettent de plus en plus les informations qu'ils possèdent à disposition de tout le monde via le web. De nombreux processus industriels sont également contrôlés par l'informatique. Et de nombreuses mesures effectuées un peu partout dans le monde, comme par exemple les mesures météorologiques qui remplissent des bases de données numériques importantes. Il existe des lors un grand intérêt à développer des techniques permettant d'utiliser aux mieux tous ces stocks d'informations tel que la classification automatique, afin d'en extraire un maximum de connaissance utile (Candillier, 2006). La résolution des problèmes de l'environnement est principalement une activité de traitement de l'information par le biais de la manipulation de bases de données volumineuses. L'informatique environnementale est une nouvelle discipline qui s'intéresse à la science de traitement de l'information (informatique) dans les sciences de l'environnement. Cette expression peut prendre un sens très large puisqu'elle englobe deux domaines séparés et différents, mais unis par un même projet d'études et d'actions (Huang et Chang, 2003 ; Kolehmainen, 2004). L'informatique environnementale vise à faciliter la recherche et la gestion environnementale en développant des techniques spécifiques pour accéder à l'information environnementale, pour intégrer les informations et connaissances provenant de différentes disciplines et pour créer des outils ou services basés sur ces informations.

Dans le cas des données météorologiques, leur analyse en utilisant les outils de la classification automatique peut aider à mieux comprendre les phénomènes généraux qui régissent le climat, afin, par exemple, d'anticiper les phénomènes extrêmes et d'agir en conséquence pour les populations concernées. L'identification des types de jours météorologiques, ou la classification des conditions atmosphériques dans des catégories (clusters), continue à être populaire, et de nombreuses méthodes ont été développées dans les deux dernières décennies. L'intérêt accru pour ce procédé est attribué à son utilité à résoudre une grande partie de problèmes climatologiques. Le souci de comprendre les impacts de la météorologie, particulièrement les implications possibles des changements climatiques a conduit la recherche pour plus et meilleures approches de classification météorologiques (Sheridan, 2002). En raison de la quantité énorme de données fournie par la station météorologique d'Annaba, des outils efficaces d'analyse de données sont indispensables afin d'extraire des connaissances utiles.

La disponibilité d'une vaste collection d'algorithmes de clustering dans la littérature peut facilement confondre un utilisateur tentant de choisir un algorithme approprié pour le problème considéré. En outre, il n'existe pas d'algorithme de clustering qui peut être universellement utilisé pour résoudre tous les problèmes. Par conséquent, il est important d'étudier soigneusement les caractéristiques du problème considéré, afin de sélectionner ou concevoir une stratégie de regroupement appropriée. Récemment, les cartes auto-organisatrices de kohonen (self-organizing maps (SOM)) ont été largement utilisées comme une méthode d'extraction et de visualisation de données complexes. Des milliers d'applications de la carte de kohonen dans différentes disciplines peuvent être trouvées dans l'étude de Oja et al, (2003). L'algorithme SOM représente une classe d'apprentissage non supervisée des réseaux de neurones dont la caractéristique principale est son capacité de mapper les relations non linéaire d'un ensemble de données multidimensionnelles dans une grille de neurones à deux dimensions facilement visualisable. Les cartes de kohonen sont également appelées cartes auto organisatrices topologiques puisque la fonction de base d'un SOM est d'afficher la topologie ou les relations entre les membres d'un ensemble de données. L'algorithme SOM a été développé par Kohonen dans les années 1980, et depuis lors, il a été utilisé comme un outil de reconnaissance de formes et de classification dans les différents domaines tels que la robotique, l'astronomie, et la chimie (Gupta et al, 2008). L'algorithme SOM a suscité beaucoup d'intérêt dans plusieurs domaines. Il a été largement analysé, un certain nombre de variantes ont été développés et, peut-être surtout, il a été largement appliqué dans des domaines aussi variés que les sciences de l'ingénieur, la médecine, la biologie et l'environnement. L'algorithme SOM est un modèle simulant le processus d'auto-organisation du cerveau humain. Toutefois, il est encore très loin pour la réalisation du mécanisme du cerveau. Dans le but de réaliser des méthodes de classification non supervisée plus puissantes, il est intéressant de proposer de nouveaux modèles inspirés du mécanisme du cerveau humain et d'étudier leurs comportements. En 1952, Hodgkin et Huxley ont montré expérimentalement, utilisant un axone géant du calmar, qu'une membrane nerveuse réel dans l'état de repos répond à une stimulation d'impulsions périodiques non seulement synchrone mais aussi chaotique en fonction des valeurs d'amplitude et de période des impulsions de stimulation (Aihara et al, 1990). Depuis lors, plusieurs chercheurs ont essayé d'exploiter les fonctionnalités chaotiques pour résoudre les problèmes d'optimisation combinatoire. Aihara et al, (1990) ont proposé le premier modèle général du neurone chaotique. Ce modèle a été utilisé pour résoudre le problème du voyageur de commerce avec une efficacité remarquable

dont la qualité des solutions est meilleure que celle du réseau de Hopfield classique (Aihara et al, 1990 ; Yamada et al, 1993). Par conséquent, il est important d'étudier la possibilité d'ajouter des fonctionnalités chaotiques à la carte de kohonen. Ce concept est appelé SOM Chaotique (CHAOSOM) (Matsushita, 2006), où le taux d'apprentissage et de coefficient voisinage du SOM sont rafraîchies par des impulsions chaotiques générés par l'équation Hodgkin-Huxley (Hodgkin et Huxley, 1952).

Notre objectif dans cette thèse consiste à proposer une approche de classification non supervisée basée sur les cartes auto-organisatrices de kohonen est d'évaluer l'utilisabilité des SOMs et d'autres méthodes d'intelligence computationnelle dans l'analyse et la modélisation des problèmes de l'informatique environnementale. La première approche, composée de deux niveaux, consiste à utiliser une carte auto-organisatrice de Kohonen (SOM) pour le premier niveau et un algorithme de classification par partition (K-means) dans le deuxième niveau. Cette approche a été également utilisée pour l'identification des types de jours météorologiques pour la région d'Annaba à partir des données météorologiques captées par la station Samasafia d'Annaba durant la période 2003 à 2004. La deuxième approche de classification à deux niveaux consiste à utiliser une carte CHAOSOM à deux dimensions dans le premier niveau pour l'apprentissage des données et dans le deuxième niveau une carte CHAOSOM unidimensionnel est utilisée pour regrouper les vecteurs prototypes générés par le premier niveau.

Dans le premier chapitre, nous présentons le contexte géographique et météorologique de la région d'Annaba et notamment les paramètres météorologiques utilisés. Ensuite, nous introduisons les concepts importants et nécessaires à la compréhension du reste de la thèse tel que l'intelligence computationnelle et l'informatique environnementale ainsi que les défis posés par les données recueillies dans les processus de ce domaine. Dans le deuxième chapitre, nous abordons un état de l'art sur la classification des données, nous commençons ce chapitre par rappeler quelques concepts et définitions essentiels pour comprendre les différentes méthodes et outils de classification automatique. Nous présentons par la suite les étapes de base de la procédure de classification automatique ainsi que les différentes applications possibles du clustering. Puis, nous discutons les différentes notions qui sont utilisées pour définir la similarité entre objets, qui constitue la base de toute méthode de clustering. Enfin, nous présentons les approches existantes permettant d'évaluer les résultats d'algorithmes de clustering. De plus amples détails sur les principales méthodes de clustering sont ensuite présentés dans le chapitre suivant. Ainsi, Dans le troisième chapitre, nous avons

choisi de présenter un tour d'horizon des méthodes de clustering les plus classiques, en mettant l'accent, plus précisément, sur les cartes au-organisatrices de kohonen. Dans le quatrième chapitre, nous présentons les deux approches de classification automatique proposées. La première approche consiste à utiliser la carte auto-organisatrice de Kohonen (SOM) pour le premier niveau et la classification par partition (K-means) dans le deuxième niveau. La deuxième approche de regroupement à deux niveaux est obtenue par la création d'un ensemble de prototypes utilisant un CHAOSOM bidimensionnel. Ces prototypes sont ensuite regroupés dans le second niveau en utilisant un CHAOSOM unidimensionnelle. Ensuite, Les clusters météorologiques obtenus ont été utilisés pour étudier l'influence des paramètres météorologiques sur la pollution atmosphérique pour la région d'Annaba.

Chapitre 1

Domaine d’application et profils météorologiques et géographiques de la région d'Annaba

1.1. Introduction

La pollution que génèrent l'urbanisation rapide, la croissance de la population et de l'industrialisation a pris des dimensions alarmantes, et constitue le plus grand fléau que l'humanité ait à affronter dans les prochaines années. En Algérie, il est évident de constater que les risques dus à la pollution sont beaucoup plus importants dans le Nord du pays, et davantage sur le littoral en raison de la concentration d'activités industrielles qui a des effets négatifs sur les équilibres écologiques et économiques de nos espaces littoraux (Mebirouk et Mebirouk-Bendir, 2007). Pour distinguer les problèmes de pollution en milieu urbain, nous nous sommes intéressés à la ville d'Annaba et sa périphérie où les effets de la pollution, notamment atmosphérique, sont gravissimes pour sa population. Pour évaluer la qualité de l'air dans la région d'Annaba, un dispositif de surveillance est mis en place, depuis juin 2002. Ce dispositif dénommé «Samasafia», permet de mesurer la concentration des principaux polluants atmosphériques dans la région d'Annaba.

Il est clair que les différents problèmes de protection de l'environnement et de la planification environnementale ne peuvent être résolus qu'avec l'implication d'une base complète et fiable d'informations relatives. La résolution des problèmes de l'environnement est principalement considérée comme une activité de traitement de données afin d'en extraire un maximum de connaissance utiles. Les solutions de nos problèmes environnementaux sont fortement liées à la qualité des ressources d'information accessibles. Ces informations sont aussi importantes pour les décisions sur les actions à entreprendre pour assurer une meilleure protection de l'environnement et pour l'acquisition de connaissances dans la recherche environnementale (Huang et Chang, 2003; Kolehmainen, 2004).

1.2 Domaine d'application

1.2.1 Intelligence computationnelle

Bien que le terme "intelligence computationnelle (IC)" a été largement utilisé depuis longtemps avec régulièrement une quantité croissante de recherches et d'applications, il n y a pas de définition globalement commune. IC peut signifier beaucoup de choses pour différentes personnes et différentes techniques sont impliquées comme appartenant à l'IC. Selon (Engelbrecht, 2007) l'intelligence computationnelle est une sous-branche de l'intelligence artificielle (IA), qui s'occupe de l'étude des mécanismes d'adaptation qui

permettent ou facilitent un comportement intelligent dans un environnement complexe et dynamique. Ces mécanismes comprennent les paradigmes de l'IA qui possède une certaine capacité d'apprentissage ou d'adaptation à de nouvelles situations, de généralisation, d'abstraction, de découverte et d'association.

Pal et Pal (2002) ont combinés plusieurs définitions en exigeant les caractéristiques que doivent remplir les composants de l'IC (réseau de neurones, logique floue, etc.):

- Une potentialité considérable dans la résolution des problèmes du monde réel.
- Capacité d'apprendre par l'expérience.
- Capacité d'auto-organisation
- Capacité d'adaptation en réponse à l'évolution dynamique des conditions et des contraintes.

La difficulté de définir l'IC d'une manière courte et complète est probablement la raison pour laquelle elle est souvent exprimée en termes de ses composantes. La liste la plus souvent rencontrée inclut les Réseaux de Neurones Artificiels (RNA), la logique floue, les Algorithmes Génétiques (AG) et les réseaux bayésiens. Il convient de noter que le rôle de ces méthodes n'est pas compétitif, mais synergique et complémentaire. Cette idée est aussi un enjeu central de ce travail, à savoir celle de la combinaison de plusieurs méthodes de l'IC.

1.2.2. La qualité de l'air urbain

Avec l'expansion industrielle et technologique, la pollution de l'air présente aujourd'hui l'un des problèmes sérieux pour la plus part des populations sur terre non seulement celles qui sont industrialisées mais aussi pour celles qui sont en voie de développement. Malgré les progrès constatés ces dernières années à l'échelle urbaine (ex : élimination du plomb des carburants), les études épidémiologiques mettent toujours en évidence des répercussions de la pollution atmosphérique sur l'appareil respiratoire et cardio-vasculaire, c'est pourquoi une panoplie de mesures est nécessaire à mettre en œuvre pour réduire la pollution atmosphérique urbaine. Pour cette raison, la planification scientifique des méthodes d'analyse et de contrôle de la pollution sont devenues obligatoires. Dans ce cadre, il est nécessaire (i) d'analyser et de spécifier toutes les sources de pollution et leurs effets sur la qualité de l'air, (ii) d'étudier les différents facteurs qui provoquent la pollution, et (iii) de développer des outils capables de réduire la pollution par l'introduction de mesures de contrôle et des alternatives aux pratiques existantes (Kolehmainen, 2004).

Les recherches ainsi effectuées visent à atteindre une meilleure compréhension des phénomènes liés aux polluants atmosphériques, aux particules solides et aux paramètres météorologiques. L'objectif spécifique est donc de développer de nouveaux modèles qui peuvent prévoir la qualité de l'air urbain dans le futur. Par exemple, la modélisation d'un processus de prévision de concentrations des polluants pour un jour en avant, sera de grande utilité pour la population concernée. Une telle modélisation s'appuie généralement sur l'historique de données des différents polluants et des paramètres météorologiques. La modélisation donc se présente comme un outil reconnu et très répandu pour étudier les questions liées à la qualité de l'air ambiant. Par la multiplicité des scénarii qu'elle propose, la modélisation est un moyen essentiel d'élaboration et d'évaluation des plans d'amélioration de la qualité de l'air.

Une des tâches principales à réaliser dans le domaine des applications météorologiques est l'utilisation des systèmes d'extraction de connaissances à partir de données pour l'étude et l'analyse des paramètres atmosphériques en utilisant la classification automatique. Cette dernière permet l'extraction d'un ensemble de prototypes (clusters) représentant les modèles météorologiques par rapport à une région d'intérêt. Ceci revient donc à identifier les différents types de jours météorologiques qui caractérisent une région d'intérêt. L'intérêt accru pour ce procédé revient à attribuer à son utilité la capacité de résoudre une grande partie de problèmes climatologiques (Sheridan, 2002). La prédiction de la qualité de l'air est un domaine de recherche typique dans le contexte de l'informatique environnementale à de nombreux égards. Premièrement, plusieurs bases de données environnementales ont été mesurées et recueillies tant pour les paramètres de la qualité de l'air (concentrations des polluants atmosphériques et des particules ultrafines) que pour les paramètres météorologiques qui sont étroitement liés à ces polluants. Cela permet l'utilisation des méthodes s'appuyant sur les données telles que les réseaux de neurones (Kolehmainen et al. 2004). Deuxièmement, les données proviennent de processus ouverts ce qui rend les caractéristiques chaotiques de nature. Troisièmement, la qualité de l'air urbain est également reliée aux principaux axes des sciences de l'environnement, à savoir les conséquences des activités humaines sur la nature et les êtres humains eux-mêmes sous la forme d'une détérioration des conditions de vie et de santé (Kolehmainen et al. 2004).

1.2.3. L'informatique environnementale: un médiateur entre les sciences de l'environnement et l'informatique

Les systèmes informatiques pour le traitement de données environnementales ont été en usage depuis plus de trois décennies. Une large gamme d'applications a été couverte par ces systèmes, y compris la surveillance et le contrôle, la gestion de l'information, l'analyse de données, ainsi que la planification et l'aide à la décision. Les progrès de l'informatique ont apporté une précieuse contribution à la capacité d'analyse des processus biologiques, chimiques et physiques qui se produisent dans l'environnement. Inversement, la nature complexe des problèmes qui se posent dans les contextes environnementaux est un grand défi pour l'informatique. De ce processus de stimulation mutuelle, une discipline particulière a émergé connu comme l'informatique environnementale. Il combine les topiques de l'informatique tels que les systèmes de bases de données, systèmes d'information géographique et la modélisation en ce qui concerne leurs applications à la recherche et à la protection de l'environnement (Hilty et al, 2006).

L'informatique environnementale est une nouvelle discipline qui s'intéresse à la science de traitement de l'information (informatique) appliquée sur les sciences de l'environnement. L'expression peut véhiculer un sens très large puisqu'elle englobe deux domaines qui sont considérés à la fois comme séparées et différents, mais unis par un même projet d'études et d'actions (Huang et Chang, 2003 ; Kolehmainen, 2004). L'informatique environnementale vise donc à faciliter la recherche et la gestion environnementales tout en développant des techniques spécifiques permettant d'accéder à l'information environnementale, intégrer les informations et connaissances provenant de différentes disciplines et créer des outils ou des services basés sur ces informations. Cette discipline a été initialement définie par (Page et Hilty, 1995) comme une sous-discipline particulière de l'informatique appliquée traitant les méthodes et les outils de l'informatique qui permettrait d'analyser et mettre en place les procédures de traitement de l'information et qui contribuent à la recherche, l'élimination, la prévention, et la réduction des charges et aux dommages environnementales.

« *environmental informatics is a special sub-discipline of applied informatics dealing with the methods and tools of computer sciences for analysing, supporting and setting up those information processing procedures which are contributing to the investigation, removal, avoidance and minimisation of environmental burden and damages.* »

Une autre définition de l'informatique environnementale a été aussi proposée par (Green et Klomb, 1998) mettant davantage l'accent sur l'intégration des sources de données globales sur les aspects des sciences de la vie.

« *The application of information technology to environmental issues is changing both theory and practice. The idea of natural computation provides new ways to understand environmental complexity across the entire range of scales, from individual phenotype to biogeography. Understanding the ways in which local interactions affect the global composition and dynamics of whole communities is crucial to the viability of strategies to manage ecosystems, especially in landscapes altered by human activity. Also environmental planning and management are increasingly dependent on accurate, up to date information that sets local decisions within a global context. The internet makes it possible to combine environmental data from many different sources, raising the prospect of creating a global information warehouse that is distributed amongst many contributing sites.* » (Green et Klomb, 1998; Kolehmainen, 2004).

D'autre part, l'informatique environnementale peut être considérée comme un médiateur entre les sciences de l'environnement et de l'informatique moderne en offrant des solutions innovantes étant principalement fondées sur des données recueillies qui sont par la suite traitées par un mécanisme qui couple d'une part les informations acquises et les connaissances nécessaires dans le processus de résolution des problèmes d'autre part. Ces concepts sont illustrés dans la figure 1.1.

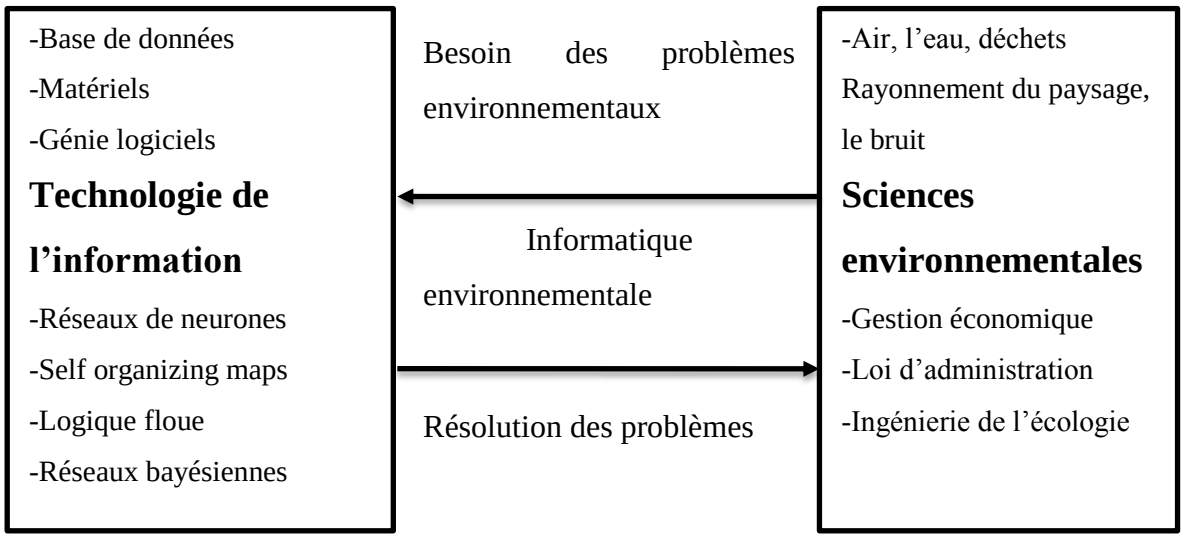


Figure 1. 1– Le rôle de l'informatique environnementale en tant que médiateur entre les sciences de l'environnement et de l'informatique moderne (Kolehmainen, 2004).

1.3 Défis dûs aux données environnementales

L'ensemble de données environnementales est généralement constitué de séries temporelles mesurées sur plusieurs années. Les principaux facteurs, qui influencent la modélisation, proviennent donc des cycles de la nature (la saisonnalité et les variations quotidiennes) et de l'activité humaine, par exemple, les variations journalières et hebdomadaires (Kolehmainen, 2004). La saisonnalité conduit généralement à la nécessité d'avoir au moins 3-5 ans de données mesurées pour la modélisation. De cette manière, une période de un à deux ans est nécessaire pour la procédure de validation des résultats de modélisation. De nombreux défis existent pour les applications des techniques de modélisation de la gestion environnementale. De façon générale, les quantités de données disponibles sont suffisantes mais la qualité des données est celle qui présente la source de la préoccupation. Le problème le plus souvent rencontré dans tel domaine est dû à l'absence de données (ou données manquantes), en raison des pannes des appareils de mesures ou des erreurs humaines. Dans le cas des séries temporelles, la plupart des méthodes de l'intelligence artificielle exigent un ensemble de données assez complet, ce qui conduit à l'imputation des valeurs manquantes (Kolehmainen, 2004). Un autre problème typique réservé à la qualité des données est dû à des erreurs de mesure commises soit par des dispositifs ou des êtres humains, ce problème est généralement appelé « problème des données aberrantes ». Par conséquent, Un défi majeur peut être soulevé depuis cette situation qui s'agit des erreurs systématiques provenant de mauvaise calibration des appareils de mesure lors de l'acquisition de données.

Après une étude de la base de données réellement captée par la station Samasafia, on a constaté qu'il y a des valeurs aberrantes et des valeurs manquantes dans cette base. La détection des valeurs manquantes est assez facile, et nous pouvons en déduire facilement la cause. Cependant, la détection des valeurs aberrantes n'est pas une tâche assez facile que la détection des valeurs manquantes. Les données réellement observées à un moment donné sont spatialement distribués, les données collectées sur la température par exemple dépendent à la fois du facteur du temps et de l'espace, peuvent parfois inclure d'autres facteurs climatiques et non climatiques (Takahashi et al, 2011). Il est à noter qu'il est difficile de détecter toutes les données bruitées de façon instantanée par une détection automatique.

Il est supposé qu'à l'issue d'une modélisation ou d'une conception basée sur les données disponibles, l'ensemble du travail est simplement à mi-chemin réalisé. Dans le reste du chemin, on examine les informations qui ne sont pas disponibles et considérées comme cachées. Ces données pourraient se présenter comme des connaissances implicites pour les

décideurs qui pourraient les acquérir grâce aux technologies de l'information innovantes telles que l'intelligence artificielle et l'exploration de données.

1.4 La région d'étude et les données utilisées

Annaba, ville côtière du nord-est de l'Algérie, à 600 km d'Alger. Cette région est constituée d'une vaste plaine bordée au sud et à l'ouest, d'un massif montagneux au Nord, et par la mer méditerranéenne à l'est sur une longueur de 50 km représentant le cordon dunaire. Elle couvre une superficie d'environ 520 km^2. La région d'étude est soumise à un climat méditerranéen caractérisé principalement par deux saisons distinctes. L'une humide, marquée par une forte pluviosité et par de faibles températures allant du mois d'octobre à mai. L'autre sèche et chaude avec de fortes températures atteignant le maximum au mois d'août (Mebirouk et Mebirouk-Bendir, 2007; Derradji et al, 2005).

1.4.1 Les mesures (Stations de surveillance "Samasafia")

Un dispositif de surveillance de la qualité de l'air qui permet l'évaluation de la pollution atmosphérique dans la région d'Annaba est mis en place, depuis juin 2002. Ce dispositif de mesures de la qualité de l'air, dénommé «Samasafia», permet de mesurer la concentration des principaux polluants atmosphérique dans la région de Annaba, notamment les rejets générés par le complexe sidérurgique MITTAL STEEL d'EL HADJAR (ex SNS et ISPAT) et le complexe de production d'engrais FERTIAL (ex ASMIDAL). Ce réseau est constitué de quatre stations implantées à Annaba ville, El Bouni, Sidi Amar et au niveau de l'aéroport Rabah Bitat (voir figure. 1.2) (Chaffai et Mourdi, 2011; Alioua et al, 2008).

Ces stations qui sont principalement installées à proximité des pôles industriels, sont dotées d'ensembles d'appareils de mesure et de transfert des données vers un poste central. Un système d'acquisition des données permet d'élaborer un bulletin précis relatif à la qualité de l'air à travers l'indice de pollution de chaque région. Ce système permet donc de détecter les pics de pollution et d'alerter les autorités concernées pour informer et sensibiliser la population durant les périodes critiques (Chaffai et Mourdi, 2011).

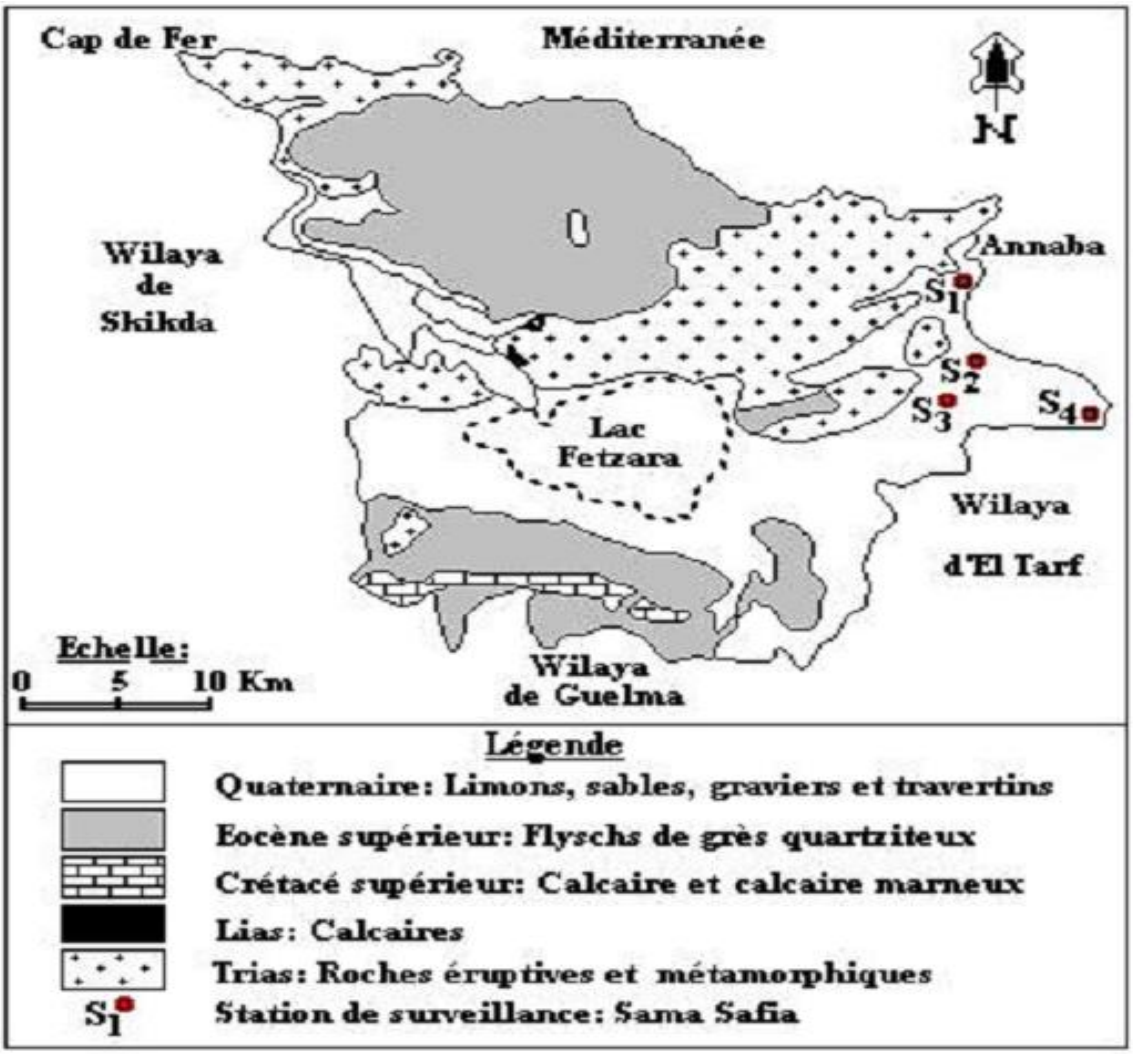


Figure 1. 2– Situation géographique de la wilaya d'Annaba (Chaffai et Mourdi, 2011).

Les missions de Samasafia se définissent dans 5 directions (Samasafia, 2006):

- Surveiller la qualité de l'air sur l'ensemble de la région;
- Analyser et expliquer les phénomènes de pollution atmosphérique;
- Informer la population et les décideurs;
- Alerter en cas de pic de pollution atmosphérique;
- Communiquer sur la qualité de l'air;

Le rôle réglementaire de Samasafia consiste à (Samasafia, 2006):

- fournir des indications sur l'ensemble des polluants réglementés (dioxyde de soufre, oxydes d'azote, monoxyde de carbone, ozone, poussières, etc.) sur les agglomérations où les seuils guides risquent d'être dépassées.
- rendre les données collectées accessibles à toute personne (mise en ligne sur internet, indice de qualité de l'air quotidien et alerte en cas de pic de pollution).

1.4.2 Base de données

Les données utilisées dans le cadre de cette étude ont été collectées par la station Samasafia de Sidi Amar. La base de données collectée par la station Samasafia d'Annaba sur une base de mesure continue de 24 heures pendant la période 2003-2004. Les polluants atmosphériques surveillés en continue incluent les concentrations du : monoxyde d'azote (NO), monoxyde de carbone (CO), l'ozone (O_3), particule en suspension (PM_{10}), oxydes d'azote (NO_x), dioxyde d'azote (NO_2), dioxyde de soufre (SO_2). Cette base de données contient également trois paramètres météorologiques : la vitesse du vent, la température et l'humidité relative.

Tableau 1. 1– Données stockées sous forme tabulaire contenant des valeurs manquantes et des valeurs aberrantes.

	CO mg/m^3	HU %	NO µg/m^3	NO_X µg /m^3	NO_2 µg/m^3	O_3 µg/m^3	PS µg/m^3	TE deg	VV m/s	SO_2 µg /m^3
01/01/2003 01:00	0,5	71	1,0	2,0	2,0	NAN	74,0	12,7	2,4	17,0
01/01/2003 02:00	0,2	72	1,0	2,0	2,0	NAN	49,0	12,4	2,3	18,0
01/01/2003 03:00	0,0	73	1,0	2,0	2,0	NAN	38,0	12,2	2,6	18,0
01/01/2003 04:00	0,0	70	1,0	2,0	2,0	NAN	38,0	12,6	2,8	18,0
01/01/2003 05:00	0,1	72	1,0	4,0	4,0	NAN	42,0	12,1	1,6	18,0
01/01/2003 06:00	0,2	71	1,0	5,0	6,0	NAN	40,0	12,1	2,6	19,0
01/01/2003 07:00	0,2	70	1,0	5,0	6,0	NAN	50,0	12,1	2,5	19,0
01/01/2003 08:00	1,0	72	1,0	10,0	13,0	NAN	61,0	12,0	2,8	20,0
01/01/2003 09:00	1,3	71	1,0	12,0	16,0	NAN	63,0	12,4	3,1	18,0
01/01/2003 10:00	1,3	69	2,0	21,0	28,0	NAN	76,0	13,3	2,9	14,0
01/01/2003 11:00	3,5	67	4,0	16,0	17,0	NAN	83,0	14,3	3,3	26,0
01/01/2003 12:00	1,1	70	0,0	6,0	8,0	NAN	104,0	14,2	3,2	18,0
01/01/2003 13:00	0,3	71	1,0	2,0	1,0	NAN	76,0	14,4	2,3	17,0
01/01/2003 14:00	0,0	58	1,0	1,0	0,0	NAN	43,0	16,6	2,8	19,0
01/01/2003 15:00	0,0	56	1,0	1,0	0,0	NAN	25,0	16,6	**-3,2**	14,0
01/01/2003 16:00	0,0	56	1,0	1,0	0,0	NAN	19,0	16,8	3,1	15,0
01/01/2003 17:00	0,0	62	1,0	1,0	0,0	NAN	24,0	16,0	2,6	11,0
01/01/2003 18:00	0,0	67	1,0	2,0	2,0	NAN	32,0	15,3	2,5	16,0
01/01/2003 19:00	0,1	69	0,0	2,0	2,0	NAN	38,0	15,0	0,9	15,0
01/01/2003 20:00	0,2	74	0,0	5,0	8,0	NAN	31,0	14,6	1,6	14,0
01/01/2003 21:00	0,2	72	0,0	5,0	7,0	NAN	32,0	14,3	2,1	18,0
01/01/2003 22:00	0,1	67	1,0	1,0	0,0	NAN	31,0	14,0	2,2	14,0
01/01/2003 23:00	0,2	72	1,0	5,0	6,0	NAN	43,0	12,9	2,3	14,0
02/01/2003 00:00	2,2	76	1,0	7,0	10,0	NAN	64,0	**121**	1,2	15,0

Les données sont représentées sous forme tabulaire telles que les lignes correspondent aux exemples, les colonnes correspondent aux attributs qui les caractérisent et les cases correspondent aux valeurs des exemples sur ces attributs. Cette représentation de données est appelée attribut-valeur ou représentation sous forme matricielle. Le tableau 1.1 représente un échantillon de données collectées par la station Samasafia pour un jour. Chaque entrée du

tableau est définie par un ensemble de caractéristiques: trois paramètres météorologiques et sept polluants atmosphériques. D'autres terminologies peuvent être utilisées pour décrire les éléments de telle base de données tabulaire. Généralement, les lignes du tableau sont appelées les exemples, et les colonnes sont appelées les attributs. Ainsi, les lignes du tableau peuvent être également considérées comme des objets décrits par des valeurs sur plusieurs dimensions, ou bien des points décrits par leurs coordonnées. Comme nous avons mentionné dans la section 1.3, les données peuvent être bruitées, c'est à dire que certaines de leurs valeurs sont aberrantes. Il peut aussi arriver que certaines valeurs ne soient pas renseignées, on parle dans ce cas des données manquantes. Le tableau 1.1 présente un exemple de valeurs aberrantes (écrites en gras) et de valeurs manquantes (mentionnées par NAN).

1.4.3 Pollution atmosphérique dans la région d'Annaba

La dégradation de la qualité de l'air liée aux activités de l'homme sont de plus en plus au cœur des préoccupations des spécialistes de la santé publique et des agences chargées de la protection de l'environnement. Cette perception se justifie par le fait que les polluants atmosphériques provoquent des effets nocifs sur la santé ou des gênes respiratoires. En Algérie, les infections respiratoires demeurent la première cause de mortalité infantile après la rougeole et la diarrhée (Noui et Boukhemis, 2011). La bronchite chronique, le cancer du poumon et l'asthme sont, entre autres, les maladies engendrées par la pollution. Au fait, un tiers de la population algérienne soufre d'une morbidité respiratoire. Bien qu'une étude épidémiologique permettant de faire une corrélation entre les niveaux de pollution atteints et les maladies respiratoires n'ait jamais été réalisée au niveau de la région d'Annaba, les concentrations élevées de certains polluants ont déjà atteint des niveaux dangereux, particulièrement pour les personnes fragiles du cœur et des poumons. Ceci est d'autant plus apparent dans certaines régions non aérées et à forte industrie. La ville d'Annaba est considérée comme étant l'une des villes les plus polluées sur le territoire national et dans le nord de l'Afrique (Alioua et al, 2008). Dans cette région, le taux de prévalence de l'asthme est supérieur au taux national, 55 % des asthmatiques ont plus d'une crise par mois et 42 % des patients ont été hospitalisés au moins une fois durant l'année 2008 (selon l'archive de Samasafia[1]).

[1] www.samasafia.dz/journaux

1.4.4 Les causes de la pollution à Annaba

Les émissions atmosphériques des principaux contaminants se répartissent de façon différente en fonction du secteur d'activité. Les principales sources émettrices de polluants atmosphériques sont (Alioua et al, 2008; Mebirouk et Mebirouk-Bendir, 2007) :

• Le complexe sidérurgique MITTAL STEEL d'EL HADJAR (ex SNS et ISPAT);

• Le complexe d'engrais FERTIAL (ex ASMIDAL);

• Le trafic routier qui évolue à un rythme inquiétant.

La principale source de pollution à Annaba réside dans sa zone industrielle et particulièrement dans ses deux complexes géants de production d'engrais phosphatés et azotés «FERTIAL» (ils se trouvent juste à proximité de la ville) et le complexe de produits sidérurgiques d'El-Hadjar, malgré l'installation d'un système de dépollution au niveau du complexe FERTIAL qui a réduit la pollution par NO_x à El-Bouni par rapport aux années précédentes. Ces activités industrielles constituent la principale source d'émission de matières particulaires et d'oxyde de soufre, tandis que les émissions de monoxyde de carbone, d'azote et de plomb sont surtout dues au secteur du transport. La pollution atmosphérique a progressé avec l'accroissement du nombre de véhicules (une hausse annuelle de 5 % en moyenne en Algérie) ainsi qu'avec l'absence totale de contrôle des émissions (Mebirouk et Mebirouk-Bendir, 2007).

Figure 1. 3--Topographie de la région d'Annaba.

Le problème de la pollution atmosphérique à Annaba a été aggravé par sa position géographique. En effet, la région d'Annaba comme le montre la figure 1.3 est entourée par le massif de l'Edough culminant à 1008 m d'altitude. Cette topographie en forme de cuvette favorise la stagnation de l'air pollué et l'accumulation des taux de concentration des polluants. Les brises de mer, terre, et pente concourent au transport des nuages de polluants. En effet, durant la nuit, les brises de terre entraînent ces nuages de polluants vers la mer. Ces derniers retournent sur la ville par effet de brise de mer en longeant la montagne de SERAIDI. Les polluants se déposent lentement sur la région sous une forme de cercle et l'on assiste à une pollution de plus en plus grave. En plus de ces facteurs de pollution, il y a aussi les déchets solides (ménagers, industriels, hospitaliers, toxiques) qui sont également considérés comme une source de pollution atmosphérique dans la mesure où ces déchets sont incinérés à l'air libre. L'incinération des déchets solides provoque le dégagement des nuages de pollution comprenant 50 à 65 % de gaz du méthane (important gaz à effet de serre, comme il peut former un mélange explosif avec l'air). Les principaux polluants urbains dans la région d'Annaba sont présentés dans le tableau 1.2 (Mebirouk et Mebirouk-Bendir, 2007).

Tableau 1. 2--Les principaux polluants urbains à Annaba (Mebirouk et Mebirouk-Bendir, 2007).

Polluant	**Origine**	**Effets Environnemen--taux**	**Effets Biologiques**	**Observations**
Monoxyde de carbone (CO)	Combustion incomplète des carburants.		Bloque l'oxygénation de tissus. A forte dose : asphyxie mortelle.	Effet de Proximité
Oxyde d'azote (No_x)	Trafic automobile	Formation d'ozone en basse Atmosphère (NO_x + Vapeur) = Contribution	Altération des fonctions respiratoires	Le monoxyde émis à l'échappement s'oxyde et se transforme en dioxyde d'azote

		aux pluies acides)		(NO_2) plus toxique.
Dioxyde de souffre SO_2	Combustion du Fioul	SO_2+ Vapeur = Acide sulfurique (Pluies Acides)	Gaz irritant : asthme et gêne respiratoire	Effet régional
Particules en suspensions	Émission des moteurs Diesel	Souillures des bâtiments	Se fixant dans les voies respiratoires	Effet de Proximité
Plomb	Nuisances en ville		Oxyde de plomb est un toxique neurologique, rénal, etc.	Effet de proximité
Dioxyde de carbone (CO_2)	Combustion des carburants	Effet de serre		Effet planétaire
Ozone (O_3)	Composant de l'air. Réaction photochimique entre oxygène de l'air, oxydes d'azote et de souffre, COV sous l'effet du rayonnement ultra violet du soleil.	Concentration en basse atmosphère	Irritation oculaire, céphalées. Altère les fonctions respiratoires et la résistance aux infections.	Protège la planète en haute altitude

1.5 Évaluation de Données

La base de données utilisée dans la présente étude couvre la période 2003-2004 et a été fournie par le réseau Samasafia d'Annaba sur une base de mesure continue de 24 heures. Une vue générale des polluants étudiés est montrée par le tableau 1.3. Les niveaux de concentration moyenne pour les polluants considérés pour cette étude sont montrés par la figure 1.4 et la figure 1.5. Ainsi les niveaux de concentration de ces polluants pendant la période 2003-2004 sont représentés par la figure 1.6. La figure 1.4, et la figure 1.5 montrent également des différences principales dans les niveaux de pollution entre les quatre saisons de chaque année. Les concentrations des polluants dans la saison du printemps tendent à être occasionnelles, des épisodes de pollution étaient très élevés dans la saison d'hiver pour la plupart des polluants.

Tableau 1. 3–Résumé de la base de données utilisée et plage de variation des polluants.

Polluant	*Mesure*	*1er trimestre*	*2eme trimestre*	*3eme trimestre*	*4eme trimestre*
2003					
CO	*Mg/m3*	*Oui*	*Oui*	*Oui*	*D-manquantes*
NO_2	*Mg/m3*	*Oui*	*Oui*	*Oui*	*Oui*
NO_x	*Ug/m3*	*Oui*	*Oui*	*Oui*	*Oui*
NO	*Ug/m3*	*Oui*	*Oui*	*Oui*	*Oui*
O_3	*Microg/m3*	*D-manquantes*	*D-manquantes*	*D-manquantes*	*D-manquantes*
PM_{10}	*Microg/m3*	*Oui*	*Oui*	*D-manquantes*	*Oui*
SO_2	*Microg/m3*	*Oui*	*Oui*	*Oui*	*Oui*
2004					
CO	*Mg/m3*	*D-manquantes*	*D-manquantes*	*D-manquantes*	*D-manquantes*
NO	*Mg/m3*	*D-manquantes*	*Oui*	*Oui*	*Oui*
NO_x	*Ug/m3*	*D-manquantes*	*Oui*	*Oui*	*Oui*
NO_2	*Ug/m3*	*D-manquantes*	*Oui*	*Oui*	*Oui*
O_3	*Microg/m3*	*Oui*	*Oui*	*Oui*	*Oui*
PM_{10}	*Microg/m3*	*Oui*	*Oui*	*Oui*	*Oui*
SO_2	*Microg/m3*	*D-manquantes*	*D-manquantes*	*D-manquantes*	*D-manquantes*

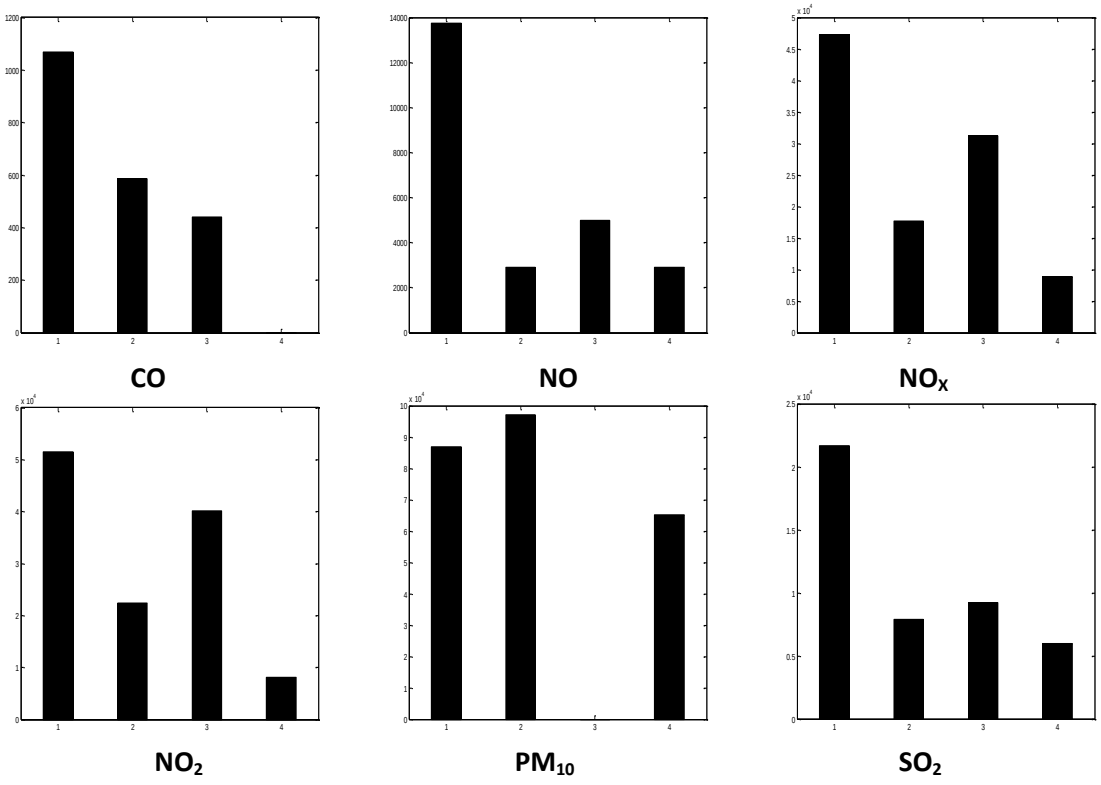


Figure 1. 4–Niveaux de concentration moyenne des polluants pendant 2003.

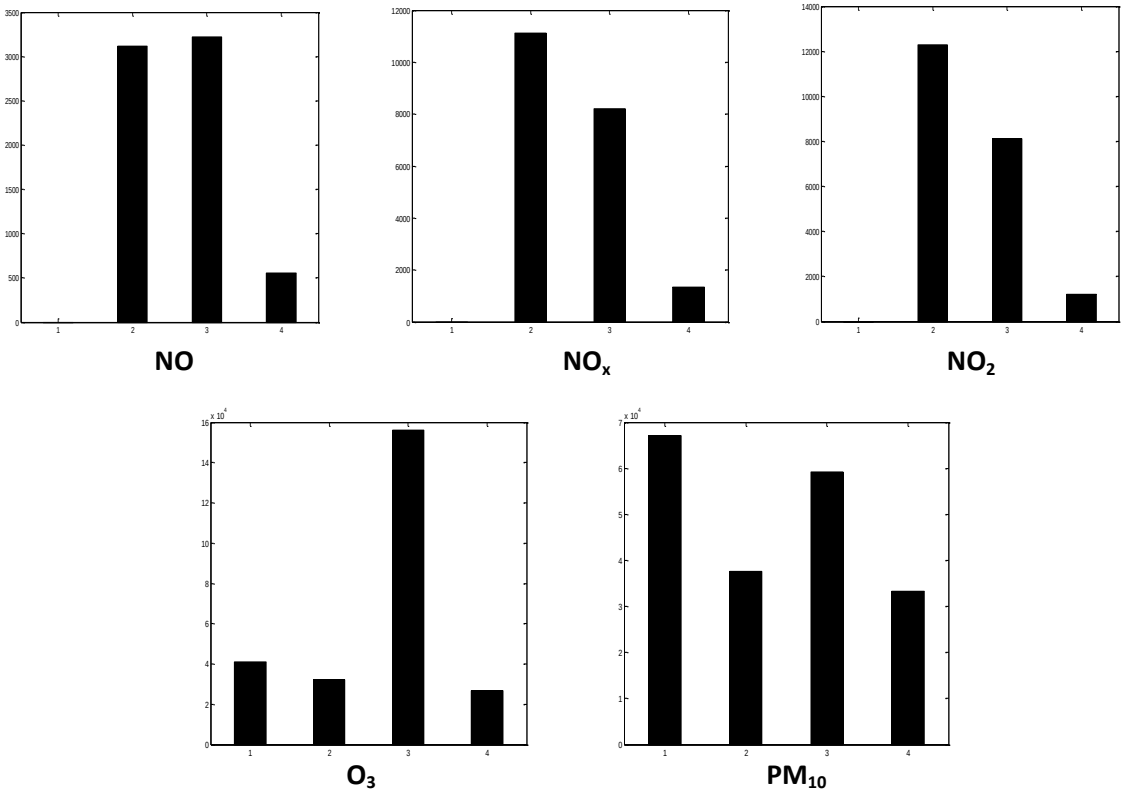


Figure 1. 5--Niveaux de concentration moyenne des polluants pendant 2004.

Le monoxyde de carbone (CO) résulte d'une combustion incomplète des combustibles et carburants. Il se fixe sur l'hémoglobine du sang à la place de l'oxygène conduisant à un manque d'oxygénation (Samasafia, 2006). Selon (Andrew et al, 1999), les concentrations du CO sont typiquement élevées dans les secteurs urbains, et atteignent leurs plus hauts niveaux à côté des routes à grand trafic ou les véhicules à essence constituent les principales sources d'émission de ce polluant. De façon saisonnière, les concentrations du CO sont très élevées en hiver pendant les conditions de stagnation. Cette conclusion peut être également extraite de la figure 1.4, où les niveaux moyens saisonniers du CO sont très élevés en hiver par rapport aux autres saisons. Concernant l'évolution journalière des concentrations de CO, deux pointes de concentration de CO sont généralement observées aux heures d'intensification du trafic du matin et du soir (Samasafia, 2007). L'ozone résulte de la transformation chimique dans l'atmosphère de certains polluants "primaires" (en particulier NO_x), sous l'effet du rayonnement solaire. Ce polluant, dit "secondaire", pénètre les voies respiratoires et provoque ainsi, des altérations pulmonaires. Les niveaux de concentration moyenne de l'ozone atteignent leurs maximums dans les mois chauds. Généralement, les pics de pollution de l'ozone ont lieu entre juin et août (Samasafia, 2007; Karatzas et Kaltsatos, 2007). Selon (Samasafia, 2007), le profil moyen journalier montre que les valeurs maximales de l'ozone sont généralement observées en début d'après-midi en raison d'un ensoleillement élevé dans cette période de la journée. Le monoxyde d'azote est dû à la réaction de l'azote et de l'oxygène dans les moteurs et les installations de combustion. Il s'avère aussi que les concentrations moyennes du monoxyde d'azote (NO) pour la période d'hiver sont plus grandes que les concentrations dans la période d'été, cette même conclusion a été aussi rapporté par (Chaloulakou et al, 2003). De nombreuses études dans la littérature ont essayé de déterminer statistiquement les sources, ainsi que les effets de la météorologie sur les concentrations du dioxyde de soufre (SO_2) (Bridgman et al, 2002; Turalioğlu et al, 2005; Gupta et al, 2005). Ce polluant provient essentiellement de la combustion des combustibles fossiles. Il est émis par les industries et le chauffage urbain, ce qui est en accord avec le caractère industriel de la région (Complexe Sidérurgique d'El-Hadjar) et les émissions domestiques (chauffage) (Mebirouk et Mebirouk-Bendir, 2007). Les niveaux de concentration du dioxyde de soufre les plus élevés sont observés en hiver en raison d'un usage accru de combustibles fossiles. Selon (Turalioğlu et al, 2005) et c'est vraiment le cas pour la région d'Annaba, ces niveaux de concentration élevés en hiver sont attribués à la consommation de plus grande quantité de carburant due à une basse température ayant pour résultat une émission élevée de SO_2. En

plus des facteurs météorologiques, un événement d'inversion est vu fréquemment dans la saison d'hiver due à la présence des hautes montagnes entourant la ville d'Annaba qui affecte négativement la distribution de l'air pollué.

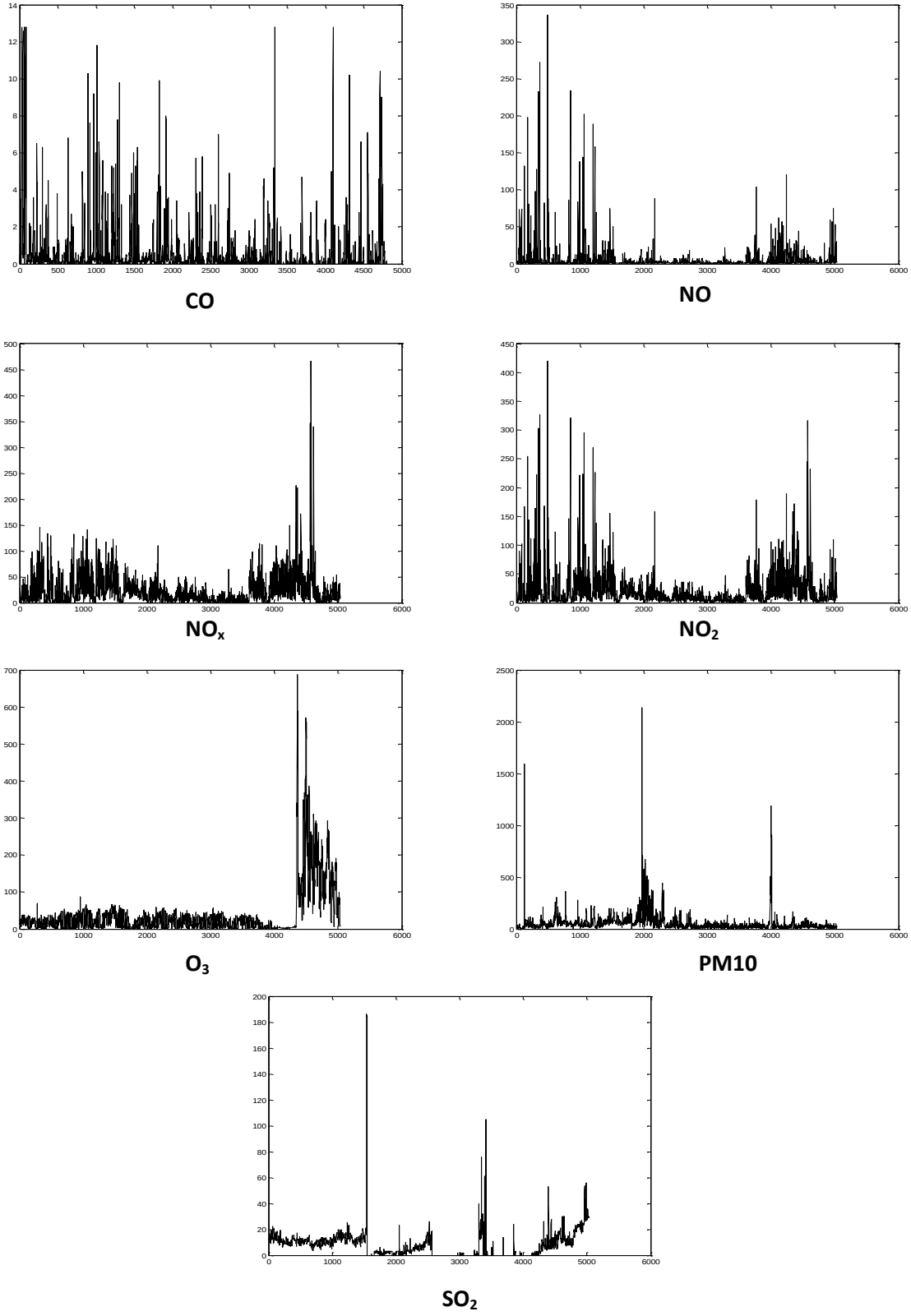


Figure 1. 6--Niveaux de concentration des polluants à Annaba pendant la période 2003-2004.

Le monoxyde d'azote (NO) émis à l'échappement s'oxyde et se transforme en dioxyde d'azote (NO_2) plus toxique. La figure 1.4 et la figure 1.5 montrent également une augmentation des teneurs en NO_2 pendant la période hivernale, ou les conditions météorologiques sont plus pénalisantes (régimes de stabilité plus fréquents) et l'activité humaine, notamment le trafic routier, est maximale. Nombreux efforts ont été consacrés à l'étude du comportement de ces polluants tels que (Hargreaves et al, 2000; Pokrovsky et al, 2002). L'oxyde d'azote (NO_x) inclut pour sa formation le NO, et le NO_2, pour cette raison des épisodes de concentrations élevées ont été marquées en hiver. Les particules en suspension (PM_{10}) peuvent provoquer une atteinte fonctionnelle respiratoire, le déclenchement de crises d'asthme notamment chez les sujets sensibles (bronchitiques chroniques, asthmatiques...). Le terme PM_{10} vient de l'anglais "Particulate Matter" et signifie donc "matière particulaire" inférieures à 10 microns (Samasafia, 2006). Les émissions de poussières sont scientifiquement mal connues. En effet, les tailles et natures des particules sont diverses, il est donc difficile de quantifier leur origine et les quantités émises. Un cycle hebdomadaire de concentrations est manifesté dans la plupart des sites urbains ou les concentrations de PM_{10} sur 24 heures sont plus faibles durant la fin de semaine que durant la semaine, cette différence est amplifiée pour les sites routiers. De façon saisonnière et d'après (Samasafia, 2006) et la figure 1.4, ainsi que la figure 1.5, les concentrations les plus élevées sont mesurées pendant l'été.

1.6 Conclusion

La dégradation de la qualité de l'air liée aux activités de l'homme est de plus en plus au cœur des préoccupations des spécialistes de la santé publique et des agences chargées de la protection de l'environnement. L'industrie est l'élément moteur de croissance et de dégradation de l'environnement dans la ville d'Annaba et sa région. En outre, la pollution atmosphérique à Annaba a été aggravée par sa position géographique qui prend la forme d'une cuvette. Cette topographie favorise ainsi la stagnation de l'air pollué et l'accumulation des taux de concentration des polluants. Par conséquent, il est extrêmement important de considérer l'effet des conditions météorologiques sur la pollution atmosphérique, puisqu'elles influencent directement la capacité de dispersion de l'air polluant.

Première partie

Etat de l’art : Clustering et classification non supervisée

Depuis l'apparition de l'informatique, l'ensemble de données stockées sous forme numérique ne cesse de croitre de plus en plus rapidement partout dans le monde. Les individus mettent de plus en plus les informations qu'ils possèdent à disposition de tout le monde via le web. De nombreux processus industriels sont également contrôlés par l'informatique. Et de nombreuses mesures effectuées un peu partout dans le monde, comme par exemple les mesures météorologiques qui remplissent des bases de données numériques importantes. Si le problème n'est pas bien défini, il s'avère qu'une augmentation de la quantité de données peut avoir un effet négatif sur la compréhension de données. Ceci est généralement le cas pour les données multidimensionnelles en particulier. Si le but est simplement l'interprétation d'un ensemble de données pour générer des hypothèses raisonnables ou de trouver de nouvelles modèles de données, il semble paradoxalement que plus les données sont disponibles, plus il est difficile de comprendre l'ensemble des données du fait que les structures sont cachées parmi les grandes quantités de données multidimensionnelles. Il existe dès lors un grand intérêt à développer des techniques permettant d'utiliser aux mieux tous ces stocks d'informations telles que la classification automatique, afin d'en extraire un maximum de connaissances utiles (Candillier, 2006).

La classification automatique - clustering - est une étape importante dans le processus d'Extraction de Connaissances à partir de Données (ECD). Elle vise à développer un partitionnement optimal, c'est à dire le regroupement des données en classes qui partagent des caractéristiques similaires où les données sont généralement représentées par des vecteurs de mesures ou des points dans un espace multidimensionnel. Intuitivement, les vecteurs appartenant à un cluster valide sont plus semblables les uns aux autres qu'un vecteur appartenant à un groupe diffèrent (Jain et al, 1999; Xu et Wunsch, 2005). En d'autres termes, l'objectif des méthodes de classification automatique (dites aussi méthodes d'apprentissage non supervisées) dans les applications de datamining est d'identifier les groupes d'un ensemble non étiqueté de vecteurs de données qui partagent des similarités sémantiques. Cela permet à l'utilisateur de construire un modèle cognitif, favorisant ainsi la détection de la structure inhérente d'un ensemble de données. Cependant, dans de nombreuses applications peu ou pas d'information préalable sur les modèles à obtenir sont disponibles. La classification non supervisée est liée à plusieurs domaines de recherche. Elle a été utilisée couramment dans les statistiques (Arabie et Hubert, 1996), la reconnaissance de formes, la segmentation et le traitement d'images (Ganti et al, 1999), et dans plusieurs autres domaines d'application tel que la recherche documentaire, forêt, agriculture (Recknagel, 2002; Suwardi

et al, 2007), qualité de l'eau (Aguilera et al, 2001; Tison et al, 2005), identification des types de jours des charges électriques (Khadir et al, 2006) et bien d'autres domaines.

Dans le cas des données environnementales, leur analyse en utilisant les outils de la classification automatique peut aider à mieux comprendre les phénomènes généraux qui régissent le climat, afin, par exemple, d'anticiper les phénomènes extrêmes et d'agir en conséquence pour les populations concernées. La classification automatique a été utilisée avec succès dans de nombreuses études météorologiques tel que dans (Eder et al, 1994). Les méthodes de classification partitives ont été utilisées dans (Ludwig et al. 1995) pour la classification des paramètres météorologiques et les concentrations de l'ozone. Aussi bien les méthodes de classification hiérarchique ont été utilisées par Yu et Chang (2001) pour analyser les niveaux de concentration des PM_{10} à Taiwan où ils ont regroupé les vecteurs de mesures de PM_{10} en cinq classes. Ainsi, Les méthodes hiérarchiques et partitives ont été utilisées, et comparées dans le but de l'identification des conditions météorologiques à Houston (Davis et al, 1998) où sept régimes météorologiques ont été extraits. D'autres méthodes d'extraction de connaissances à partir de données ont été utilisées afin d'analyser les données environnementales telle que l'analyse en composante principale (ACP) (Reljin et al, 2003). L'ACP a prouvé son utilité comme une technique puissante d'ECD (kwan et al, 2001), cependant cette méthode n'est pas bien appropriée pour la visualisation des structures complexes d'un ensemble de données (Laitinen et al, 2002).

Au cours des deux dernières décennies les Réseaux de Neurones Artificiels (RNA) ont prouvé leurs utilités dans le traitement des bases de données volumineuses à plusieurs variables. Les tâches pour lesquelles les RNA ont été jugés particulièrement efficaces sont : la modélisation non linéaire et le clustering/classification. L'algorithme des cartes auto-organisatrices de Kohonen (Self-Organizing Maps (SOM)) représente la classe d'apprentissage non supervisé des réseaux de neurones dont la caractéristique principale est son capacité à mapper les relations non linéaire d'un ensemble de données multidimensionnelles dans une grille de neurones à deux dimensions facilement visualisable. Les réseaux SOM sont également appelés cartes auto organisatrices topologiques puisque la fonction de base d'un SOM est d'afficher la topologie d'un ensemble de données, c'est à dire les relations entre les membres de l'ensemble. L'algorithme SOM a été développé par Kohonen dans les années 1980, et depuis lors, il a été utilisé comme un outil de reconnaissance de formes et de classification dans différents domaines tels que la robotique, l'astronomie, et la chimie (Candillier, 2006 ; Oja et al, 2003). Les cartes auto-organisatrices

(SOM) ont suscité beaucoup d'intérêt des travaux de recherches dans plusieurs domaines. Par conséquent l'algorithme SOM a été largement analysé, ainsi qu'un certain nombre de variantes ont été développés et, surtout, il a été largement appliqué dans des domaines aussi variés que les sciences de l'ingénieur, la médecine, la biologie et l'économie (Oja et al, 2003). SOM est largement utilisé pour l'analyse et la visualisation de données complexes. Des milliers d'applications de la carte de kohonen dans différentes disciplines peuvent être trouvées dans l'étude de Kaski et al. (1998). La dernière décennie a connu une augmentation énorme des applications SOM. Aujourd'hui, SOM est souvent utilisé comme un outil statistique pour l'analyse multivariée, car il s'agit à la fois d'une méthode de projection permettant de mapper les données de grande dimension à un espace de faible dimension, ainsi qu'une méthode de classification qui regroupe les vecteurs de données similaires dans une même unité ou dans des unités voisine. Les applications du SOM sont devenues de plus en plus très utiles dans plusieurs domaines tels que la reconnaissance de formes, le Webmining et les géosciences, où il a été prouvé qu'il s'agit d'une technique d'extraction de caractéristiques efficace qui présente de nombreux avantages par rapport aux autres méthodes d'analyses données classiques.

Le SOM a été introduit pour la météorologie et les sciences climatiques vers la fin des années 1990 comme une méthode de reconnaissance des formes et de clustering (Liu et Weisberg, 2011). En tant que nouveau algorithme de clustering, un certain nombre d'articles ont mis l'accent sur la démonstration de l'algorithme SOM et ses sorties, avec l'application de cet algorithme sur une problématique de recherche particulière comme un objectif secondaire (Hewitson et Crane, 2002 ; Turias et al, 2006). L'algorithme SOM est alors apparu comme un outil très utile pour les applications météorologiques à différentes échelles spatiales et temporelles: la climatologie synoptique, les conditions météorologiques extrêmes et l'analyse des modèles de précipitations, classification de nuage, ainsi que l'analyse des changements climatiques (Liu et Weisberg, 2011). De nombreux types de données météorologiques ont été analysés en utilisant les SOMs, par exemple, la pression observée et modélisée au niveau de la mer, la hauteur du géopotentiel à différents niveaux de pression, la température de l'air, l'humidité, les précipitations, l'évaporation, la neige, la glace de mer, etc. Géographiquement, les applications météorologiques utilisant la méthode SOM se trouvent partout dans le monde: Amériques, Afrique, Asie, Europe de l'Arctique et de l'Antarctique.

Au cours des dernières années, les SOM ont gagnés en popularité dans les communautés météorologique et océanographique comme des outils puissants d'extraction de

caractéristiques. Au-delà des avantages traditionnels des anciennes méthodes, l'auto-organisation qui est une partie inhérente du processus du SOM, offre plusieurs autres avantages. Peut-être le développement le plus important avec SOM est les nouvelles possibilités de visualisation, les modèles peuvent être beaucoup plus facilement compris par les experts quand ils sont affichées dans le SOM réduisant ainsi le temps nécessaire pour le traitement d'un ensemble de données quelconque. À ce jour, peu d'études ont comparé directement l'efficacité du SOM avec d'autres méthodes. Parmi quelques études, Reusch et al (2007) ont comparé directement les modèles de l'ACP aux modèles du SOM, et ils ont découvert que les deux méthodes présentent des avantages uniques, bien que les SOMs aient montré une meilleure visualisation des données par rapport à l'ACP. Les performances du SOM ont été également comparées avec celles de l'algorithme de clustering des K-means. Bien que les différences ne sont pas très importantes, il a été remarqué que la carte de Kohonen a obtenu de meilleurs résultats par rapport à l'algorithme K-means (Turias et al, 2006). Il est à noter que SOM et l'algorithme K-means sont rigoureusement identiques lorsque le rayon de la fonction voisinage dans le SOM est égal à zéro (Boinee, 2006). Dans ce cas, l'adaptation des poids lors de l'apprentissage se produit uniquement dans l'unité gagnante comme il arrive dans les K-means. SOMs ont l'avantage d'afficher facilement les sorties en grilles bidimensionnelles. Toutefois K-means fonctionne comme un SOM sans préservation de topologie et donc sans visualisation facile.

Dans cette partie, nous ne nous intéressons pas à fournir une liste exhaustive de l'ensemble des notions et méthodes existantes dans le cadre du clustering, mais plutôt un aperçu général des différentes problématiques dans ce cadre tout en focalisons sur la carte au-organisatrice de kohonen qui constitue notre centre d'intérêt. En premier lieu, nous présentons dans le deuxième chapitre une vue générale de la classification automatique. En dernier lieu, nous présentons dans troisième chapitre les principales méthodes de clustering les plus classiques.

Chapitre 2

Vue générale du problème

2.1 Vue générale du problème

Le but de ce chapitre est de présenter une vue générale du problème du clustering, en introduisant les notions et les concepts de base sur lesquels s'appuiera la suite de cette thèse, et de mettre en évidence la diversité qui existe parmi les différentes méthodes de classification non supervisée en mettant l'accent sur les cartes auto-organisatrices de kohonen qui constitue l'axe principal de la présente étude. Le lecteur intéressé est invité à consulter l'une des nombreuses références disponibles (Jain et al, 1999; Berkhin, 2002; Liu et Weisberg, 2011; Xu et Wunsch, 2005; Grabmeier et Rudolph, 2002 ; Guérif, 2006 ; Boubou, 2006) pour approfondir son étude.

Nous commençons ce chapitre par rappeler quelques concepts et définitions essentiels pour comprendre les différentes méthodes et outils de classification automatique. Nous présentons par la suite les étapes de base de la procédure de classification automatique ainsi que les différentes applications possibles du clustering. Puis, nous discutons les différentes notions qui sont utilisées pour définir la similarité entre objets, qui constitue la base de toute méthode de clustering. Enfin, nous présentons les approches existantes permettant d'évaluer les résultats d'algorithmes de clustering. De plus amples détails sur les principales méthodes de clustering sont ensuite présentés dans le chapitre suivant.

2.2 Concepts et définitions utiles

La classification non supervisée ou classification automatique - clustering - est une étape importante de l'analyse de données; dont l'objectif est d'identifier des groupes d'objets similaires appelé clusters d'un ensemble de données sans en connaître la structure au préalable (Guérif, 2006). Le concept identification des types de jours météorologiques utilisé dans cette thèse est étroitement lié à la notion de partition ou classification d'un ensemble fini et nous utiliserons ces deux termes de manières interchangeables tout au long de cette thèse. La définition qui suit correspond à la notion de classification dure mais ce qualificatif ne sera plus précisé dans la suite du travail.

2.2.1 Définition d'une partition

Étant donnée un ensemble fini d'objets noté *I*, on appelle partition de *I* toute famille de parties *P* non vides disjointes deux à deux dont l'union forme l'ensemble *I*.

P= $\{C_i, i \in I\}$, tel que C_i : partie de *I* (ou une classe) possédant les propriétés suivantes:

1. $\forall\, i \in I, C_i \neq \emptyset$ 2. $\forall\, i \in I, \forall\, j \in I,\ i \neq j \rightarrow C_i \cap C_j = \emptyset$

3. $\cup_{i \in I}\, C_i = I$ 4. $\forall\, i \in I, C_i \neq \emptyset$

2.3 Classification supervisée ou non supervisée

La classification de données situées dans un espace de grande dimension est un problème délicat qui apparaît dans de nombreuses sciences dont l'objectif général est d'être capable d'étiqueter des données en leur attribuant une classe. Il existe deux types d'approches : la classification supervisée et la classification non supervisée. Ces deux approches se différencient par leurs méthodes et par leur but.

2.3.1 La classification supervisée (ang. classification)

Si les classes possibles sont connues et si les exemples sont fournis avec l'étiquète de leur classe on parle de classification supervisée (ang. classification), L'objectif est alors d'apprendre à l'aide d'un modèle d'apprentissage à partir de l'ensemble des exemples (appelé ensemble d'apprentissage) des règles qui permettent de prédire la classe des nouveaux exemples ce qui revient à découvrir la structure des classes afin de pouvoir généraliser cette structure sur un ensemble de données plus large.

Un exemple d'application de la classification supervisée concernant les voitures peut être tiré de (Candillier, 2006), ou il peut s'agir par exemple de déterminer si une nouvelle voiture rencontrée fait partie de la classe des citadines, des voitures intermédiaires ou des voitures confortables, en se basant sur des caractéristiques, et sur la classe connue des voitures déjà rencontrées (exemples d'apprentissage). Le tableau 2.1 présente le problème dans ce cadre. À chaque exemple est associée la classe de voiture à laquelle il appartient. L'objectif est alors d'être capable d'estimer la classe la plus appropriée à tout nouvel exemple rencontré (par exemple la voiture 7 dans le tableau 2.1).

Tableau 2. 1--exemple d'application de l'apprentissage supervisée1

identifiant	carburant	cylindres	longueur	puissance	Classe
1	gpl	8	186	6000	confort
2	Essence	4	170	5800	intermédiaire
3	diesel	6	172	5500	intermédiaire
4	diesel	4	156	5200	citadine
5	Essence	12	190	5500	confort
6	Essence	4	175	5800	intermédiaire
7	diesel	6	170	6000	?

2.3.2 La classification non supervisée (ang. clustering)

Si seulement les exemples, sans étiquettes, sont disponibles, et si les classes et leurs nombres sont inconnus, on parle alors de classification non supervisé appelée aussi "classification automatique", "clustering" ou encore "regroupement". La classification non supervisée (ang. clustering) consiste à diviser un ensemble d'exemples en sous-ensembles, appelés classes (clusters), tels que les objets d'une classe sont similaires et que les objets de classes différentes sont différents, afin d'en comprendre la structure (Blansché, 2006). Autrement dit, il s'agit à ce niveau de rechercher la distribution sous-jacente des exemples dans leurs espaces de description. Comme le montre le tableau 2.2, il n'y a aucune information de classe associée aux exemples dans ce cas (Candillier, 2006).

Tableau 2. 2--Exemple d'application d'apprentissage supervisée

identifiant	carburant	cylindres	longueur	puissance	Classe
1	gpl	8	186	6000	?
2	Essence	4	170	5800	?
3	diesel	6	172	5500	?
4	diesel	4	156	5200	?
5	Essence	12	190	5500	?
6	Essence	4	175	5800	?
7	diesel	6	170	6000	?

2.4 Les étapes d'une classification automatique

Les étapes de base du processus de classification automatique sont présentées dans la figure 2.1 et peuvent être récapitulées comme suit (Boubou, 2006 ; Halkidi et al, 2001, Jain et al, 1999; Xu et Wunsch, 2005) :

Sélection/extraction des caractéristiques. La sélection des caractéristiques est le processus d'identification d'un sous ensemble optimal de caractéristiques d'origine pertinentes pour un critère fixé auparavant pour les utiliser dans le regroupement. La sélection de ce sous ensemble de caractéristiques permet d'éliminer les informations non pertinentes et redondantes selon le critère utilisé. Tandis que l'extraction des caractéristiques vise à l'utilisation d'une ou plusieurs transformations des caractéristiques d'entrées pour produire de nouvelles caractéristiques saillantes. Chacune de ces techniques ou les deux peuvent être utilisées pour obtenir un ensemble approprié de caractéristiques à utiliser dans le clustering.

Algorithme de classification automatique. L'étape de regroupement ou classification peut être effectuée de plusieurs façons. La classification de données (ou clustering) peut être dure (une partition de données en groupes) ou floue (où chaque modèle à un degré d'appartenance à chacun des clusters de sortie). Ainsi, notre objectif à travers cette étape est de choisir l'algorithme de clustering le plus approprié pour le regroupement de l'ensemble de données où chaque algorithme de clustering est caractérisé principalement par une mesure de proximité et un critère de regroupement.

- La mesure de proximité est une mesure qui quantifie à quel degré deux points de données quelconque sont "similaire" (vecteurs des caractéristiques). Dans la plupart des cas nous devons nous assurer que tous les variables choisis contribuent également au calcul de la mesure de proximité.
- Critère de regroupement peut être exprimé par une fonction de coût ou d'autre type de règles. il est nécessaire de prendre en compte le type des clusters attendus par le regroupement de l'ensemble de données. Ainsi, nous pouvons définir un "bon" critère de regroupement, menant à un partitionnement qui représente le mieux que possible l'ensemble de données.

Validation des résultats. Les algorithmes de regroupement permettent d'extraire des clusters qui ne sont pas connus à priori. En outre, différentes approches conduisent généralement à différents groupes et même pour le même algorithme. Par conséquent une classification finale d'un ensemble de données exige un certain genre d'évaluation dans la plupart des applications. L'exactitude des résultats obtenus par les algorithmes de regroupement est vérifiée en utilisant des techniques et des critères bien appropriés (Halkidi et al, 2001; Xu et Wunsch, 2005).

Interprétation des résultats. Le but ultime du regroupement est de fournir aux utilisateurs un aperçu significatif des données d'origine, afin qu'ils puissent résoudre efficacement les problèmes rencontrés.

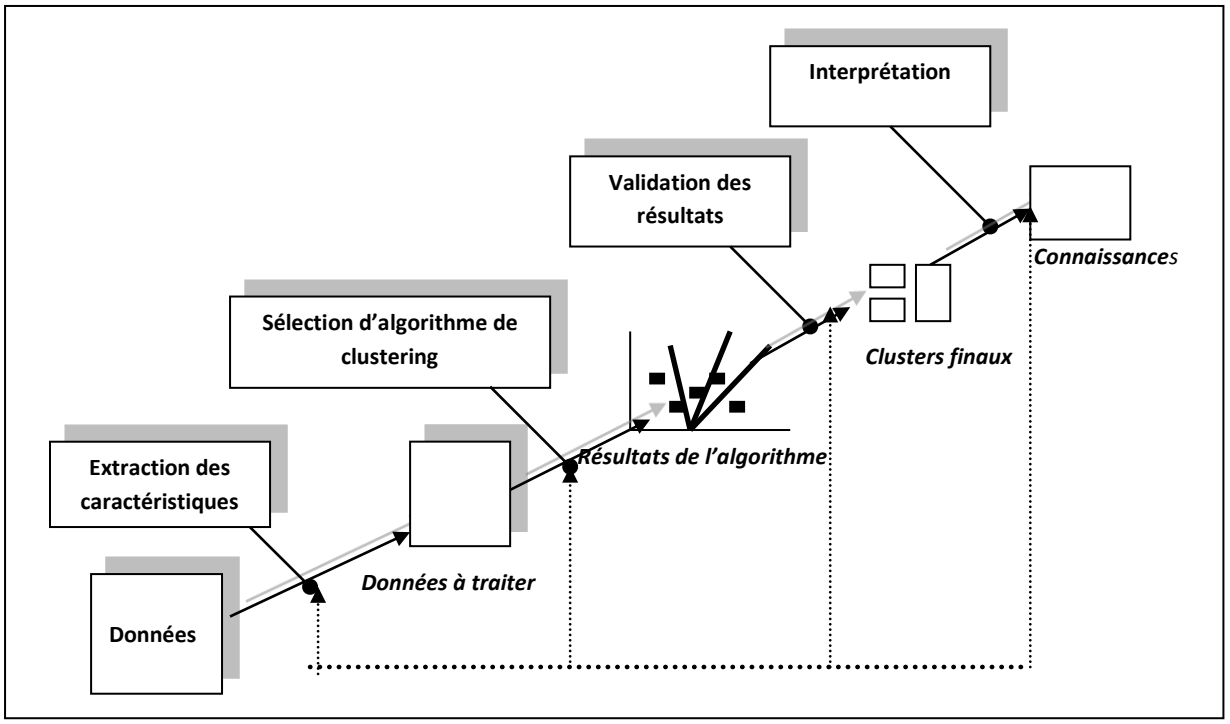


Figure 2. 1--Les étapes d'un processus de classification automatique.

2.5 Application du clustering

La classification non supervisée a été utilisées dans plusieurs domaines, allant de l'ingénierie (apprentissage automatique, intelligence artificielle, reconnaissance des formes, génie mécanique, génie électrique), l'informatique (Webmining, base de données spatiale, la collection de documents textuels, la segmentation des images), sciences médicales et vie (génétique, la biologie, la microbiologie, la paléontologie , la psychiatrie, clinique, pathologie), sciences de la terre (Géographie. géologie, télédétection), sciences sociales (sociologie, la psychologie, l'archéologie, l'éducation), et l'économie (marketing, commerce). Cette diversité reflète la position importante du regroupement dans la recherche scientifique. D'autre part, cette diversité peut être une source de confusion, en raison des terminologies et objectifs différents. Les algorithmes de clustering ont été développés pour résoudre des problèmes particuliers, dans des domaines spécifiques, et ils sont généralement basés sur des hypothèses et des suppositions sur l'ensemble de données à traiter. Ces suppositions affectent inévitablement les performances de ces algorithmes dans d'autres problèmes qui ne satisfont pas ces hypothèses. Par exemple, l'algorithme K-means basé sur la distance euclidienne et, par conséquent, il tend à générer des clusters hyper sphériques. Mais si les clusters réels sont dans une autre forme géométrique, K-means peut ne pas être efficaces, ce qui conduit à chercher d'autres algorithmes de clustering. Dans ce qui suit, nous décrivons quelques domaines

d'applications où le clustering a été utilisé comme une étape essentielle (Berkhin, 2002 ; Jain et al, 1999 ; Candillier, 2006)

- La segmentation,
- La reconnaissance de formes et,
- Le datamining.

La segmentation. La segmentation d'une base de données a pour objectif principal de réduire la taille des données afin de faciliter leur traitement. On parle dans ce cas de compression de données. Par exemple cette technique est très utile en segmentation d'image qui peut être défini comme un partitionnement exhaustive d'une image d'entrée en plusieurs régions, chacune d'elles étant considérées homogènes par rapport à une propriété d'intérêt de l'image considérée (par exemple, l'intensité, la couleur ou la texture). L'utilisabilité de la méthodologie de regroupement dans le problème de segmentation d'image a connu plus de trois décennies, et les paradigmes sous-tendent les efforts des premiers chercheurs sont encore en usage aujourd'hui. Enfin, la segmentation est parfois utilisée pour discrétiser une base de données, c'est à dire transformer la description complexe des objets par un unique attribut caractérisant leurs appartenance à une classe identifiée automatiquement.

Le datamining. Vise à traiter les bases de données volumineuses qui imposent au clustering des exigences computationnelles supplémentaires. Ces défis ont conduit à l'émergence de puissantes méthodes de classification non supervisée largement utilisées afin de faciliter le regroupement de grands ensembles de données caractérisés par de nombreux attributs de différents types. La naissance du clustering dans le data mining est due principalement aux développements intenses dans les domaines de recherche d'informations et de Texte Mining, les applications de bases de données spatiales par exemple les données astronomiques, l'analyse de données, les applications Web, l'analyse de l'ADN en bioinformatique, et bien d'autres applications spécifiques.

La reconnaissance de formes. Dans de nombreuses applications de reconnaissance de formes, il est extrêmement difficile ou coûteux, voire impossible, d'étiqueter de manière fiable un échantillon d'apprentissage avec sa véritable catégorie. Les applications typiques incluent la reconnaissance vocale et de caractère. Les algorithmes d'apprentissage de classification non supervisée ont été également utilisés pour la segmentation des images et vision par ordinateur (Jain et al, 1999). L'analyse typologique est une technique très importante et très utile. La vitesse, la fiabilité et la consistance avec laquelle un algorithme de clustering peut

organiser de grandes quantités de données constituent de fortes raisons écrasantes de l'utiliser dans des applications de reconnaissance de formes.

2.6 Présentation des méthodes de classification de données

La classification non supervisée est un domaine très actif qui a engendré un nombre très important de travaux de recherches. De nombreuse méthodes et approches ont été définies, et il serait difficile de présenter une liste exhaustive ici. Cependant nous pouvons distinguer différentes méthodes couramment utilisées. Les premières approches proposées étaient algorithmiques, heuristiques ou géométriques et reposaient essentiellement sur la dissimilarité entre les objets à classer. Plus récemment les modèles probabilistes sont utilisés par l'approche statistique. Différents points de vue et critères de départ conduisent généralement à différentes taxonomies des algorithmes de clustering. Ces algorithmes peuvent être classé selon (Boubou, 2006 ; Halkidi, 2000) :

- Le type de données d'entrée à l'algorithme de classification.
- Le critère de regroupement définissant la similarité entre les objets.
- Les théories et les concepts fondamentaux sur lesquels les techniques de regroupement sont basées (par exemple la théorie floue, statistique).

2.6.1 Résultat de la classification

Les résultats d'une classification peuvent être représentés de différentes façons, selon qu'il y ait des chevauchements entre les classes ou non (classification dure ou floue), et selon le fonctionnement de l'algorithmique de clustering (agglomératif vs divisif). Par ailleurs, les méthodes de classification peuvent être déterministes ou stochastiques (Jain et al, 1999; Blansché, 2006).

Agglomératif vs divisif. Cet aspect concerne la structure et le fonctionnement de l'algorithmique de clustering. Une approche agglomérative commence en assignant chaque élément à un cluster distinct (singleton), et fusionne successivement les clusters jusqu'à ce qu'un critère d'arrêt soit satisfait. Tandis qu'une méthode divisive commence en prenant tous les objets comme un seul cluster et effectue ensuite le fractionnement (division) jusqu'à ce qu'un critère d'arrêt soit atteint.

Dure vs floue. Dans une classification dure (ang. Hard clustering), chaque objet est attribué à une et une seule classe. Tandis que dans une méthode de classification floue (ang. Fuzzy clustering), chaque objet peut appartenir à plusieurs classes avec un certain degré

d'appartenance. La classification floue est difficilement interprétable par l'utilisateur et souvent transformée en classification dure en assignant chaque objet au cluster dont son degré d'appartenance est maximum.

Déterministe vs stochastique. Cette question est plus pertinente aux approches de classification par partition conçues pour optimiser une fonction d'erreur quadratique. Cette optimisation peut être réalisée en utilisant des techniques traditionnelles ou par une recherche aléatoire de l'espace d'état composé de tous les étiquètes possibles.

Incrémentale vs non-incrémental. Ce problème se pose lorsque la base de données à regrouper est volumineuse, et que les contraintes du temps et espace d'exécution affectent l'architecture de l'algorithme. L'histoire des méthodologies de regroupement ne contient pas de nombreux exemples d'algorithmes de clustering conçus pour fonctionner avec des ensembles de données volumineux, mais l'avènement du data mining a favorisé le développement d'algorithmes de clustering qui minimisent le nombre de balayages sur l'ensemble de données, et réduit par conséquent, le nombre d'objets examinés au cours de l'exécution.

Monothétique vs polythétique. Cet aspect concerne l'utilisation séquentielle ou simultanée des caractéristiques dans le processus de regroupement. La plupart des algorithmes sont polythétique, c'est-à-dire, toutes les caractéristiques sont impliquées dans le calcul des distances entre les exemples et les clusters. Un algorithme monothétique simple considère les caractéristiques de façon séquentielle lors de la classification de données.

Ainsi selon la méthode adoptée pour définir les clusters, les algorithmes de regroupement peuvent être largement classés dans les catégories définies ci-après.

2.6.2 Quelques approches classiques

Algorithmes hiérarchiques

Cette classe d'algorithmes consiste à créer une décomposition hiérarchique d'un tableau de données. On peut envisager deux stratégies : ascendante ou descendante. La classification hiérarchique ascendante procèdent successivement par fusionnement de plus petits clusters dans les plus grands, Le résultat de l'algorithme est un arbre de clusters, appelés le dendrogramme, qui montre comment les clusters sont reliés. Tandis que l'approche descendante démarre avec tous les objets dans une seule et même classe. A chaque itération, une classe est décomposée en classes plus petites, jusqu'à n'avoir plus qu'un seul objet dans chaque classe, ou éventuellement qu'une condition d'arrêt soit vérifiée.

- Classification Hiérarchique Ascendante (CHA),
    -Clustering Using Representatives (CURE)
    -Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)
    -Robust Clustering using links (ROCK)
- classification hiérarchique descendante (CHD)
    - Williames et Lambert
    - Tree Structured Vector Quantization (TSVQ)

Algorithmes par partition

Tente à décomposer directement l'ensemble de données en un ensemble disjoint de clusters. Plus spécifiquement, ils essayent de déterminer un nombre entier de partitions qui optimisent une fonction objective.

- K-means
- K-médoides,
- Partition Around Medoid(PAM).
- Clustering large applications based upon randomized search (CLARANS)
- Clustering LARge Applications (CLARA)

Algorithmes basées sur la densité

L'idée principale de ce type de regroupement est de regrouper les objets voisins d'un ensemble de données dans des clusters basés sur des états de densité.

Classification basée sur la quantification par grille

L'idée de ces méthodes est qu'on divise l'espace de données en un nombre fini de cellules formant une grille. Ce type d'algorithme est conçu pour des données spatiales. Une cellule peut être un cube, une région, un hyper rectangle. Ces deux derniers types de méthodes ne seront pas détaillés par la suite.

Autres méthodes …

2.7 Les mesures de similarité

L'objectif principal d'une classification est de fournir des groupes homogènes et bien séparés, en d'autre terme des groupes d'objets tel que (Guérif, 2006 ; Boubou, 2006 ; Blansché, 2006):

- Les objets soient les plus similaires possibles au sein d'un groupe
- Les groupes soient aussi dissemblables que possible

En raison de la variété des types des caractéristiques et des échelles, la mesure de distance (ou mesures) doivent être choisis avec précaution. Malheureusement, trop souvent, il s'agit d'un choix arbitraire, sensible à la représentation des objets, et qui traite tous les attributs de la même manière. Il est plus fréquent de calculer la dissimilarité entre deux objets en utilisant une mesure de distance définie sur l'espace des caractéristiques (voir le tableau 2.3).

Un objet est décrit par un ensemble de caractéristiques, généralement représentées par un vecteur multidimensionnel. Les caractéristiques peuvent être quantitatives ou qualitatives, continues ou binaires, nominales ou ordinales, qui déterminent les mécanismes de mesure correspondante. La distance ou la fonction de dissimilarité d'un ensemble de données ***X*** est définie pour satisfaire les conditions suivantes (Xu et Wunsch, 2005) :

1) Symétrie $d(x_i, x_j) = d(x_j, x_i)$;

2) Positivité $d(x_i, x_j) >= 0$ pour tout x_i et x_j. Si les conditions

3) Inégalité triangulaire $d(x_i, x_k) \le d(x_i, x_j) + d(x_j, x_k)$

4) Réflexivité $d(x_i, x_j) = 0$ si $x_i = x_j$ sont vérifiées, alors il est appelé une métrique

De même, une fonction de similarité est définie pour satisfaire aux conditions ci-après.

1) Symétrie $S(x_i, x_j) = S(x_j, x_i)$;

2) Positivité $S(x_i, x_j) >= 0$ pour tout x_i et x_j. Si les conditions

3) Inégalité triangulaire $S(x_i, x_k) \le S(x_i, x_j) + S(x_j, x_k)$

4) Réflexivité $S(x_i, x_j) = 0$ si $x_i = x_j$ sont vérifiées, alors il est appelé métrique de similarité

Pour un ensemble de données avec *N* objet d'entrées, nous pouvons définir une matrice symétrique de taille $N \times N$, appelée matrice de proximité, dont le *(i,j)*ème élément représente la similarité ou la mesure de dissimilarité pour le ième et le jème objets (*i,j=1,...N*).

En règle générale, les fonctions de distance sont utilisées pour mesurer les variables continus, tandis que les mesures de similarités sont plus importantes pour les variables qualitatives. Quelques mesures typiques pour les fonctions continues sont représentées dans le tableau 2.3.

Tableau 2. 3--Mesures de similarité et de dissimilarité

Mesures	Formes	Commentaires	Exemples et applications
La distance de Minkowski	$D_{ij} = \left(\sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|^{1/n} \right)^n$	métrique invariante à toute rotation seulement pour n = 2 (distance euclidienne). Les caractéristiques de grandes valeurs et variances ont tendance à dominer les autres caractéristiques.	c-means floue

La distance euclidienne	$D_{ij} = \left(\sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|^{1/2} \right)^2$	La distance la plus connue, ce n'est qu'un cas particulier pour p = 2 de la distance de Minkowski. Tend à former des clusters hyper sphériques.	K-means
La distance de Manhattan	$D_{ij} = \sum_{l=1}^{d} \left\| x_{il} - x_{jl} \right\|$	c'est un cas particulier pour p = 1 de la distance de Minkowski. Tend à former des clusters hyper-rectangulaires.	ART floue
la distance de Tchebychev	$D_{ij} = \max_{1 \le l \le d} \left\| x_{il} - x_{jl} \right\|$	cas particulier pour $p = \infty$ de la distance de Minkowski,	c-means floue avec la distance de Tchebychev
La distance de mahalanobis	$D_{ij} = (x_i - x_j)^T S^{-1} (x_i - x_j)$, où S est la matrice de covariance à l'intérieur du groupe.	S est calculé sur la base de tous les objets. Tend à former des clusters hyper-ellipsoïdes. lorsque les caractéristiques ne sont pas corrélées, la distance de mahalanobis carrée est équivalente à la distance euclidienne au carré. Peut causer une certaine charge de calcul.	ART ellipsoïde, algorithme de clustering hyper-ellipsoïde.
Corrélation de Pearson	$\boldsymbol{D_{ij} = \frac{1 - r_{ij}}{2}, ou}\ r_{ij} = \frac{\sum_{l=1}^{d} (x_{il} - \bar{x}_j)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{d} (x_{il} - \bar{x}_j)^2 \sum_{l=1}^{d} (x_{il} - \bar{x}_j)^2}}$	N'est pas une métrique. Calcule la distance entre un objet x_i et un point de référence x_r. d_{ir} est minimisé quand un objet symétrique existe.	largement utilisé comme une mesure pour l'analyse de données d'expression génique
Cosinus de similarité	$S_{ij} = \cos \alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$	Mesure indépendante de la longueur du vecteur. Invariant à la rotation, mais pas à la transformation linéaire	la mesure la plus utilisée dans le clustering des documents.

2.8 Évaluation et critères de validités

Le clustering est un processus non supervisé où les données ne sont pas étiquetées, et aucune information structurelle n'est disponible. Ainsi, l'évaluation des résultats des algorithmes de clustering est une tâche très importante. Dans le processus de regroupement, il n'y a pas donc de classes prédéfinies, de plus la plupart des algorithmes ne fournissent pas les moyens pour la validation et l'évaluation du regroupement. Par conséquent, il est difficile de trouver une mesure appropriée pour vérifier l'exactitude du regroupement obtenu. Dans ce cas-là, plusieurs questions peuvent être posées telles que :

- Quel est le nombre optimal de clusters?,
- Quel est le meilleur cluster ?,
- Quel est le meilleur regroupement de données?

Dans la plupart des algorithmes, des évaluations expérimentales des données bidimensionnelles sont utilisées afin que le l'utilisateur soit en mesure de vérifier visuellement la validité des résultats (c'est-à-dire à quel point l'algorithme de clustering a découvert les clusters de l'ensemble de données). Il est clair que la visualisation de l'ensemble de données est une vérification cruciale des résultats de clustering. Dans le cas de base de données multidimensionnels (par exemple plus de trois dimensions) une visualisation efficace de l'ensemble de données pourrait être difficile. En outre, la perception de clusters est une tâche difficile pour les êtres humains qui ne sont pas habitués à des espaces de dimensions élevés (Halkidi et al, 2001, Wang et Zhang, 2007).

2.8.1 Détermination du nombre de clusters

La classification automatique vise à identifier des groupes d'objets similaires, ce qui permet donc à découvrir la distribution des objets et les corrélations intéressantes dans des ensembles de données volumineux. Cependant, la plupart des algorithmes de clustering ont besoin de savoir le nombre de classes à rechercher. Il s'agit d'une méthode non supervisée et dans la plupart des cas, l'utilisateur n'aura pas de connaissances préalables sur le nombre de classes qui se trouve dans l'ensemble de données, ce qui conduit donc à une séparation de données en un certain nombre de classes *k* qui est généralement plus grand ou plus petit que le nombre réel de classes. Si le nombre de clusters de partitionnement *k* est supérieur au nombre optimal de classes (*k'*), alors une ou plusieurs bonnes classes compactes peuvent être partitionnées. Cependant, si *k* est inférieur à *k'*, au moins un cluster distinct peut être fusionné. Ainsi, trouver le bon nombre de clusters est un problème très important (Halkidi et al, 2001, Wang et Zhang, 2007). Par exemple, si on prend l'ensemble de données montré par la figure 2.2 (a). Il est évident que le nombre de clusters optimal est trois (Halkidi et al, 2001). Le partitionnement de l'ensemble de données précèdent par un algorithme de classification non supervisée (par exemple K-means) en quatre clusters est présenté par la figure 2.2(b). Dans cet exemple l'algorithme K-means a trouvé les meilleurs quatre clusters dans lesquels l'ensemble des données peut être divisé. Cependant, cette classification n'est pas optimale pour la base de données considérée. Nous définissons, ici, le terme regroupement "optimal" comme le résultat d'exécution d'un algorithme de classification automatique (c.-à-dire, un regroupement) qui ajuste le mieux que possible les partitions inhérentes de l'ensemble de données.

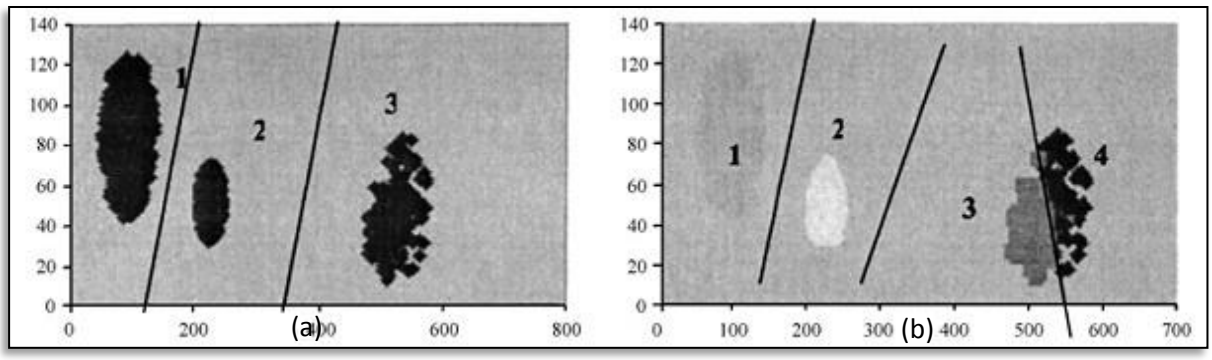


Figure 2. 2--(a) ensemble de données qui se compose de trois clusters, (b) le résultat de regroupement des données par K-means (POUR K=4).

Il est évident que le regroupement représenté par la figure 2.2 (2b) n'est pas une classification optimale de la base de données. Le regroupement optimal pour cet ensemble de données peut être obtenu par une partition des données en trois clusters. Par conséquent, si les valeurs affectées aux paramètres d'un algorithme de classification sont inexactes, la méthode de regroupement peut avoir comme résultat, un partitionnement qui n'est pas optimal pour l'ensemble de données considéré, ce qui conduit à des fausses décisions. Le problème de décider le nombre optimal de clusters pour un ensemble de données ainsi que l'évaluation des résultats d'un processus de regroupement ont été sujet de plusieurs efforts de recherches (Dave, 1996; Dimitriadou et al, 2002; Bradley et al, 1998; Chen et al, 2002).

2.8.2 Concepts fondamentaux de la validité des clusters

Dans cette section, nous présentons quelques concepts de base de la validation de clustering, ainsi que les indices de validation les plus utilisés. Comme nous avons mentionné précédemment, l'évaluation des résultats de clustering est l'une des questions les plus importantes dans le processus de la classification non supervisée. Ce sont généralement les indices de validité qui sont utilisés pour mesurer la qualité des résultats de clustering. Deux types d'indices peuvent être utilisés pour l'évaluation de la qualité d'une classification: les indices externes et les indices internes (Halkidi et al, 2001). Les indices externes mesurent l'accord entre deux partitions où la première partition est connue à priori, et la deuxième partition est le résultat d'un processus de regroupement (Dudoit et al, 2002). Les indices internes sont utilisés pour mesurer la qualité d'une structure de regroupement sans aucune information externe (Halkidi et al, 2001). C'est-à-dire, les indices externes permettent d'évaluer les résultats d'un algorithme de clustering en se basant sur une structure de regroupement connu au préalable d'un ensemble de données (ou des étiquettes de cluster).

Tandis que, les indices internes permettent d'évaluer une partition en utilisant des quantités et des caractéristiques inhérentes de l'ensemble de données considéré. Le nombre optimal de clusters, pour une classification quelconque, est généralement déterminé par un ou plusieurs indices de validité interne.

Le but d'un tel regroupement est de réaliser une classification telle que les objets à l'intérieur d'une même classe sont les plus similaires que possible et les objets de différentes classes sont les plus dissimilaires que possible, en d'autre terme, obtenir des classes Compactes et Bien Séparées ou classes CBS(en anglais, CWS clusters : Compact Well Separated Clusters). (Halkidi et al, 2001 ; Liu et al, 2010) :

I. La compacité. Plusieurs mesures basées sur la variance permettent d'évaluer la compacité d'un cluster. Basse mesure de variance indique une meilleure compacité. En outre, il existe de nombreuses mesures basées sur la distance pour estimer la compacité d'un cluster, tels que la distance maximale ou moyenne par paires.

II. La Séparabilité. Mesure à quel degré un cluster est distinct ou bien séparés par rapport aux autres clusters. Par exemple, les distances par paire entre les centres de classes ou les distances minimales par paire entre les objets dans différents groupes sont largement utilisées comme des mesures de séparation. En outre, les mesures basées sur la densité sont utilisées dans certains indices.

La procédure générale pour déterminer une meilleure partition d'un ensemble d'objets à l'aide des mesures de validation interne est la suivante (Liu et al, 2010) :

Étape 1: initialiser la liste des algorithmes de classification qui sera appliqué à l'ensemble de données.

Étape 2: Pour chaque algorithme de clustering, utiliser différentes combinaisons de paramètres pour obtenir des résultats de clustering différents.

Étape 3: calculer l'indice de validation interne correspondant à chaque partition obtenue à l'étape 2.

Étape 4: Choisir la meilleure partition et le nombre de cluster optimal selon les indices de validités utilisés.

On peut choisir un indice de validité pour estimer le nombre optimal de clusters, où la classification optimale peut être extraite parmi plusieurs classifications sous différents nombres de clusters. Toutefois, la meilleure solution de clustering pour une tâche de classification ne dépend pas uniquement d'un indice de validité, mais aussi bien de la procédure de regroupement appropriée. Un cas évident est l'utilisation de différentes

méthodes de classification et différents indices de validité des résultats dans différentes solutions de clustering pour une tâche spécifique de classification. Par conséquent, il reste encore beaucoup de travail complexe à faire dans le processus de validation de cluster (Kaijun et al, 2009). Les principes de certains indices largement utilisé pour l'estimation du nombre optimal de cluster et l'évaluation de la qualité de clustering sont introduits dans ce qui suit.

2.8.2.1 Erreur quadratique moyenne

L'erreur quadratique moyenne -*Mean Squared Error, MSE*- est une mesure de compacité très utilisée, elle est notamment équivalente à la fonction de coût de l'algorithme de K-means décrite précédemment (Guérif, 2006) :

$$MSE = \frac{1}{N} \times \sum_{i=1}^{N} \sum_{j=1}^{K} c_{ij} \times \left\| x_i - w_j \right\|^2 \qquad \textbf{(2.1)}$$

Ou k est le nombre de groupes et $c_{ij}=1/c_j(i)$ indique si $x_i \in C_j$.

2.8.2.2 Indice de Davies-Bouldin

L'indice de Davies-Bouldin tient compte à la fois de la compacité et de la séparabilité des groupes, la valeur de cet indice est d'autant plus faible que les groupes sont compacts et bien séparés (Wang et al, 2007). L'indice de Davies-Bouldin peut être utilisé pour estimer le nombre des classes telles que le nombre optimal de clusters est donné par le point de minimum global de cet indice. Cet indice favorise les groupes hypersphériques et il est donc particulièrement bien adapté pour une utilisation avec la méthode des K-means (Guérif, 2006).

$$I_{DB} = \frac{1}{K} \sum_{K=1}^{K} \max_{j \neq K} \left\{ \frac{S_c(C_K) + S_c(C_j)}{D_{ce}(C_K, C_j)} \right\} \qquad \textbf{(2.2)}$$

ou $S_c(C_k)$ est la distance moyenne entre un objet du groupe C_i et son centre, et $D_{ce}(C_k,C_j)$ est la distance qui sépare les centres des groupes C_k et C_j:

$$S_c(C_i) = \frac{1}{N_i} \sum_{i=1}^{N_i} \left\| x - w_i \right\| \qquad \textbf{(2.3)}$$

$$D_{ce}(C_i, C_j) = \left\| w_i - w_j \right\| \qquad \textbf{(2.4)}$$

2.8.2.3 Indice de silhouette

Kaufman et Rousseeuw suggèrent choisir le nombre de groupes k>2 qui donne la plus grande valeur de silhouette, la silhouette d'un cluster A est mesurée selon sa compacité et à quelle

distance ce cluster est loin de son prochain cluster le plus proche. Prenons *i* un objet arbitraire dans *A*. On définit *a(i)* comme la distance moyenne entre l'objet *i* et tous les autres objets dans le même cluster (Famili et al, 2003; Kaufman et Rousseeuw, 1990)

$$a(i) = \frac{\sum_{j \in A, j \neq i} d(i, j)}{|A| - 1} \quad \textbf{(2.5)}$$

pour tout autre cluster $C \neq A$, on définit

$$d(i, C) = \frac{1}{|C|} \sum_{J \in C} d(i, j) \quad \textbf{(2.6)}$$

et

$$b(i) = \min_{C \neq A} \{ d(i, C) \ \} \quad \textbf{(2.7)}$$

alors la silhouette objet de l'objet i est donnée par :

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad \textbf{(2.8)}$$

Ce qui entraîne que la valeur de *S(i)* soit entre -1 et 1.

La silhouette du cluster est la moyenne de la silhouette objet pour tous les objets du cluster *A* elle est donnée par :

$$cluster\_silhouette = \frac{\sum_{i=1}^{|A|} s(i)}{|A|} \quad \textbf{(2.9)}$$

La silhouette générale d'un résultat regroupement avec *c* cluster est donnée par :

$$general\_silhouette = \frac{1}{c} \sum_{j=1}^{c} cluster\_silhouette_i \quad \textbf{(2.10)}$$

Le nombre optimal de clusters est obtenu pour la valeur de la silhouette générale la plus grande. Plus la valeur de silhouette est grande, plus la qualité du cluster est meilleure.

2.8.2.4 Homogénéité et séparation

L'homogénéité et Séparation sont deux index proposés par Shamir et Sharan (2002). L'homogénéité est calculée en tant que la distance moyenne entre chaque objet et le centre du cluster dont il appartient :

$$H_{ave} = \frac{1}{N} \sum_{i} d(i, C(i)), \quad \textbf{(2.11)}$$

où *i* est un objet et *C(i)* est le centre du cluster qui contient l'objet *i*, *N* est le nombre total d'objets; *d* est la fonction de distance. La séparation est calculée comme la distance moyenne des poids entre les centres des groupes (Chen et al, 2002).

$$S_{ave} = \frac{1}{\sum_{i \neq j} N_{Ci} N_{Cj}} \sum_{i \neq j} N_{Ci} N_{Cj} D(C_i, C_j) \qquad \textbf{(2.12)}$$

Ou *Ci* et C_j sont les centres du i^{eme} et j^{eme} cluster, et N_i et N_j sont le nombre d'objets dans le i^{eme} et j^{eme} cluster. Ainsi H_{ave} reflète la compacité des clusters tandis que S_{ave} reflète la distance globale entre les clusters. Décroître H_{ave} où accroître S_{ave} permet d'améliorer les résultats de regroupement. Il est à noter également que *H* et *S* ne sont pas indépendant l'un de l'autre, *H* est étroitement lié à la variance dans le cluster, *S* est étroitement liés à la variance entre les clusters. Pour un ensemble de données quelconque, la somme de la variance intra-clusters et la variance inter-clusters est une constante.

2.8.2.5 Indice inter-intra poids

Cette technique procède par une recherche en avant et s'arrête à la première marque vers le bas de l'indice, qui indique le nombre optimal de groupes (Strehl, 2002). Cet indice vise à maximiser similarité intra-cluster et minimiser similarité inter-cluster.

2.9 Conclusion

Ce chapitre avait pour but de présenter une vue générale du problème du clustering en introduisant les notions de base en classification non supervisée, mais aussi de mettre en évidence la diversité qu'il existe parmi les différentes méthodes. Au cours de ce chapitre nous avons également introduit les notions possibles de la similarité entre objets ou le choix de la fonction de distance utilisée pour évaluer la similarité entre objets doit être effectué au préalable en fonction du problème considéré. Dans ce chapitre, nous avons aussi rappelé Les étapes de base pour développer un processus de classification automatique. Ainsi nous avons constaté qu'il existe différentes applications au clustering, chacune ayant un objectif différent. La première étape de la mise en œuvre ou de l'utilisation d'une méthode de clustering doit par conséquent être d'identifier l'application visée afin d'identifier ses besoins. Il existe aussi de nombreuses méthodes de clustering avec chacune ses atouts et ses limites. Le chapitre suivant est consacré à la présentation plus détaillée des méthodes existantes les plus classiques en clustering et plus précisément les cartes auto-organisatrices de kohonen.

Chapitre 3

Les méthodes de classification automatique

3.1 Introduction

La classification non supervisée est un domaine très actif qui a engendré un nombre très important de travaux de recherches. De nombreuses méthodes et approches ont été définies, et il serait difficile de présenter une liste exhaustive ici. Cependant nous pouvons distinguer différentes méthodes couramment utilisées. Dans ce chapitre, nous avons choisi de présenter un tour d'horizon des méthodes de clustering les plus classiques, en mettant l'accent, plus précisément, sur les cartes au-organisatrices de kohonen.

3.2 L'approche neuromémitique

3.2.1 Quantification vectorielle

La tâche de l'identification d'un sous-ensemble approprié qui décrit et représente un ensemble plus grand de vecteurs de données est appelée quantification vectorielle (QV) (Pölzlbauer, 2004). En d'autres termes, la quantification vectorielle vise à réduire le nombre de vecteurs ou à les remplacer par des centres de gravité. La figure 3.1 montre le principe des méthodes de QV, en réduisant l'ensemble original de 5 échantillons à 2 échantillons. Les centres de gravité qui en résultent ne doivent pas nécessairement être continus dans l'ensemble des échantillons, mais peuvent aussi être un rapprochement des vecteurs qui leur sont assignées, par exemple leur moyenne. La QV est étroitement liée au regroupement, dont K-means est l'une des techniques de quantification vectorielle les plus importantes.

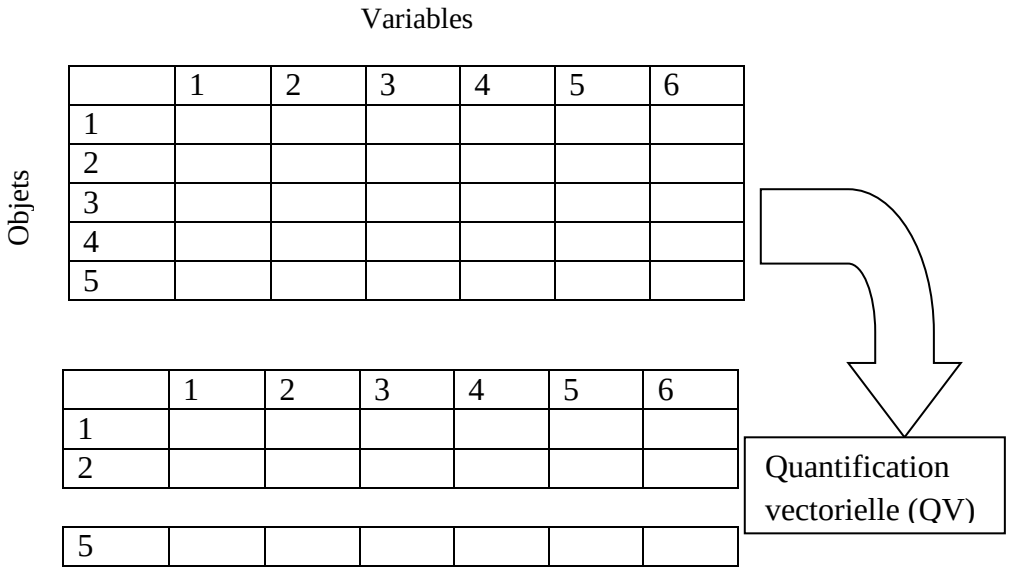


Figure 3. 1--Représentation schématique de la quantification vectorielle

Evidemment, le SOM est un algorithme de QV puisque les vecteurs de données d'entrées sont mappés à un certain nombre de vecteurs prototypes (plus petit). En raison de ses autres fonctionnalités telles que la projection vectorielle, le SOM présente deux principaux inconvénients du point de vue QV en tant qu'effets secondaires de la fonction de voisinage:

- *l'effet de frontière*: les unités sur les bords et dans les coins n'ont pas le même nombre de voisins, mais ces unités frontalières sont choisies plus souvent comme BMU (Pölzlbauer, 2004).

- *problème des unités d'interpolation*: si les vecteurs d'entrées sont largement séparés, les neurones situés entre les régions qui sont très différents sur la carte de kohonen sont mis à jour par la fonction de voisinage.

Pour mesurer la qualité d'un algorithme de QV (et bien sûr de la SOM), l'erreur de quantification vectorielle peut être calculée.

3.2.2 Projection vectorielle

La visualisation est une tâche très importante dans le processus de data mining. Ainsi, la représentation graphique d'un ensemble de données peut fournir des indications sur la structure et la distribution sous-jacente des données que la table numérique de données ne peut pas fournir. Cependant, les données ne peuvent pas être visualisées sur une feuille de papier ou sur un écran si leurs dimensions sont supérieures à 2. La projection vectorielle (PV) vise à réduire la dimension de l'espace d'entrée, et mapper les vecteurs d'entrée à cet espace de dimension réduite. L'espace de dimension réduite est habituellement de deux dimensions, pour permettre une visualisation des données sur un écran ou pour l'impression sur papier. Cependant, la projection vectorielle conduit inévitablement à une perte d'informations dans presque tous les cas. La cartographie de la PV doit se produire d'une manière à ce que les distances dans l'espace d'entrée soient conservées aussi bien que possible. Ainsi, les vecteurs similaires dans l'espace d'entrée sont mappés sur la même position ou sur des positions proches dans l'espace de sortie, et les vecteurs qui sont distants dans l'espace d'entrée sont mappés sur des positions différentes dans l'espace de sortie. Les algorithmes de PV tentent de préserver les distances des vecteurs qui sont proches les uns des autres, sans nécessairement préserver des distances relativement grandes. En outre, les méthodes PV peuvent être classées en deux catégories linéaire ou non linéaire (Pölzlbauer, 2004).

Les projections linéaires sont généralement des projections géométriques à un plan (2 dimensions), tandis que, les projections non linéaires essayent de découvrir les structures complexes dans l'ensemble de données. Habituellement, les algorithmes de projections non linéaires sont moins sensibles aux valeurs aberrantes, mais ils sont plus difficiles à calculer et à évaluer, car la majorité des algorithmes de cette catégorie incluent des techniques d'optimisation non déterministes. La figure 3.2 montre le principe de la PV, où l'ensemble de données de dimension 6 est réduit à 3.

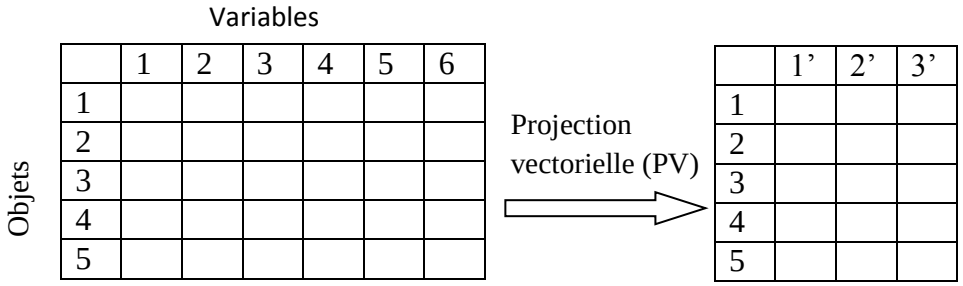


Figure 3. 2--Représentation schématique de la projection vectorielle

3.2.3 Source historique et principes

Les cartes auto-organisatrices communément désignées par SOM (pour Self Organising Maps) ont été introduites par T.Kohonen en 1981 en s'inspirant du fonctionnement des cartes topographiques du cerveau humain, tel que, les points proches qui se trouvent dans le corps humain sont représentés par des groupes de neurones proches dans le cerveau. Ces cartes ne sont pas uniformes, à savoir, la surface la plus sensible du corps humain est représentée par un cluster contenant le plus grand nombre de neurones. D'un point de vue informatique, on peut traduire cette propriété de la façon suivante : supposons que l'on dispose d'un ensemble de données que l'on désire classifier. On cherche un mode de représentation tel que les objets voisins soient classés dans la même classe ou dans des classes voisines. Ce type de réseaux de neurones artificiels a largement montré son efficacité dans la classification de données multidimensionnelles, mais malheureusement il a été ignoré pour de nombreuses années malgré son grand intérêt. Le principe des cartes de Kohonen est de projeter un ensemble complexe de données sur un espace de dimension réduite (2 ou3). Cette projection permet d'extraire un ensemble de vecteurs dites référents ou prototypes. Ces prototypes sont caractérisés par des relations géométriques simples. La projection de données par SOM se produit tout en conservant la topologie et les métriques les plus importantes des données d'entrée lors de l'affichage, c'est-à-dire les données proches (dans l'espace d'entrée) vont

avoir des représentations proches dans l'espace de sortie et vont donc être classés dans le même cluster ou dans des clusters voisins (Dreyfus et al, 2004 ; Kohonen, 1990; El Golli, 2004; Boinee, 2006; kolehmainen, 2004).

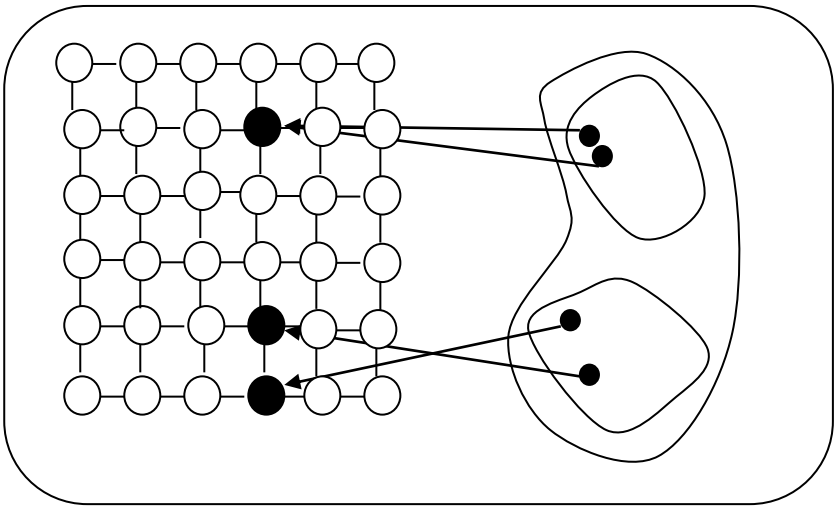


Figure 3. 3--Des neurones voisins sur la carte représentent des objets assez "proche" dans l'espace des données d'entrées

3.2.4 Architecture des cartes de Kohonen

Le SOM est une grille (ou réseau) de dimension faible qui contient un nombre *M* de neurones. Dans cette thèse, uniquement les grilles bidimensionnelles seront considérées, puisque les grilles de dimensions supérieures à 3 sont difficiles à être visualisées (Vesanto, 1999). Si la visualisation n'est pas nécessaire, les grilles dont la dimension est supérieure à trois peuvent être utilisées (Himberg, 2000). Les neurones sont généralement disposés soit dans un hexagone ou dans un rectangle (figures 3.5 a, b), d'autres topologies sont possibles, mais ne seront pas discutés dans cette thèse. La carte de kohonen est habituellement composée de deux couches de neurones, une couche d'entrée et une couche de sortie. Dans la couche d'entrée, chaque objet à classer (dans notre cas, les paramètres météorologiques sur 24h) est représenté par un vecteur multidimensionnel (section 1.4). La couche (topologique) d'adaptation ou la couche de sortie est composée d'un treillis de neurones selon la géométrie prédéfinie (Reljin et al, 2003; Turias et al, 2006). Chaque neurone de la couche topologique est totalement connecté aux neurones de la couche d'entrée $w_{.i} = (w_{1i}, \ldots w_{ni})$. Les vecteurs poids de ces connexions forment le vecteur référent ou prototype associé à chaque neurone, qui est de même dimension que les vecteurs d'entrées. La dimension des vecteurs d'entrées (appelée "dimension d'entrée") est généralement beaucoup plus grande que celle de la grille (appelée «dimension de sortie"). Par conséquent, le SOM est dit un algorithme de "Projection

Vectorielle", car il réduit la dimension de l'espace d'entrée (plus de 2 dimensions) à la dimension de la grille (dimension 2)).

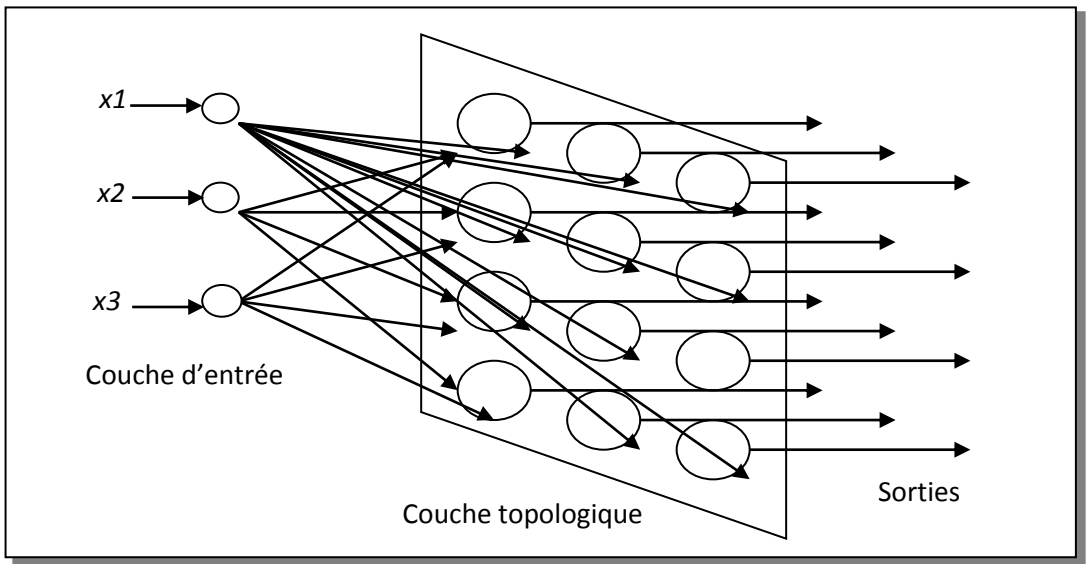


Figure 3. 4--Structure d'une carte auto-organisatrice.

C'est l'utilisation de la notion de voisinage qui introduit les contraintes topologiques dans la géométrie finale des cartes de kohonen. Le réseau hexagonal est à la base de la plupart des applications des cartes de kohonen, les grilles rectangulaires sont aussi utilisées, mais leurs topologie diffère des réseaux traditionnels (figure 3.5). La position des neurones dans le réseau, en particulier les distances entre eux et les relations de voisinage sont très importantes pour l'algorithme d'apprentissage. Ainsi, l'architecture d'un réseau de kohonen bidimensionnel de grille rectangulaire est montrée par la figure 3.4, cette architecture est composée d'une couche d'entrée de dimension *M*=3 et une couche topologique de dimension L=4*3=12 neurones. Un vecteur d'entrée $x(t) = [x_1,..x_M]^T$ est projeté à la couche de sortie. Chaque entrée de la SOM est connectée à tous les neurones par des poids correspondants (*wji*) ou $j = 1,..L$ et $i = 1,..,M$. Ainsi à chaque neurone de la SOM un vecteur poids de dimension *M* est affecté $w_j = [w_{j1},..,w_{jM}]^T$.

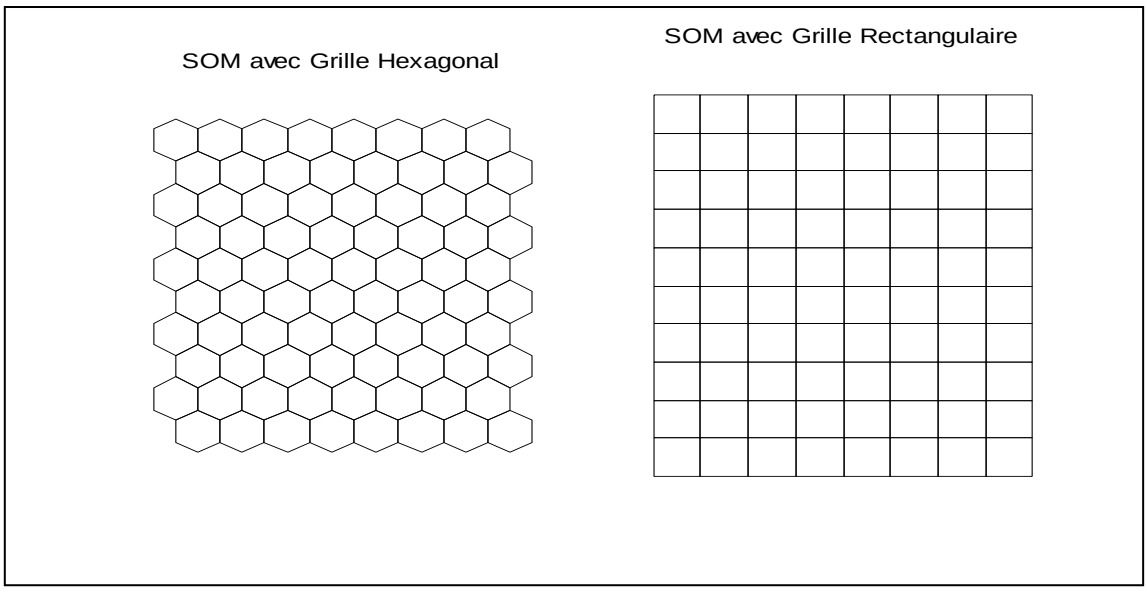


Figure 3. 5--Les formes topologiques les plus utilisées des cartes de kohonen : (a) rectangulaire (b) hexagonal.

3.2.5 Apprentissage du SOM

Une fois les vecteurs référents ou prototypes sont initialisés, l'apprentissage commence. Le SOM est très robuste en ce qui concerne les paramètres d'initialisations, mais une bonne initialisation permet à l'algorithme d'apprentissage de converger plus rapidement à une bonne solution. L'ensemble d'individus d'apprentissage est présenté à l'algorithme SOM, ce processus est répété pour *t* étapes d'apprentissage. Un tour complet de l'apprentissage (lorsque tous les échantillons ont été présentées) est appelée une «époque». Le nombre d'itérations *t* de l'apprentissage est un nombre entier multiple du nombre d'époques. Chaque itération est composée de deux étapes : une étape de compétition entre les neurones qui détermine la région de la grille à ajuster, et une étape d'adaptation des poids de la zone sélectionnée à l'individu projeté. Le principe des deux phases d'apprentissage est illustré sur la figure 3.7. L'algorithme SOM est une variante de l'algorithme des K-means, ainsi, lors d'une itération *t*, l'algorithme d'apprentissage du SOM ne modifie pas seulement le centre sélectionné comme étant le plus proche d'un individu en entrée, mais aussi les centres voisins pour un graphe de voisinage fixé. Pendant la phase d'apprentissage, le processus d'auto-organisation permet de concentrer l'adaptation des poids des connexions essentiellement sur la région de la carte la plus «active». Cette région d'activité est choisie comme étant le voisinage associé au neurone dont l'état est le plus actif on parle ainsi de neurone gagnant. Le critère de

sélection du neurone gagnant est de chercher celui dont le vecteur poids est le plus proche en termes de distance euclidienne de l'individu présenté (Pölzlbauer, 2004; Dreyfus et al, 2004).

Algorithme 3.1 : Pseudo code de l'algorithme d'apprentissage du SOM. L'apprentissage se termine quand le nombre *N* des époques est atteint (Dreyfus et al, 2004).

1: pour toutes les époques faire
2: pour toutes les entrées faire
3: calculer les distances entre le vecteur d'entrée et le poids de tous les neurones dans la carte;
4: [compétition] sélection du neurone vainqueur;
5: [Coopération] Adapter les poids du neurone vainqueur et ses neurones voisins;
6: Adapter les paramètres d'apprentissage;
7: fin pour
8: fin pour

3.2.5.1 Phase de compétition

La phase de compétition entre les neurones est basée sur une fonction discriminante où les neurones calculent leurs valeurs pour chaque vecteur d'entrée. Le vainqueur de la compétition est le neurone qui a la plus grande valeur retournée par la fonction discriminante. Si on prend un espace de dimension *n* pour l'ensemble de données d'entrées. Un vecteur d'entrée *x* sélectionné de façon aléatoire à partir des données d'entrées tel que $\underline{x}=[x_1, x_2, ...,x_n]^T$ où T dénote la matrice transposée. Le vecteur de poids synaptiques de chaque neurone a la même dimension que les vecteurs d'entrées sera représenté par $\underline{w}_j= [w_{j1}, w_{j2},...,w_{jn}]$, $j=1,2,...,m$, où le nombre total de neurones dans le réseau est représenté par *m,*. A un instant *t,* un vecteur *x(t)* tiré de la distribution de l'espace d'entrée est sélectionné, tous les neurones de la grille sont alors mis en compétition. Cette compétition revient à chercher le neurone vainqueur, c'est-à-dire, le plus proche du vecteur d'entrée. En d'autre termes, parmi tous les neurones de la carte, le neurone vainqueur noté *c,* est celui dont la distance entre son vecteur poids synaptiques et le vecteur d'entrée est le plus faible. Ce neurone dit "neurone gagnant" et souvent noté par BMU (Best Matching Unit). Formellement, le BMU est défini comme le neurone qui vérifie l'équation suivante:

$$\left|x(t)-w_j(t)\right| = \min_{j\in m}\left\|x(t)-w_j(t)\right\| \qquad \textbf{(3.1)}$$

Où $\|.\|$ est une mesure de distance.

Le neurone vainqueur, pour un vecteur d'entrée est également appelé centre d'excitation de la carte. La distance généralement utilisée entre les vecteurs *x* et *w* est la distance euclidienne, mais tout autre type de distance peut être utilisé.

3.2.5.2 Phase d'adaptation

Pour que les vecteurs similaires en entrées soient mappés sur le même neurone ou sur des neurones proches de la carte, non seulement le neurone vainqueur, mais aussi ses neurones voisins doivent être mis à jours. Cette action vise à retirer les poids des neurones voisins à se rapprocher du vecteur d'entrée, fournissant ainsi des vecteurs poids plus similaires et par conséquent, des entrées similaires sont mappées sur des neurones proches de la carte (figure 3.7). Les vecteurs poids synaptiques w_j du neurone d'indice *j* et ses voisins de la carte auto-organisatrice sont mis à jour par correction d'erreur (l'erreur est définie comme la distance entre le vecteur *x* et le vecteur de référence w_j du neurone considéré) :

$$w_j(t+1) = w_j(t) + \Delta w_j(t) \quad \textbf{(3.2)}$$

$$= w_j(t) + a(t)h_{cj}(r(t))[x(t) - w_j(t)] \quad \textbf{(3.3)}$$

Avec :

$$\Delta w_j(t) = a(t)h_{cj}(r(t))[x(t) - w_j(t)]. \quad \textbf{(3.4)}$$

Dans les expressions (3.3) et (3.4), *a(t)* représente le coefficient d'apprentissage qui diminue dans le temps pour permettre un meilleur ajustement des poids. $h_{cj}(r(t))$ est le noyau de voisinage autour du neurone gagnant *c*, avec un rayon de voisinage *r (t).* L'adaptation des poids de chaque neurone est réalisée en fonction de la position du neurone dans la grille par rapport au neurone gagnant.

Au cours de l'apprentissage, la taille de voisinage du BMU, qui détermine la zone active, décroît avec le temps. L'évolution temporelle du coefficient d'apprentissage est illustrée sur la figure 3.6. Ainsi, les paramètres d'apprentissage sont modifiés progressivement en commençant par une phase initiale plus grossière avec une grande zone d'influence et évolution rapide des vecteurs prototypes jusqu'à arriver à une phase de mise à jours fin avec un petit rayon de voisinage et vecteurs prototypes qui s'adaptent lentement aux échantillons.

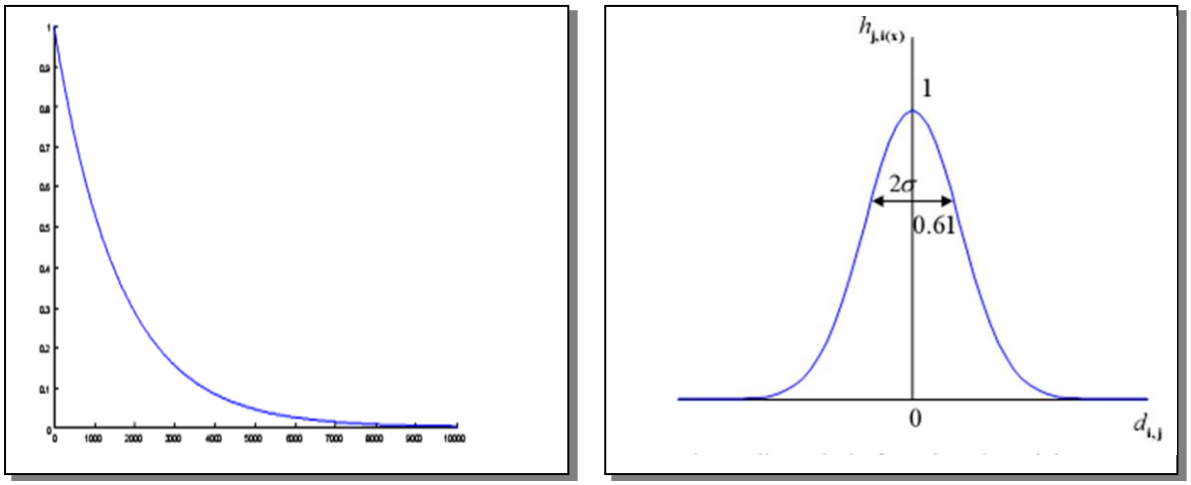


Figure 3. 6--Évolution des paramètres d'une carte de kohonen au cours de l'apprentissage. (a) l'évolution du coefficient d'apprentissage au cours de l'apprentissage. (b) L'allure de la fonction de voisinage pour un rayon donné (s=0.61).

La fonction de voisinage généralement utilisée est la fonction gaussienne, cette fonction est centrée sur le neurone déclaré vainqueur après la phase de compétition qui a suivi la présentation d'un vecteur d'entrée. Le rôle de la modification appliquée sur le voisinage choisi revient à rapprocher les vecteurs poids sélectionnés de l'exemple présenté. Ainsi le neurone dont le vecteur poids est proche du vecteur d'entrée est mis à jour pour qu'il soit plus proche. Le résultat est que le neurone gagnant est plus probable de gagner la compétition une autre fois si un vecteur d'entrée similaire est représenté, et moins probable si le vecteur d'entrée est totalement différent du vecteur précédent. Comme précisée précédemment, la fonction de voisinage tient compte de la distance des neurones par rapport à la position du neurone vainqueur pour pondérer la correction des poids synaptiques delta du neurone *i* à l'instant *t*. Soit δ_{ci} la distance entre le neurone vainqueur d'indice *c* et un neurone voisin d'indice *i*. Cette distance n'est pas calculée dans l'espace des entrées mais dans l'espace topologique de la carte :

$$\delta_{ci}^2 = \|c - i\|^2 \qquad \textbf{(3.5)}$$

La fonction de voisinage $h_{ci}(t)$ s'écrit alors :

$$h_{ci}(t) = e^{-\frac{\delta^2_{ci}}{2r(t)^2}} \qquad \textbf{(3.6)}$$

Où *r(t)* est le rayon de voisinage. Ce rayon peut être exprimé par l'expression suivante :

$$\delta_k = \delta_i \left(\frac{\delta_f}{\delta_i} \right)^{\frac{k}{k_{\max}}} \quad \textbf{(3.7)}$$

Le processus d'apprentissage est interrompu si l'une des conditions est rencontrée : le nombre maximum d'époques est atteint, la performance est minimisée à un but, ou un temps maximum d'apprentissage est excédé.

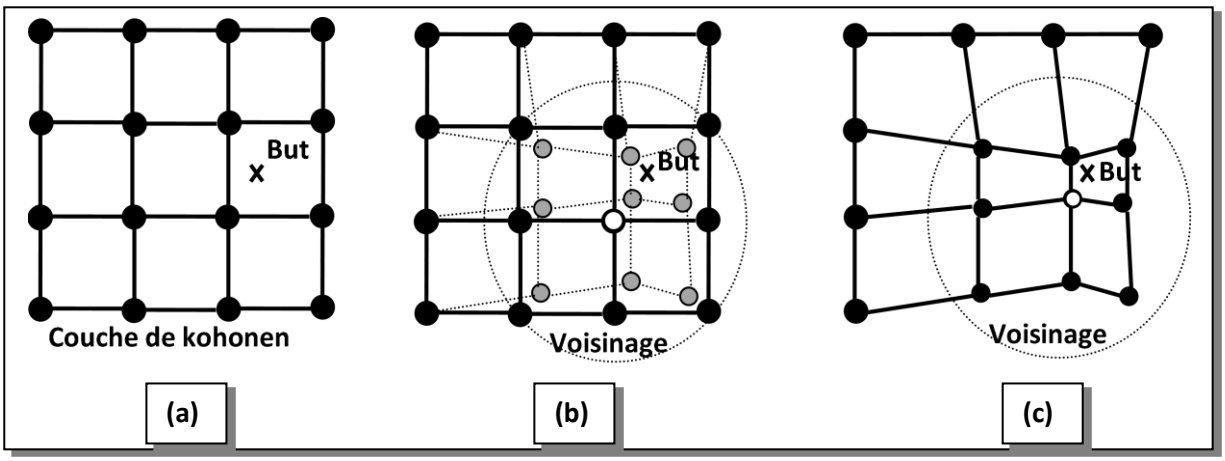


Figure 3. 7--Illustration de l'apprentissage de la méthode SOM : (a) État initial, (b) état à l'étape k, (c) état à l'étape k+1.

Les cartes de Kohonen ont prouvé leur utilité pour la classification des bases de données multidimensionnelles traitant des problèmes non linéaire. L'algorithme SOM est capable d'extraire les propriétés statistiques des paramètres météorologiques présentes dans la base de données d'entrée et c'est la raison pour laquelle ce type de réseau a été choisi pour la présente étude. Afin d'obtenir de bons résultats, un apprentissage du réseau par des données statistiquement représentative de la totalité des données doit être effectué. Dans notre cas, les propriétés statistiques des données météorologiques ne sont pas claires, donc la base de données entière est nécessaire pour une bonne modélisation.

L'algorithme décrit ci-dessus est appelée "apprentissage séquentiel» ou «SOM basique». Une autre règle importante d'apprentissage est appelé «apprentissage par lot", qui est basée sur l'itération du point fixe, est nettement plus rapide en termes du temps de calcul. A chaque étape, les BMU pour tous les échantillons d'entrée sont calculés à la fois, et les vecteurs prototypes sont mis à jour comme suit:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic(j)}(t) x_j}{\sum_{j=1}^{n} h_{ic(j)}(t)} \quad \textbf{(3.8)}$$

3.2.6 Initialisation et paramétrage de la carte auto-organisatrice

En dehors de l'algorithme d'apprentissage, il y a des choix à faire qui peuvent être considérées comme paramétrages du SOM, à savoir le choix des fonctions *a(t)*, et $h_{ck}(t)$, la topologie du réseau, et le nombre de vecteurs prototypes (et leur état initial). L'initialisation des vecteurs prototypes est habituellement réalisée par l'une des méthodes suivantes:

- *Initialisation aléatoire*. Les vecteurs prototypes sont initialisés aléatoirement, ce qui n'est souvent pas la bonne politique à tenir, mais cette politique a montrée qu'elle converge vers une bonne carte topographique pour un grand nombre d'époques d'apprentissage.
- *Initialisation linéaire*. Les vecteurs prototypes sont initialisés en ordre ascendant ou descendant le long des axes *x* et *y* du réseau. Ce type d'initialisation dépend habituellement des composantes principales des échantillons de données (ce sujet sera discuté dans la section 3,5). C'est la méthode qui sera utilisée pour l'initialisation de la carte de Kohonen.
- *Permutation aléatoire d'un sous-ensemble des échantillons*. Similaire à la méthode d'initialisation aléatoire, des échantillons aléatoires sont prélevés comme des vecteurs modèles.

L'initialisation linéaire a également l'avantage d'être déterministe, réduisant ainsi le caractère aléatoire de l'algorithme d'apprentissage du SOM. Cela permet une reproduction plus facile des résultats.

On définit le voisinage de rayon *r* d'une unité *u*, noté *V(u)*, comme l'ensemble des unités *u* situées sur le réseau à une distance inférieure ou égale à *r* (Rousset, 1999 ; Lemaire, 2006). Le noyau de voisinage $h_{ci}(t)$ peut être n'importe quelle fonction qui diminue avec l'augmentation de la distance sur le réseau. Un exemple typique d'un noyau de voisinage est dérivé de la courbe de Gauss. La figure 3.9 montre quatre fonctions de voisinage différentes sur un réseau. Le neurone gagnant est sélectionné en tant que BMU et son influence sur ses voisins est déterminée par les fonctions de voisinage suivantes:

$$h_{ci}(t) = e^{-\frac{\delta^2_{ci}}{2r(t)^2}} \qquad \textbf{(3.9)}$$

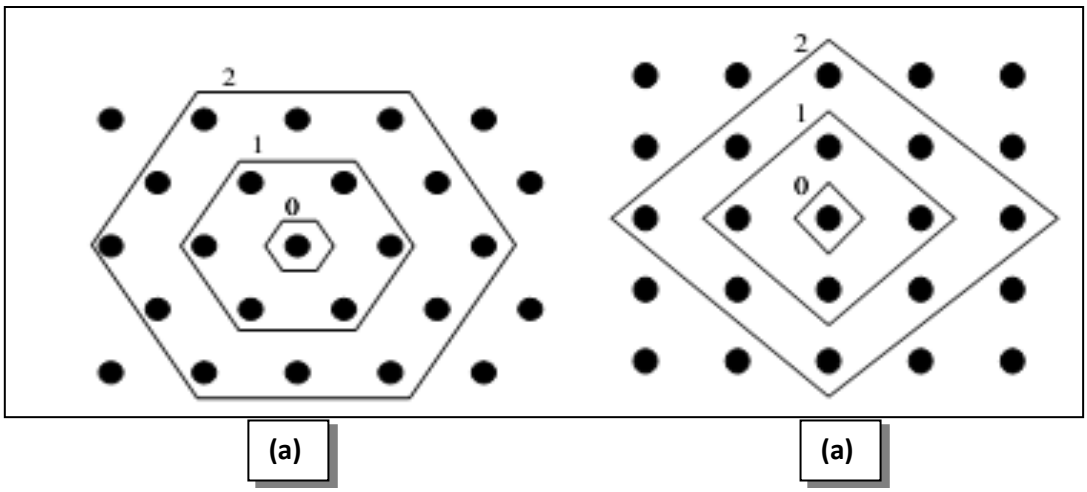


Figure 3. 8-- Voisinages distincts (de taille 0, 1 et 2) du neurone gagnant: (a) réseau hexagonal, (b) réseau rectangulaire. Le polygone le plus intérieur correspond au voisinage d'ordre 0, le second au voisinage d'ordre 1 et le plus grand correspond au voisinage

Un exemple de SOM peut être donnée en utilisant la librairie somtoolbox[2], l'utilisateur pourra librement fixer tous les paramètres précèdent, mais pour réduire ses efforts, des valeurs par défaut sont fournies pour ces paramètres, ces valeurs par défaut sont (Vesanto, 2000) :

— Le nombre de neurones de la couche topologique peut être défini approximativement par l'équation: $m = 5\sqrt{n}$ ou n est le nombre d'échantillons de données.

— La forme par défaut de la carte est une feuille rectangulaire avec un treillis hexagonal. Le rapport longueurs/largeur de la carte est déterminé selon le rapport calculé entre les deux plus grandes valeurs propres à la matrice de covariance de données.

— La fonction de voisinage par défaut est la fonction gaussienne $h_{ci}(t) = e^{-\frac{\delta^2_{ci}}{2r(t)^2}}$, ou δ_{ci} est la distance entre les nœuds *c* et *i* de la carte, et *r (t)* est le rayon de voisinage au temps *t*.

— Le rayon d'apprentissage, ainsi que le taux d'apprentissage, sont des fonctions monotoniquement décroissante dans le temps. Le rayon initial dépend de la taille de la carte, mais le rayon final est 1. Le taux d'apprentissage commence à partir de 0.5 et finit (presque) à zéro.

— La longueur d'apprentissage est mesurée en époques : Le nombre d'époques est directement proportionnel avec le rapport entre le nombre d'unités de la carte et le nombre d'échantillons de données.

[2] Disponible gratuitement sur http://www.cis.hut.fi/projects/somtoolbox.

L'algorithme SOM est très robuste en ce qui concerne le choix des paramètres d'initialisation. Ainsi, les résultats sont presque identiques pour différents choix de fonctions et paramètres discutés ci-dessus.

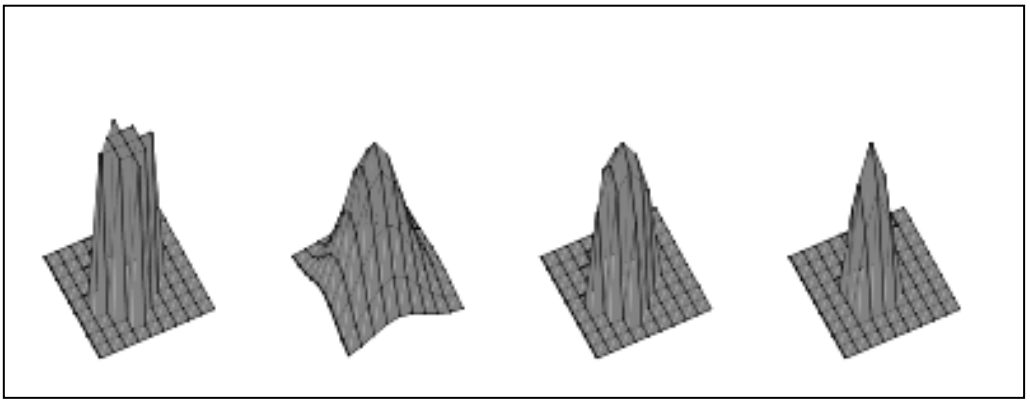


Figure 3. 9--Fonctions de voisinage : bulle, gaussienne, coupe gaussienne et epanechicov.

3.2.7 Visualisation

Une fois le processus d'apprentissage du SOM est terminé, on passe à l'étape suivante qui est le post-traitement et la visualisation des résultats. Cette étape est particulièrement importante, ainsi, les visualisations intuitives et significatives sont en fait l'un des atouts les plus importants du SOM. Bien que les réseaux de neurones artificiels sont généralement très difficiles à être visualiser, le SOM est une exception remarquable, ceci est l'une des raisons de la popularité des SOMs. Les performances de visualisation du SOM sont principalement dues à ses capacités de rapprocher un ensemble de données et de les représenter en deux dimensions (Simula et al, 1999; Kaski et al, 1997). Dans les techniques de visualisation des clusters et des variables, les vecteurs prototypes du SOM sont considérés comme un échantillon représentant les données: les données originales sont remplacées par un ensemble plus petit où l'effet du bruit et de données aberrantes est diminué. On suppose que les propriétés observées lors de la visualisation des prototypes sont également respectées pour les données originales. Pour cette raison, une certaine prudence s'impose avant de tirer des conclusions ambitieuses basées sur les visualisations du SOM. Dans ce qui suit, les méthodes de visualisation du réseau de Kohonen les plus utilisées sont présentées (Pölzlbauer, 2004).

3.2.7.1 Visualisation des clusters

Les techniques qui permettent de visualiser la forme et la structure des classes d'un nuage de points sont généralement basées sur des projections vectorielles. Comme la forme de la grille du SOM est prédéfinie, cette projection n'est pas vraiment utile tel qu'elle est. Par conséquent, la carte des vecteurs prototypes doit être projetée sur un espace de faible dimension. En plus des coordonnées physiques, des techniques de codage de couleur ont été utilisées pour la visualisation des clusters (Vesanto, 1999 ; Vesanto, 2002). Cependant, la technique la plus couramment utilisée pour visualiser les clusters sur le SOM est la matrice des distances. Dans ces techniques, les distances entre chaque unité i et les unités qui sont dans son voisinage N_i sont calculées comme suit:

$$D_i = \{\|m_i - m_j\| j \in N_i, j \neq i \qquad \textbf{(3.10)}$$

Les distances, ou la moyenne des distances, pour chaque unité cartographique sont généralement visualisées en utilisant des couleurs. Ainsi, d'autres techniques de visualisation sont possibles (Vesanto, 1999 ; Vesanto, 2002). Prenons par exemple un ensemble de données composé de trois attributs aléatoires. Comme les données (et donc les prototypes de la carte) sont tridimensionnelles, ils peuvent être directement tracées sur un graphique. Sur la figure 3.10, les objets de données sont tracés avec le caractère 'O' bleu et les vecteurs prototypes de la carte de kohonen sont tracés avec le caractère '+' noir.

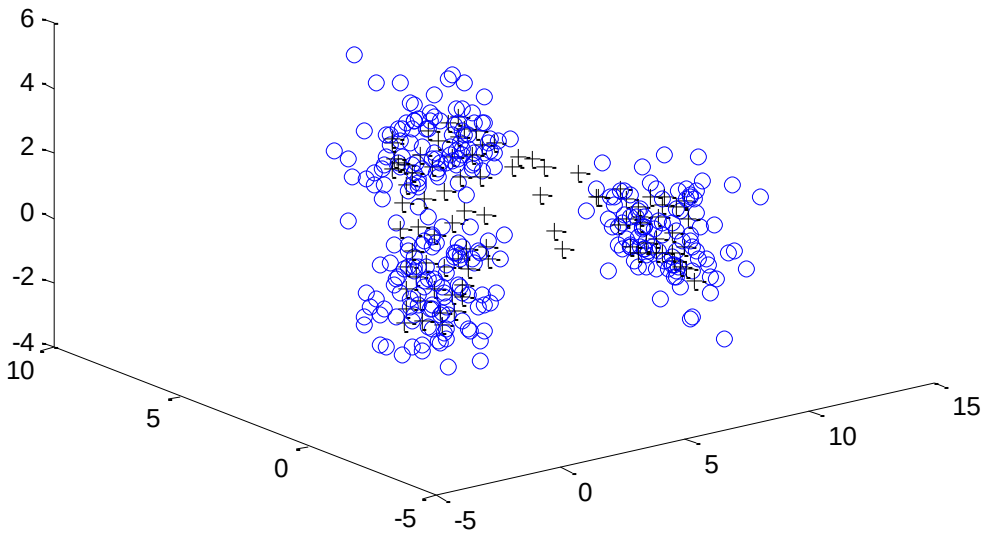


Figure 3. 10--La distribution de données sur les vecteurs prototypes de la carte de Kohonen.

Par visualisation directe, il est très facile d'identifier la distribution de données, et la façon avec laquelle les vecteurs prototypes sont positionnés. On peut facilement remarquer trois groupes de données bien séparés, ainsi que des vecteurs prototypes entre les groupes, mais en réalité aucune donnée n'est trouvée dans ces emplacements. Les unités de la carte correspondant à ces vecteurs prototypes sont appelés «mortes» ou unités d'interpolation de la carte.

La matrice de distance unifiée de Ultsch (Ultsch et Siemon, 1990), appelée U-matrix, permet de visualiser toutes les distances de chaque unité avec ses voisins. Ceci est rendu possible grâce à la structure régulière de la carte de kohonen, ainsi, il est facile de positionner un seul marqueur visuel entre chaque unité de carte de kohonen et chacun de ses unités voisines. Les valeurs élevées sur la barre de couleur signifient de grandes mesures de distance entre les unités voisines dans carte, et indiquent par conséquent les frontières entre les clusters. Un cluster sur la carte U-matrix est généralement représenté par une région uniforme, avec des valeurs de distance très faible entre les unités qui le compose (Se référer à la barre de couleur pour voir quelles couleurs signifient des valeurs élevées). La carte U-matrix montrée par la figure 3.11 pour un ensemble de données composé de trois attributs aléatoires semble y avoir trois clusters.

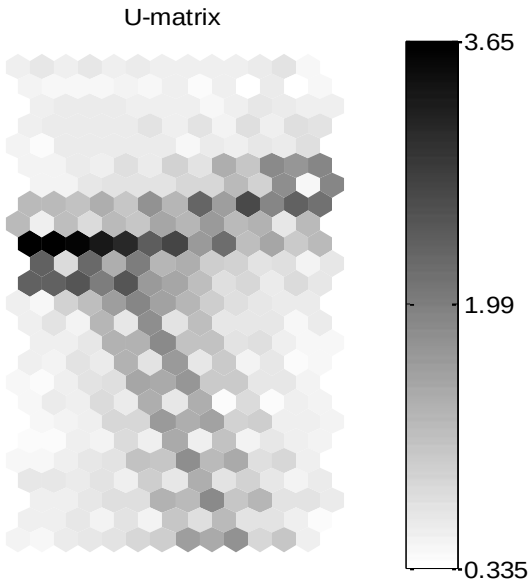


Figure 3. 11--La carte U-matrix de l'ensemble de données

La matrice de distance unifiée contient les distances entre les unités voisines de la carte, ainsi que la distance moyenne de chaque unité de la carte avec ses voisines. Ces distances

moyennes correspondant à chaque unité de la carte peuvent être facilement extraites. Le résultat est une matrice de distance moyenne appelé communément D-matrix.

Une technique similaire consiste à attribuer des couleurs aux unités cartographiques telles que les unités cartographiques similaires (proche en termes de distance) auront des couleurs similaires. Quatre sous figures représentant les méthodes de visualisation des clusters discutées précédemment sont montrées par la figure 3.12, l'U-matrix, la matrice de distance moyenne (avec niveaux de gris), la matrice de distance moyenne (avec la taille de l'unité de la carte) et la matrice code-couleur.

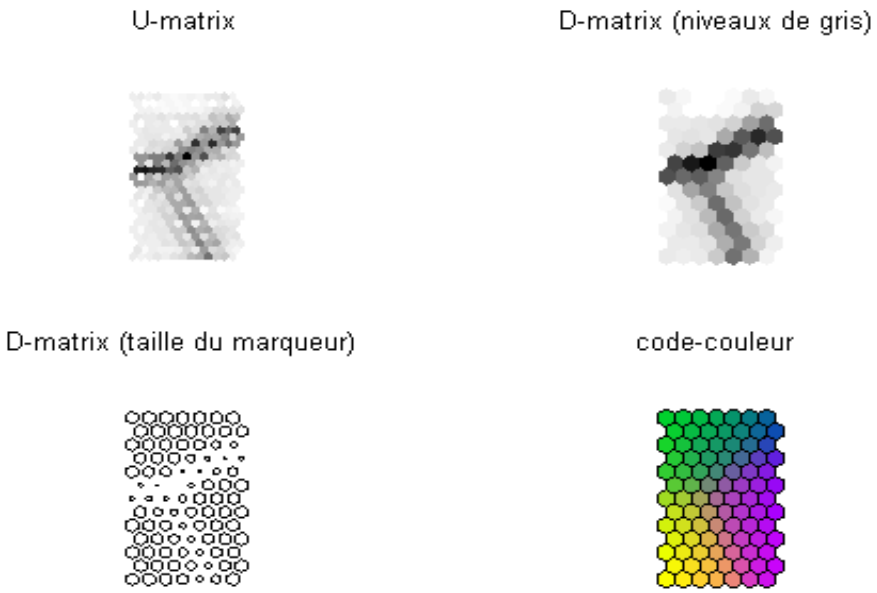


Figure 3. 12-- la carte U-matrix de l'ensemble de données

3.2.7.2 Visualisation des composants: le plan des composants

La visualisation des composants peut être considérée comme une version découpée en tranches de la carte, où chaque « plan » permet de représenter la distribution du vecteur poids d'un composant. En utilisant la distribution des composants, les différentes dépendances entre les paramètres de la base de données peuvent être étudiées. Par exemple, (Tryba et al, 1989) ont utilisé ce genre de visualisation pour étudier les variations des paramètres de conception d'un circuit VLSI. L'organisation des cartes de distributions est basée sur la mesure des dépendances entre les variables. Dans (Vesanto et Ahola, 1999), quatre méthodes ont été utilisées pour mesurer les dépendances entre les variables sont comparées. La méthode basée sur la corrélation a donné de meilleurs résultats.

Les cartes de distributions de la base de données sont montrées par la figure 2.13, ou les noms des composants sont inclus comme titres des figures secondaires. Les cartes de distributions

('X-cord', 'Y-cord', et 'Z-cord') montrée par la figure 3.13 représentent le type de valeurs des vecteurs prototypes de la carte de kohonen. Les valeurs sont indiquées avec des couleurs, et la barre de couleur à la droite montre ce que signifient ces couleurs.

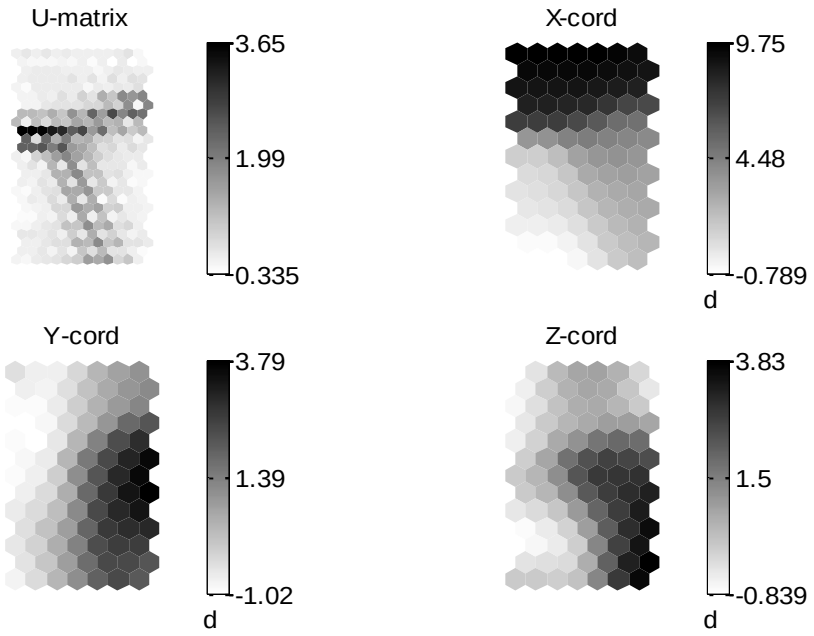


Figure 3. 13--Différentes visualisation de la SOM : U-matrix et les cartes de distribution.

Il existe également trois autres fonctions spécialement conçus pour afficher les cartes de distributions à savoir, les graphiques à barres, les graphiques circulaires et les graphiques linéaires (figure 3.14).

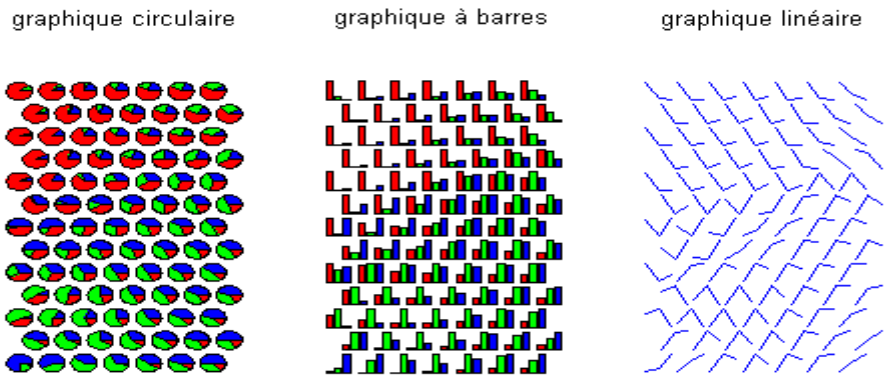


Figure 3. 14--Cartes de distribution: graphique à barres, graphique circulaire et un graphique linéaire.

3.3 Quelques approches classiques

Les techniques de clustering sont globalement divisées en deux modes, la classification hiérarchique et la classification par partition. Les méthodes hiérarchiques peuvent être aussi divisées est aussi divisé en algorithmes agglomératifs et divisifs, correspondant aux stratégies ascendantes et descendantes (Boubou, 2006 ; Berkhin, 2002; Guérif, 2006).

3.3.1 Méthodes de classification hiérarchique

Historiquement, elles furent les premières méthodes développées, principalement en raison de leur simplicité de calculs. L'avènement des ordinateurs puissants leur a fait perdre une certaine popularité au profit des méthodes non hiérarchiques. Les méthodes de regroupement hiérarchique représentent le résultat d'une classification arborescente dans laquelle les petits groupes d'objets qui sont fortement semblables les uns aux autres sont imbriqués dans de plus grands clusters qui contiennent des objets moins similaires.

Supposons que X est un ensemble de données contenant des vecteurs météorologiques à regrouper, $X = \{x_1, x_2, ..., x_N\}$. Chaque vecteur x_i est un vecteur à n dimensions, où chaque dimension correspond typiquement à un paramètre météorologique. Le regroupement des $x_i \in X$ en m groupes peut être défini en tant que $R = \{C1, C2, ..., Cm\}$, de sorte que les conditions suivantes sont satisfaites:

- Chaque cluster Ci contient au moins un vecteur météorologique: $Ci \neq \emptyset$, $i = 1, ..., m$

- L'union de tous les groupes est l'ensemble X: $\cup_{i=1}^{m} C_i = X$

- Les clusters n'ont pas des vecteurs météorologiques en commun: $C_i \cap C_j = \emptyset, i \neq j, i, j = 1,..., m$

Un regroupement $R1$ qui contient k clusters est dit imbriqué dans le regroupement $R2$, qui contient $r < k$ groupes, si chaque cluster dans $R1$ est un sous-ensemble d'un cluster de $R2$, et au moins un groupe de $R1$ est un sous-ensemble propre de $R2$.

Par exemple, le $R1$ de clustering = {{x1, x3}, {x4}. {x2, x5}} est emboîtée dans $R2$ = {{x1, x3, x4}, {x2, x5}}. D'autre part, $R1$ n'est pas emboîté dans $R3$ = {{x1, x4}, {x3}, {x2, x5}} (Tombros, 2002).

Les méthodes de classification hiérarchique sont divisées en deux grandes catégories, agglomerative et divisive selon la façon dont le dendrogramme hiérarchique est formé (Khodja, 1997 ; Chessel, 2004). La stratégie des méthodes agglomératives est de procéder à travers une série de (N-1) fusionnement des clusters. Pour une collection de N vecteurs

météorologiques, les méthodes agglomératives considèrent les singletons (les classes formées uniquement d'un seul vecteur) et en procèdent par fusionnement des classes selon une mesure de similarité pour former une nouvelle classe. Le processus est itéré jusqu'à l'obtention d'une seule classe contenant les *N* vecteurs météorologiques. D'autre part, dans une stratégie de classification divisive, un seul regroupement initial est subdivisé progressivement en petits groupes d'objets jusqu'à l'obtention des classes formées uniquement d'un seul individu. Cette méthode est très coûteuse pour être utilisée sur des volumes de données volumineuses. En effet, la division d'un groupe à *N* élément nécessite l'évaluation des $(2^{N-1} - 1)$ divisions possibles.

Le résultat d'une méthode de classification hiérarchique peut être présenté sous la forme d'un dendrogramme (Tombros, 2002) (figure 3.15). Un dendrogramme est généralement représentée comme un arbre avec des niveaux numériques associés à ses branches. Les valeurs numériques sont des niveaux de similarité pour laquelle les clusters sont formés. A tout niveau de similarité, on peut tracer une ligne perpendiculaire à l'axe de similarité. De cette façon, chaque branche de l'arbre coupé par une ligne représente un groupe constitué par les éléments du sous-arbre enraciné à cette branche. Au plus bas niveau de similarité, tous les objets sont dans un seul cluster.

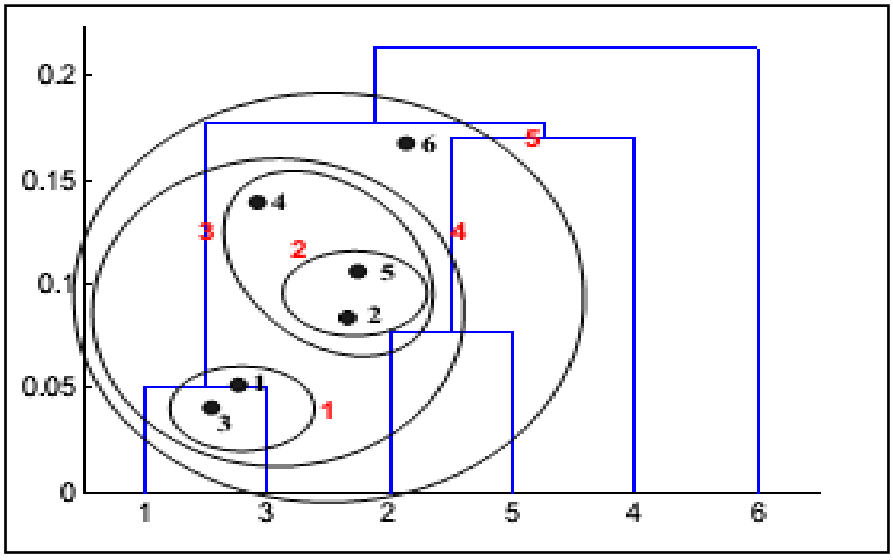


Figure 3. 15--Diagramme de cluster imbriquée avec un dendrogramme.

3.3.1.1 Les méthodes hiérarchiques agglomératives

Le principe de construction des méthodes agglomératives est d'élaborer, pas à pas, une suite de partitions emboîtées depuis la partition la plus fine formée de *N* classes jusqu'à la partition la plus grossière ($\{X\}$). Les méthodes hiérarchiques agglomératives suivent généralement la procédure générique suivante (Tombros, 2002):

1. Déterminer toutes les similarités entre les vecteurs d'entrés.

2. Former un cluster de deux objets ou clusters les plus proches

3. Redéfinir les similarités entre le nouveau cluster et tous les autres objets ou groupes déjà construit, en laissant tous les autres similarités inchangés.

4. Répéter les étapes 2 et 3 jusqu'à ce que tous les objets soient regroupés dans un seul cluster.

Les différentes méthodes agglomératives disponibles actuellement diffèrent selon la façon dont ils mettent en œuvre l'étape 3 de la procédure ci-dessus. A chaque étape *t* du processus de regroupement, la taille de la matrice de similarité *S(X)* (qui est initialement *NxN*) devient (N_t) x (N_t).

La matrice $S_t(X)$ de l'étape *t* du processus est dérivée de la matrice $S_{t-1}(X)$ par suppression des deux lignes et des colonnes qui correspondent aux objets (ou groupes) nouvellement fusionnés, et en ajoutant une nouvelle ligne et la colonne qui contiendra la nouvelle similarité entre le groupe nouvellement crée et tous les autres objets ou clusters.

Dans les paragraphes suivants, nous présentons les méthodes de classification hiérarchiques qui ont été largement utilisés dans le passé, à savoir, le lien simple, lien complet, lien moyen (représentés par la figure 3.16), et la méthode de Ward.

3.3.1.1.1 Lien simple ou *single linkage*

Dans le procédé du lien simple (saut minimum), la similarité entre deux clusters est le maximum de similarités entre tous les paires d'objets tels que le premier objet se trouve dans un groupe et l'autre objet est dans un autre groupe. Par exemple, à une étape de regroupement quelconque, le cluster *i* et le cluster *j* ont été fusionné, alors la similarité entre le nouveau cluster (appelée *p*) et un autre cluster *r* est déterminé comme suit:

$$S_{pr} = \max(S_{ir}, S_{jr}) \qquad \textbf{(3.11)}$$

Les clusters sont reliés à chaque étape par le seul lien le plus fort entre eux. Pour tout cluster réalisé par la méthode du lien simple, chaque membre est plus semblable à un autre membre du même groupe que de tout autre objet qui n'est pas dans le groupe, et par conséquent chaque

objet doit être dans le même groupe avec son document le plus proche (ou le plus proche voisin).

3.3.1.1.2 Lien complet ou *complète linkage*

La définition de la méthode dite lien complet (ou agrégation par le diamètre) est à l'opposé de la méthode du lien simple: la similarité entre deux groupes est le minimum de similarités entre toutes les paires d'objets, tels que le premier objet se trouve dans un groupe, et l'autre objet dans l'autre cluster. Par exemple, si à une étape de regroupement, le cluster *i* et le cluster *j* ont été fusionné, alors la similarité entre le nouveau cluster (appelée *p*) et un autre cluster *r* est déterminé comme suit:

$$S_{pr} = \min(S_{ir}, S_{jr}) \quad \textbf{(3.12)}$$

En raison de la façon dont ils sont formés, les groupes de liens complets ont tendance à être étroitement lié, ce qui est exactement l'opposé des clusters de lien simple. De plus, et contrairement à la méthode de lien simple, le plus proche voisin d'un objet peut être dans un groupe différent, cependant les plus proches voisins mutuels seront toujours dans le même cluster.

3.3.1.1.3 Lien moyen ou *average linkage*

La similarité entre deux groupes dans le procédé de regroupement de lien moyen est la moyenne des similarités entre tous les paires d'objets, tels que le premier objet se trouve dans un groupe et l'autre objet est dans un autre groupe. Les deux clusters dont la similarité moyenne est la plus haute sont fusionnés ensemble pour former un nouveau cluster. Le lien moyen ou lien d'agrégation définit la similarité entre deux parties par :

$$S_{pr} = \text{moyenne}(S_{ir}, S_{jr}) \quad \textbf{(3.13)}$$

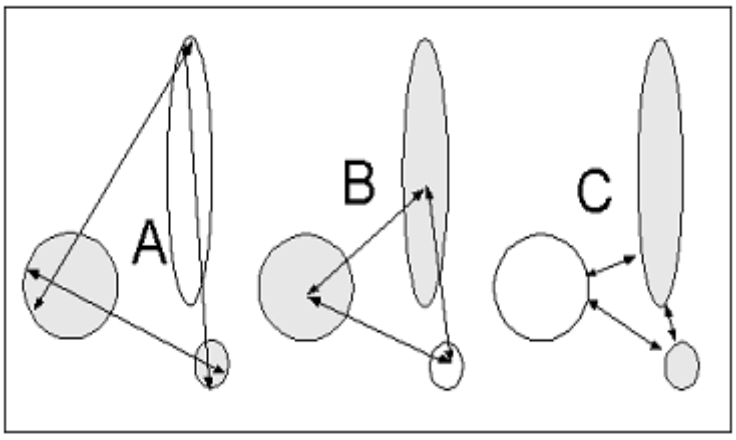


Figure 3. 16--A lien du diamètre. B lien moyen. C lien du saut minimum.

3.3.1.1.4 La méthode de Ward

Selon cette méthode proposée par Ward (1963), les fusions entre les clusters à n'importe quelle étape de la méthode sont choisies de façon à minimiser une fonction objective de sorte que la somme des inerties des groupes obtenus reste la plus petite possible : cela revient à favoriser les regroupements les plus compacts possible dans l'espace (euclidien) de données (Bradley et al, 1998). Ce critère est définit par (Boubou, 2006) :

$$D(C_1, C_2) = \frac{n_{C1} n_{C2}}{n_{C1} + n_{C2}} d^2(g_{C1}, g_{C2}) \quad \textbf{(3.14)}$$

Avec :

- g_{C1} : le centre de gravité de C_1
- g_{C2} : le centre de gravité de C_2

La distance entre 2 classes C_1 et C_2 est définie par la distance entre leurs centres de gravité.

$$D(C_1, C_2) = d(g_{C1}, g_{C2}) \quad \textbf{(3.15)}$$

Les méthodes présentées ci-dessus sont de type agglomeratives. Ce type de classification hiérarchique présente l'avantage d'être facile à implémenter. Cependant cette méthode est très coûteuse et génère un temps de calcul très élevé. Parmi les algorithmes agglomeratifs on peut citer :

- **C**lustering **U**sing **RE**presentatives(**CURE**)
- **B**alanced **I**terative **R**educing and **C**lustering using **H**ierarchies(**BIRCH**)
- **RO**bust **C**lustering using links (ROCK).

Les algorithmes de classification descendante (approche de haut en bas) commence par un cluster contenant tous les objets de données, et procède par division successive jusqu'à obtenir une partition formée uniquement de singletons. Parmi les algorithmes les plus anciens, l'algorithme de Williams et Lambert qui divise la plus grande classe en deux classes, l'algorithme de Hubert qui propose de diviser la classe de plus grand diamètre et l'algorithme TSVQ (Tree Structured Vector Quantization) qui a été proposé par Gersho et Gray. Malgré ses nombreux inconvénients la classification descendante présente quelques avantages par rapport aux algorithmes de classification automatique. La méthode de la classification hiérarchique descendante ne nécessite pas l'utilisation d'un seuil arbitraire pour la formation

des classes qui peut éventuellement mener la recherche d'une classification d'un ensemble de données à une fausse direction.

3.3.2 La classification par partition

Dans cette section, nous présentons les principaux algorithmes de partitionnement de données. Ces algorithmes divisent directement un ensemble de données en *k* classes telles que chaque classe doit contenir au moins un individu et chaque individu doit appartenir à une classe unique contrairement à la classification dite floue qui n'impose pas cette condition.

Un algorithme de clustering par partition obtient une partition unique de données au lieu d'une structure de regroupement, tel que le dendrogramme produit pour une méthode hiérarchique. Les algorithmes de classification par partition sont avantageux dans les applications impliquant des ensembles de données volumineux pour lesquels la construction d'un dendrogramme nécessite un calcul complexe. Un problème qui accompagne l'utilisation d'un algorithme de classification par partition est le choix du nombre de classes de sortie désirés. Les techniques de clutering par partition produisent généralement des clusters en optimisant une fonction objective définies localement (sur un sous-ensemble de vecteurs de données) ou globalement (définie sur tous les vecteurs) qui traduit que les objets doivent être similaires au sein d'une même classe, et dissimilaires d'une classe à l'autre. Pour une classification en *k* classes, ces algorithmes génèrent une partition initiale, puis cherchent à l'améliorer en réaffectant les individus d'une classe à l'autre ce qui permet la possibilité qu'une pauvre partition initiale pourrait être corrigée ultérieurement (Boubou, 2006, Jain et al, 1999).

Contrairement aux méthodes traditionnelles hiérarchiques qui ne révisent pas les classes (intermédiaire) une fois qu'elles sont construites, les algorithmes de classification par partition améliorent progressivement les clusters avec les vecteurs de données appropriés, cela se traduit par des clusters de haute qualité. En pratique, l'algorithme de clustering est généralement exécuté plusieurs fois avec différents états et paramètres d'initialisation, et la meilleure classification obtenue, selon les indices de validité de clustering, pour toutes les partitions est retenue.

La plupart des méthodes de partitionnement de données nécessitent de définir le nombre de groupes cherché au préalable par l'utilisateur, bien que certaines méthodes autorisent de ne pas définir le nombre final de clusters au préalable qui peut être déduit pendant l'analyse, il peut également faire partie d'une fonction d'erreur (Vesanto et Alhoniemi, 2000). L'algorithme général d'une classification par partition comprend les étapes suivantes :

1. Déterminer le nombre de clusters

2. Initialiser les centres des clusters
3. Partitionner l'ensemble de données
4. Calculer les centres des clusters (faire une mise à jour)
5. Si le partitionnement est inchangé (ou l'algorithme a convergé), arrêt ; sinon aller à l'étape 3.

Les algorithmes de classification par partition sont les plus utilisés parmi les algorithmes de clustering. Ces algorithmes minimisent un critère de classification donnée en déplaçant itérativement les points de données entre clusters jusqu'à ce qu'une partition optimale (localement) soit atteinte. La fonction la plus fréquemment utilisée dans les techniques de clustering par partition est l'erreur quadratique, qui a tendance à bien fonctionner avec les groupes isolés et compact. L'erreur quadratique pour un regroupement *R* d'un ensemble d'objets *X* (contenant K clusters) est donnée par :

$$e^2(X,R) = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i - c_j\right\|^2 \qquad \textbf{(3.16)}$$

où x_i est le *i*-ème objet appartenant au *j*-ième cluster et c_j est le centroïde du *j*-ième cluster.

K-means est l'algorithme le plus simple et le plus couramment utilisé parmi les méthodes utilisant l'erreur quadratique comme critère de regroupement (Xu et Wunsch, 2005 ; Jain et al, 1999). Les classes sont représentées par leurs centroïdes, qui correspondent à la moyenne de l'ensemble des objets contenus dans la classe, il offre donc la possibilité de manipuler de plus grandes bases de données que les méthodes hiérarchiques. Il commence par une partition initiale aléatoire et maintient la réaffectation des objets à des groupes en se basant sur la similarité entre les objets et les centres des classes jusqu'à ce qu'un critère de convergence soit atteint (par exemple, il n'y a pas de réaffectation de n'importe quel objet d'un cluster à l'autre, ou l'erreur quadratique se stabilise après un certain nombre d'itérations. L'algorithme K-means est très populaire, car il est facile à mettre en œuvre, et sa complexité algorithmique est également intéressante. Cependant, ses inconvénients sont nombreux. Un problème majeur rencontré lors de l'utilisation de cet algorithme est la sensibilité de la partition finale aux données initiale. Ainsi, K-means peut converger vers un minimum local si la partition initiale n'est pas correctement choisie. De plus cet algorithme nécessite évidement de définir un centroïde pour chaque cluster dont le calcul de ces moyennes est très sensible aux données aberrantes. Pour pallier ces limitations, de nombreux travaux ont été proposées dans la littérature. Par exemple, (Hasan et al, 2009) introduisent l'initialisation alternative à la méthode aléatoire utilisée par K-means pour améliorer la qualité et la vitesse de convergence.

Ainsi pour résoudre notamment le problème de sensibilité aux données aberrantes, un autre type de méthodes a été développé, à savoir les K-medoïdes, dont l'algorithme PAM (Partitioning Around Medoids) est un exemple typique. La méthode K-means utilise un centroïde pour représenter chaque cluster, cela veut dire, qu'un objet avec de grandes valeurs de paramètres peut perturber la distribution des données. En utilisant des médoïdes pour représenter les clusters plutôt qu'un centre de gravité, les méthodes K-médoïdes sont moins sensibles aux données aberrantes. La médoïde est l'objet le plus centralement situé dans un cluster. Pour les méthodes K-médoıdes, *k* objets de données sont sélectionnés aléatoirement comme médoïdes représentant *k* cluster et chaque objet de l'ensemble de données est affecté au groupe ayant la plus proche médoïde (ou le plus similaire) pour cet objet de données. Après le traitement de tous les objets de données, une nouvelle médoïde est déterminée pour chaque cluster et tout le processus est répété. Une autre fois tous les objets de données sont affectés aux clusters en se basant sur les nouvelles médoïdes. A chaque itération, les médoïdes changent leurs localisations étape par étape. En d'autres termes, les médoïdes se déplacent à chaque itération. Ce processus se poursuit jusqu'à ce qu'il n'y ait aucun mouvement de médoïdes. Par conséquence, les *k* clusters qui représentent l'ensemble de données sont identifiés (Singh et Chauhan, 2011 ; Boubou, 2007).

La représentation des clusters par des médoïdes est moins sensible aux données aberrantes. Cependant, l'absence d'un centroïde qui représente les données se fait au détriment de la complexité. De plus comme pour les K-means, il est nécessaire de spécifier préalablement le nombre de clusters *k.* afin de résoudre partiellement le problème de complexité temporelles pour les algorithmes de type K-médoïdes, l'algorithme CLARANS (Clustering Large Applications bases upon randomised search) est très souvent utilisé. CLARANS ne considère qu'un échantillon des données (le voisinage du médoïde que l'on cherche à échanger) à chaque itération pour remplacer le médoïde existant (Singh et Chauhan, 2011 ; Labroche, 2012).

3.3.2.1 La méthode de K-means

K-means est l'un des algorithmes de classification non supervisées les plus simples et les plus utilisés. La procédure de classification d'un ensemble de données en un certain nombre de clusters (k clusters) fixé a priori est aussi simple et facile. L'idée principale est de définir *k* centres initiaux, un pour chaque cluster. Ces centres devraient être placés d'une manière rusée, car différentes emplacements causes des résultats différents. Donc, le meilleur choix est de les placer loin les uns des autres autant que possible. La prochaine étape est de prendre chaque point appartenant à l'ensemble de données et de l'associer au centre le plus proche. Apres

avoir affecter tous les points aux clusters représentés par les centres de gravités, la première étape est terminée. L'étape suivante consiste à recalculer *k* nouveaux centres de gravité comme barycentre des groupes issus de l'étape précédente. Après avoir calculé ces *k* nouveaux centres de gravité, une nouvelle phase d'affectation de chaque objet de données au nouveau centre de gravité le plus proche. A chaque itération, les *k* centres de gravité changent leurs localisations étape par étape. Ou en d'autres termes, les centres se déplacent à chaque itération. Ce processus se poursuit jusqu'à ce qu'il n'y ait aucun mouvement des centres de gravité. Par conséquence, les k clusters qui représentent l'ensemble de données sont alors identifiés (Kanungo, 2002; Boubou, 2007). Enfin, cet algorithme peut être résumé comme suit (voir figure 3.16):

Algorithme 3.2 : Pseudo code de l'algorithme d'apprentissage de K-means (Candillier, 2006).

1. choisir aléatoirement *k* objets de la base de données comme centroïdes initiaux représentant les *k* clusters recherchés;
2. assigner chaque objet au cluster dont le centroïde est le plus proche;
3. puis tant qu'au moins un objet change de cluster d'une itération à l'autre:

- mettre à jours les centroïdes des clusters en fonction des objets qui leur sont associés:

$$\vec{\mu}_k = \frac{1}{N_k} \sum_{i \in D_k} \vec{x_i}$$

- assignations des objets en fonction de leur proximité aux nouveaux centroïdes

$$D_K = \left\{ i \in D \middle| \forall j \in [1,k] et,\ j \neq k, dist(\vec{x}_i, \vec{u}_k) < dist(\vec{x}_i, \vec{u}j) \right\}$$

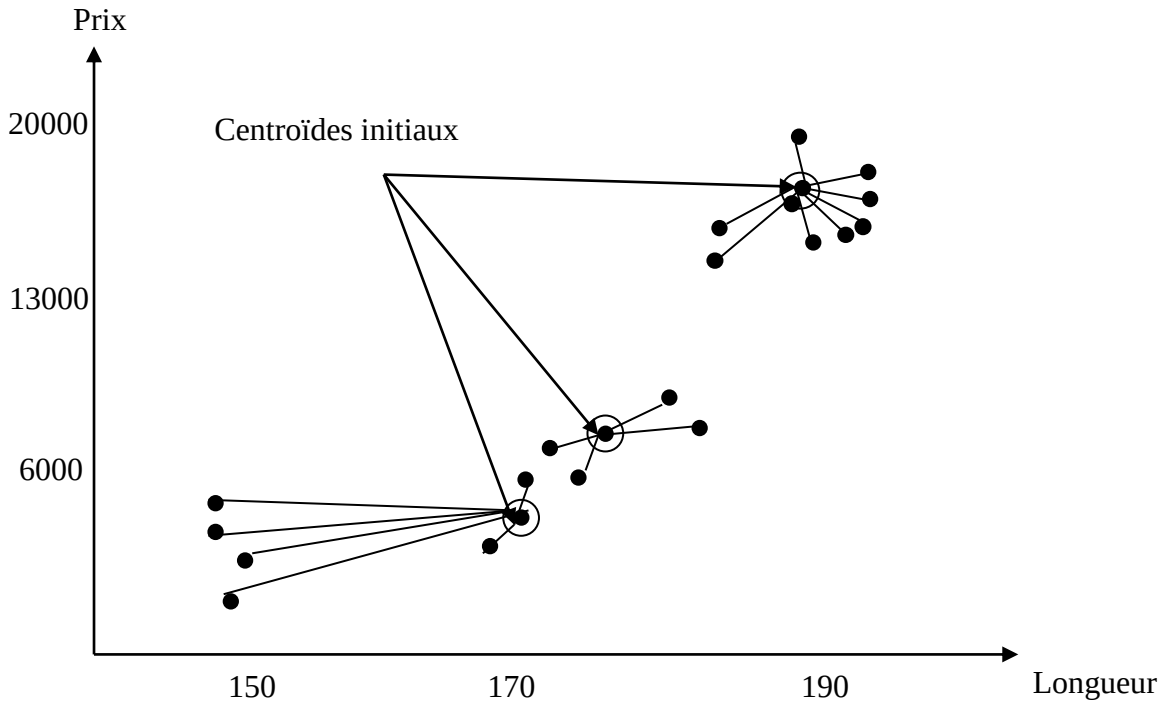
L'immense popularité de l'algorithme K-means est bien méritée. Il est simple, direct, et est basé sur un solide fondement de l'analyse des variances. L'algorithme K-means souffre aussi de tous les inconvénients habituels:

- Le résultat dépend fortement du choix initial des centres de gravité.
- risque de découvrir un minimum local au lieu d'une solution optimale

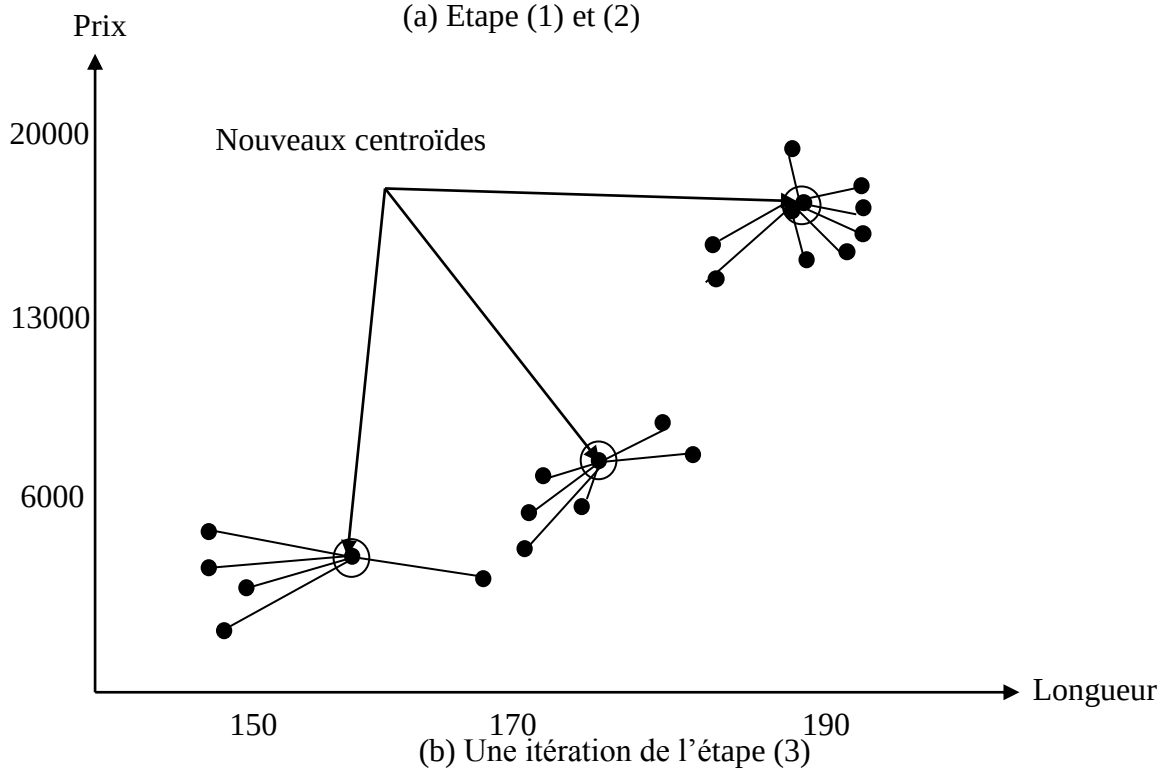
-le nombre de cluster *k* n'est pas connu au préalable.

-Le processus est sensible aux valeurs aberrantes.

-Seuls les attributs numériques sont couverts



(a) Etape (1) et (2)



(b) Une itération de l'étape (3)

Figure 3. 17--Clustering K-means (Candillier, 2006).

La popularité de l'algorithme K-means a donnée naissance à de nombreuses autres extensions et modifications (Bradley et Fayyad, 1998). La distance de Mahalanobis peut être utilisée pour couvrir les groupes hyper-ellipsoïdes (Mao et Jain, 1996). Avec l'algorithme K-means, les centres sont recalculés après chaque assignation d'un objet à un cluster, au lieu d'attendre

l'affectation de tous les objets pour mettre à jour les centres de gravités. Cette approche conduit généralement à de meilleurs résultats comparés à ceux de la méthode des centres mobiles et la convergence est également plus rapide.

3.3.2.2 La méthode K-médoïdes

L'objectif du regroupement K-médoïdes est de trouver un ensemble de clusters bien séparés de telle sorte que chaque groupe soit représenté par un de ses individus qui sont appelés médoïdes. Nous avons déjà mentionné que c'est une solution facile puisqu'elle couvre n'importe quels types de données. De plus les médoïdes sont moins sensible aux données aberrantes. Le processus des méthodes K-médoïdes commence en choisissant *k* points de façon aléatoire qui représentent les médoïdes initiaux. Après avoir sélectionné les médoïdes, les clusters sont définis comme le sous-ensemble de points respectivement le plus proche aux médoïdes, et la fonction objective est définie par la distance moyenne ou une autre mesure de dissimilarité entre un point et un médoïde (Boubou, 2006 ; Berkhin, 2002).

Deux principaux algorithmes de type K-médoïdes sont l'algorithme PAM (Partitioning Around Medoids) et l'algorithme CLARA (Clustering LARge Applications) sont discutés dans ce travail. PAM est l'optimisation itérative qui combine la relocalisation de points entre les clusters avec renomination des points comme médoïdes potentiels. Le principe du processus est basé sur une fonction objective, qui est évidemment une stratégie coûteuse. Le principe général des méthodes K-médoïdes est le suivant :

1. Choisir un ensemble de médoïdes,
2. Affecter chaque individu au médoïdes le plus proche,
3. Itérativement remplacer chaque médoïdes par un autre si cela permet de réduire la distance globale.

L'algorithme PAM est plus robuste que la méthode K-means en présence du bruit, cependant cette méthode est d'une complexité de calcul élevé. En effet, l'algorithme PAM forme des groupes en examinant tous les objets qui ne sont pas médoïdes. Ce qui impose un coût de calcul coûteux de $O(k(n-k)^2)$ à chaque itération. Par conséquent, cette méthode est très efficace pour le regroupement de petites bases de données et très coûteuse en cas de bases de données volumineuses. D'où la proposition d'autres algorithmes, tels que CLARA pour traiter les données de taille moyenne, pourrait faire l'objet d'un travail important.

CLARA est un algorithme introduit par (Kaufman et Rousseeuw, 1990(b)) pour traiter les données multidimensionnelles, cette méthode effectue une recherche locale des représentants

en opérant sur plusieurs échantillons de données de taille S extraits de l'échantillon total. Ensuite l'algorithme PAM est appliqué à chacun d'entre eux et le résultat retenu est le meilleur parmi les différents résultats. La complexité de chaque itération est de l'ordre $O(ks^2+k(n-k))$, où s est la taille de l'échantillon, k le nombre de groupes, et dépend de la méthode d'échantillonnage et la taille de l'échantillon (Singh et Kumar, 2013). L'inconvénient principal de cette méthode est que les paramètres d'échantillonnage sont choisis expérimentalement. En effet, CLARA ne peut pas trouver le meilleur regroupement si un éventuel médoïde échantillonnée n'est pas parmi les meilleurs k médoïdes. L'efficacité de CLARA dans le regroupement de grandes bases de données se fait au détriment de la qualité de clustering.

Pour regrouper les données de taille volumineuse, une méthode appelée CLARANS a été proposée par (Ng et Han, 1994). CLARANS combine à la fois les algorithmes PAM et CLARA en recherchant uniquement le sous-ensemble de la base de données (Singh et Kumar, 2013). Cette méthode fait une recherche stochastique basée sur différents paramètres permettant de borner le nombre d'itérations de la méthode, ainsi que sur l'échantillonnage aléatoire. Etant donné k le nombre de clusters recherchés, un partitionnement de données consiste en un ensemble de k médoïdes, auxquels sont associés l'ensemble des objets en fonction de leur proximité avec ces médoïdes. Les étapes principales de la méthode sont les suivantes (Candillier, 2006):

1. sélectionner un échantillon représentatif des données;
2. itérer un certain nombre fixé de fois;
    - (a) choisir une solution aléatoire : un ensemble de k médoïdes;
    - (b) itérer un certain nombre fixé de fois :
    - choisir une solution voisine de la solution courante, par modification aléatoire de l'un des médoïdes de la solution;
    - conservation du voisin comme nouvelle solution courante si l'inertie globale de la partition est inférieure à celle de la solution précédente;
    - (c) stocker la solution optimale locale trouvée
3. retourner la meilleure des solutions optimales locales trouvées.

CLARANS permet d'extraire des classes de meilleure qualité par rapport aux méthodes PAM et CLARA; cependant cette méthode est sensible aux paramètres choisis et d'une complexité de l'ordre $O(k.n^2)$.

3.4 Comparaison des algorithmes de la classification automatique

Le sujet de clustering a été largement étudié dans de nombreuses disciplines scientifiques et au fil des années, une variété de différents algorithmes ont été développés. Cependant, la classification automatique demeure un problème difficile, qui combine des concepts de divers domaines scientifiques (tel que les bases de données, apprentissage artificiel, reconnaissance des formes, statistiques). Deux études récentes sur le sujet de clustering (Jain et al, 1999 ; Xu et Wunsch, 2005) ont proposé un résumé complet des différentes applications et algorithmes de clustering. Ces algorithmes sont différents suivant les mesures de proximité qu'ils utilisent, la nature des données qu'ils traitent et les objectifs finals de la classification. Chacune de ces méthodes possède ses points forts et ses points faibles (Boubou, 2006).

Au cours des dernières années, plusieurs chercheurs ont conclu que les méthodes partitives sont meilleures par rapport aux méthodes hiérarchiques dans le sens qu'elles ne dépendent pas des clusters précédemment trouvés. Les méthodes hiérarchiques agglomératives sont utilisées en cas des données de petites tailles à cause de leur complexité très élevée. Cependant, s'il y a des problèmes de temps d'exécution, alors la méthode K-means est utilisée. D'autre part, les méthodes partitives font des estimations implicites sur la forme des données. Par exemple, K-means essais pour trouver les clusters sphériques (Boubou, 2006 ; Halkidi et al, 2001).

Les algorithmes partitifs sont principalement applicables aux données numériques. Cependant, il y a quelques variantes de l'algorithme K-means tel que K-mode, qui manipule des données catégorielles. K-mode est basé sur la méthode K-means pour identifier les clusters tandis qu'il adopte de nouveaux concepts afin de manipuler les données catégorielles. Ainsi, les centres des clusters sont remplacés par des "modes", une nouvelle mesure de dissimilarité utilisée pour traiter les objets catégoriels (Halkidi et al, 2001). De plus, les méthodes partitives sont basées sur certaines hypothèses de partitionnement de l'ensemble de données. Ainsi, ils doivent préciser le nombre de clusters à l'avance, sauf pour CLARANS, qui doit préciser en entrée le nombre maximum de voisins d'un nœud ainsi que le nombre de minimum locaux qui seront trouvés afin de définir un partitionnement d'un ensemble de données. Une autre caractéristique des algorithmes partitive est qu'ils sont sensibles aux données bruités. Ainsi, ces méthodes ne sont pas bien appropriées pour découvrir les clusters avec des formes non convexes. Le résultat du processus du regroupement est un ensemble de points représentatifs des clusters obtenus. Ces points peuvent être des centres ou des médoïdes des clusters selon l'algorithme de regroupement. En ce qui concerne les critères de regroupement, l'objectif des algorithmes partitifs est de réduire au minimum la distance des

objets au point représentatif du cluster. Ainsi, K-means vise à minimiser la distance des objets appartenant à un cluster au centre de ce cluster (medoïdes pour PAM.). CLARA et CLARANS comme mentionné ci-dessus, sont basées sur le critère de regroupement des PAM. Cependant, ces algorithmes considèrent des échantillons de données sur lesquels un regroupement est appliqué et par conséquent ils peuvent traiter de plus grande base de données que l'algorithme PAM. L'inconvénient principal de cette approche est que son efficacité dépend de la taille de l'échantillon. En outre, les résultats du clustering sont basés uniquement sur des échantillons d'un ensemble de données.

La méthode des cartes auto-organisatrices peut être vue comme une extension de l'algorithme des K-means. Comme K-means, SOM vise à minimiser une fonction objective convenablement choisie. Cette fonction de coût doit tenir compte, d'une part, l'inertie interne de la partition, et chercher, d'autre part, à assurer la conservation de la topologie. Une manière de réaliser ce double objectif consiste à généraliser la fonction d'inertie utilisée par l'algorithme des K-means en introduisant dans l'expression de cette fonction des termes spécifiques qui sont définies à partir de la carte. Cela est réalisé par l'intermédiaire de la distance définie sur la carte et de la notion de voisinage qui lui est attachée. Si la notion de voisinage lui est particulière, la carte de Kohonen a des analogies avec certaines méthodes présentées précédemment, et des propriétés qui permettent d'envisager des éventuels couplages. La carte de Kohonen est robuste – au sens où le résultat ne peut être grandement modifié par l'ajout d'un nouvel élément à la base de données si celui-ci n'est pas une valeur erronée (aberrante). Cette propriété est aussi vérifiée par la méthode des centres mobiles mais n'est pas partagée par la classification ascendante hiérarchique dont le résultat peut être remis en cause par l'apport d'un individu supplémentaire. Par contre, cette dernière est la seule à fournir exactement le même résultat quand on relance l'algorithme car les autres – qui aboutissent à un minimum local de la somme des carrés des écarts aux centres de classes – dépendent de l'ordre de présentation des individus et des paramètres d'initialisations. Ces méthodes peuvent être complémentaires et donner naissance à des combinaisons hybrides du type centres mobiles – classification hiérarchique (dont on peut trouver une présentation dans (Wong, 1982), ou carte de Kohonen – classification hiérarchique.

3.5 Conclusion

Nous venons de présenter dans ce chapitre les algorithmes de clustering les plus classiques. Plusieurs méthodes sont proposées dans littérature pour le problème générale de la classification non supervisée. Ces méthodes diffèrent par les mesures de proximité qu'ils

utilisent, la nature des données qu'ils traitent et l'objectif final de la classification. Chacune de ces méthodes possède ses points forts et ses points faibles. Malgré le nombre important de méthodes de clustering, plusieurs problématiques restent encore ouvertes dans le cadre de clustering. Un problème très souvent rencontré concerne la difficulté de fixer les paramètres en entrée des méthodes par l'utilisateur, ainsi que l'évaluation des résultats et la comparaison des différentes méthodes. La disponibilité d'une vaste collection d'algorithmes de clustering dans la littérature peut facilement confondre un utilisateur tentant de choisir un algorithme approprié pour le problème considéré. En outre, il n'existe pas d'algorithme de clustering qui peut être universellement utilisé pour résoudre tous les problèmes. Par conséquent, il est important d'étudier soigneusement les caractéristiques du problème considéré, afin de sélectionner ou concevoir une stratégie de regroupement appropriée.

Partie 2

Classification automatique des paramètres météorologiques de la région d'Annaba: approches proposées

Chapitre 4

Approche de classification à deux niveaux

4.1 Introduction

Notre objectif dans cette thèse consiste à proposer une approche de classification non supervisée basée sur les cartes auto-organisatrices de kohonen est d'évaluer l'utilisabilité des SOMs et d'autres méthodes d'intelligence computationnelle dans l'analyse et la modélisation des problèmes de l'informatique environnementale. SOM a été choisi comme une méthode principale en raison de ses propriétés, qui sont adaptées à la fois pour le regroupement et la visualisation. L'approche de clustering développée pour l'identification des types de jours météorologiques pour la région d'Annaba est représentée par la figure 4.1. Cette approche est composée de deux niveaux. Dans le premier niveau de classification, nous avons utilisé une carte de kohonen bidimensionnelle. Le nombre de neurones de sortie est significativement plus grand que le nombre désiré de groupes. Cela nécessite plusieurs neurones pour représenter un seul groupe, plutôt qu'un seul neurone. Ainsi, dans le deuxième niveau de classification les neurones de sortie sont regroupés de tel sorte que les neurones sur la carte sont divisés en autant de régions différentes que le nombre désiré de groupes. Chaque vecteur de données peut être affecté à un groupe quelconque en fonction de son vecteur prototype.

Les algorithmes classiques de clustering peuvent être utilisés pour le regroupement des neurones de sortie du SOM. Toutefois, en raison des inconvénients des différents types de ces algorithmes, nous devons choisir une méthode bien appropriée pour regrouper les unités du SOM. Pour découvrir l'algorithme le plus approprié pour regrouper à priori les vecteurs prototypes de la carte de kohonen, nous avons proposé un schéma de regroupement complet basé sur la comparaison des performances entre les algorithmes de clustering candidats tels que: PAM, K-means, et la classification hiérarchique (méthode de Ward) (Khadir et al, 2010; Khedairia et Khadir, 2008c). Ainsi dans (Khedairia et Khadir, 2008a; Khedairia et Khadir, 2008b) différentes approches de classification non supervisée basées sur SOM, K-means et l'ACP ont été proposées afin d'effectuer une classification des données météorologiques pour la même région en utilisant une base de données, composée de quatre paramètres météorologiques, collectés durant la période de 1995 à 1999. Il est à noter qu'une phase de prétraitement des données environnementales est nécessaire pour préparer les données à la phase de clustering en éliminant le bruit, corrigeant les erreurs et normalisant les données. En effet les données manquantes et les données aberrantes constituent les principaux problèmes qu'on doit faire face.

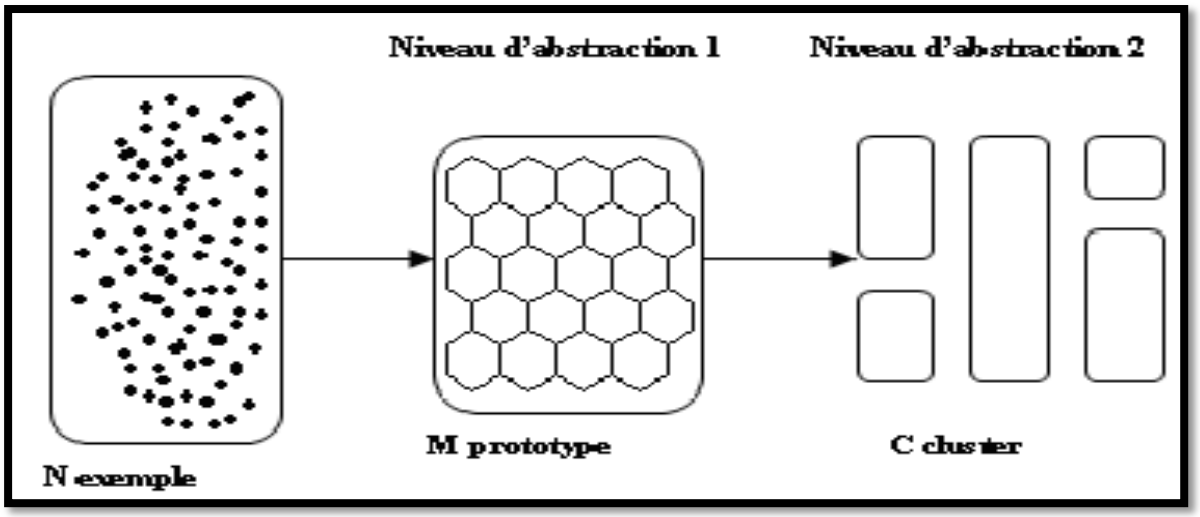


Figure 4. 1–Approche de classification à deux niveaux utilisant une carte de kohonen (Vesanto et Alhoniemi, 2000).

La première phase de l'approche de classification à deux niveaux (classification par la carte de kohonen) peut être résumée comme suit:

Algorithme 4.1 : Pseudo code de La première phase de l'approche de classification à deux niveaux (premier niveau)(Dreyfus et al, 2004; Dean, 2010).

1– Initialiser tous les poids w_{ij}. Initialiser le compteur du nombre d'itération à 0, initialiser la fonction de voisinage (d_0). Initialiser le taux d'apprentissage initial (a_0) et le nombre total des itérations (T). Par ailleurs, choisir la topologie et le nombre de neurone de la carte de kohonen.

2– Pour l'itération courante (t), calculer les valeurs du taux d'apprentissage, et la fonction de voisinage.

3– Choisir un vecteur x d'entrée parmi l'ensemble des données.

4– Calculer la distance entre le vecteur d'entrée sélectionné et tous les neurones de la carte, et identifier le neurone gagnant.

5– Identifier les neurones voisins du neurone gagnant par le biais de la fonction de voisinage.

6– Mettre à jour les poids du neurone gagnant et ses neurones voisins selon la fonction :

$$w_j(t+1) = w_j(t) + a(t)d(t)[x(t) - w_j(t)].$$

7– Calculer $t = t + 1$ et si $t < T$ aller étape 2, et sinon arrêt.

Pour faire face aux neurones situés entre les clusters de données et qui ne participent pas à la compétition, nous avons considéré une propriété «âge» pour chaque neurone. L'âge d'un neurone est incrémenté d'une unité à chaque fois ou le neurone a gagné la

compétition pour cette itération. A la fin de la première étape de la classification, les neurones qui n'ont pas participé à l'apprentissage (ou neurones morts) auront un âge nul. Nous enlevons ensuite ces neurones morts qui pourront perturber la classification dans le deuxième niveau. L'algorithme global proposé se résume comme suit:

Algorithme 4.2 : Pseudo code l'algorithme de regroupement du SOM (deuxième niveaux).

1– Apprentissage des données d'entrées par le SOM.

2– suppression des neurones morts.

3– Clustering the SOM en utilisant un algorithme de classification par partition.

4– Sélection de la solution optimale. Pour identifier le nombre de clusters *k* le plus approprié, l'algorithme partitif peut être répété pour un ensemble de différents nombres de clusters. Typiquement il peut être répété pour un intervalle de 2 jusqu'à $\sqrt{N}$ (Vesanto et Alhoniemi, 2000), ou *N* est le nombre d'individus dans l'ensemble de données.

Selon les résultats de comparaison, la procédure de classification à deux niveaux semble être plus efficace qu'une approche de clustering directe en utilisant seulement SOM ou K-means, et concernant le deuxième niveau de classification, K-means semble être l'algorithme le plus approprié pour regrouper les vecteurs prototypes de la carte de kohonen. Le tableau 4.1 et le tableau 4.2 montrent consécutivement les résultats de l'indice Davies-Bouldin et les indices de validité externe. Les résultats obtenus de ces indices montrent également que l'approche de classification à deux niveaux SOM et K-means donne de meilleurs résultats pour tous les indices de validité.

Tableau 4. 1-- Comparaison des résultats de l'indice Davies-Bouldin entre K-means et l'approche de classification à deux niveaux pour la classification des paramètres météorologiques.

Nombre de classe	K=2	K=3	K=4	K=5	K=6	K=7
k-means	1.28	1.14	1.13	1.03	1.06	1.13
SOM et K-means	1.08	0.90	0.84	0.79	0.82	0.90

Tableau 4. 2-- Comparaison des résultats des indices de validité externes entre K-means et l'approche de classification à deux niveaux pour l'identification de cinq clusters météorologiques.

Nombre de classe	Rand	Mirkin	Hubert
k-means	0.74	0.26	0.49
SOM et K-means	0.76	0.24	0.52

Le premier avantage de l'approche proposée est la diminution du coût de calcul. La classification des données météorologiques uniquement par K-means engendre un temps de calcul beaucoup plus grand que celui généré par l'approche à deux niveaux de classification. Même avec un nombre d'échantillons relativement petit, plusieurs algorithmes de classification automatique (spécialement les algorithmes hiérarchiques) deviennent très lourds. Pour cette raison, il est nécessaire de regrouper un ensemble de prototypes plutôt que de regrouper directement les données. Considérons un regroupement de *N* échantillons de données en utilisant K-means. Ceci implique de faire plusieurs expériences de classification pour les différentes k-regroupement. La complexité de calcul est proportionnelle à la quantité $\sum_{k=2}^{C_{\max}} NK$, où C_{max} est le nombre maximum présélectionné de clusters. Si on utilise un ensemble de prototypes comme étape intermédiaire, la complexité totale est proportionnelle à $NM + \sum_k MK$, ou *M* est le nombre de prototypes (Vesanto et Alhoniemi, 2000). Avec $C_{max}= \sqrt{N}$ et $M=5\sqrt{N}$, la réduction du temps de calcul est environ $\sqrt{N}$ /15. Certainement, ceci est une évaluation très grossière, puisque beaucoup de considérations pratiques sont ignorées. La réduction est encore plus grande pour les algorithmes agglomératifs, puisqu'elles ne peuvent pas commencer à partir de $\sqrt{N}$ clusters, mais doit commencer par *N* clusters tout en essayant de réduire ce nombre. Un autre avantage de l'approche de classification à deux niveaux est la réduction du bruit. Les prototypes sont des moyennes locales de données et donc moins sensibles aux variations aléatoires que les données originales.

4.2 Prétraitement de données

La base de données utilisée dans ce travail de thèse couvre la période 2003-2004, cette base de données nous a été fournie par la station Samasafia sur une base continue de 24 heures. Les polluants surveillés en permanence comprennent les concentrations des polluants suivants: l'oxyde nitrique (NO), le monoxyde de carbone (CO), l'ozone (O_3), les particules en suspension (PM_{10}), l'oxyde d'azote (NO_x), le dioxyde d'azote (NO_2) et le dioxyde de soufre (SO_2). Un aperçu général de l'ensemble de données utilisées est représenté par le tableau 4.3, où la plage de variation de chaque polluant est calculée. L'ensemble de données utilisé comprend également trois paramètres météorologiques à savoir: la vitesse du vent, la température et l'humidité relative.

Tableau 4. 3--Résumé de la base de données utilisée et la plage de variation des polluants.

2003	Janvier-mars	Avril-Juin	Juillet-Sept	Octobre-Dec
CO(mg/m3)	0-12.8	0-12.8	0-12.8	Pas de données
NO(mg/m3)	0-366	0-88	0-120	0-106
NO_x(Microg/m3)	0-420	0-158	0-316	0-142
NO_2(Microg/m3)	0-145	0-110	0-467	0-89
O_3(Microg/m3)	Pas de données	Pas de données	Pas de données	Pas de données
PM_{10}(Microg/m3)	2-309	0-2137	Pas de données	0-671
SO_2(Microg/m3)	0-79	0-186	0-53	0-56
2004	Janvier-mars	Avril-Juin	Juillet-Sept	Octobre-Dec
CO(mg/m3)	Pas de données	Pas de données	Pas de données	Pas de données
NO(mg/m3)	Pas de données	0-49	0-43	0-32
NO_x(Microg/m3)	Pas de données	0-182	0-59	0-52
NO_2(Microg/m3)	Pas de données	0-293	0-51	0-30
O_3(Microg/m3)	0-87	0-59	0-688	0-116
PM_{10}(Microg/m3)	0-1188	0-122	0-184	0-313
SO_2(Microg/m3)	Pas de données	Pas de données	Pas de données	Pas de données

La détection des valeurs aberrantes peut être considérée comme un processus de prétraitement. Pour cette raison, nous commençons d'abord par la suppression des valeurs aberrantes pour éviter leurs impacts sur les autres méthodes de classification non supervisée. Comme toutes les données environnementales, la phase de prétraitement doit faire face à deux problèmes de données évidentes: les données manquantes et les valeurs aberrantes. Les valeurs aberrantes sont principalement dues au fonctionnement incorrect des instruments ou à des méthodologies incorrectes lors de la collection et de l'analyse. La détection des valeurs manquantes est assez facile, et nous pouvons en déduire facilement la cause. Cependant, la détection des valeurs aberrantes n'est pas une tâche assez facile que la détection des valeurs manquantes. Les données réellement observées à un moment donné sont spatialement distribués, les données collectées sur la température par exemple dépendent à la fois du facteur du temps et de l'espace, et peuvent parfois inclure d'autres facteurs climatiques et non climatiques (Takahashi et al, 2011). Il est à noter qu'il est difficile de détecter toutes les données bruitées de façon instantanée par une détection automatique. Dans notre cas, la détection de valeurs aberrantes est basée sur l'identification des valeurs moyennes pour

chaque paramètre, en tenant compte de l'écart type des paramètres météorologiques. En outre, la visualisation de données peut être aussi très utile pour la détection des valeurs aberrantes. La figure 4.2 illustre des exemples de valeurs aberrantes détectées à l'attribut de la température. Les valeurs erronées détectées sont des valeurs négatives suivies généralement par des valeurs manquantes.

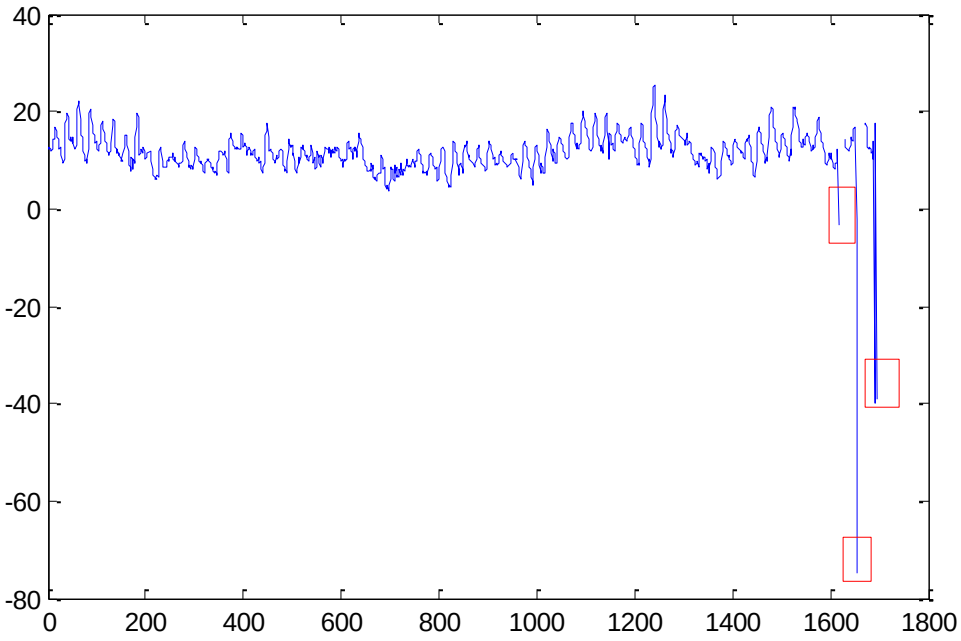


Figure 4. 2--Exemple de valeurs aberrantes dans l'attribut de température

Les données manquantes sont principalement dues à des défaillances des instruments de mesure. Surtout à cause des coupures de courant au niveau des stations Samasafia ou l'acquisition de quelques blocages (Samasafia, 2004). La présence de données manquantes peut invalider l'analyse statistique par la présentation des composants d'erreurs systématiques lors de la classification de données ou l'estimation des paramètres du modèle de prédiction. En outre, si nous estimons les paramètres du modèle en exploitant uniquement les données observées, sans prendre en compte la présence de données manquantes, les estimations obtenues ne pourraient pas être fiables parce que beaucoup d'informations concernant les données manquantes disparaîtraient (Elminir et Abdel-Galil, 2006). Les pourcentages de données manquantes présents dans la base de données pour chaque année sont les suivants: 2003 (08.31%) et 2004 (23,67%). Il est important de noter que l'algorithme d'apprentissage du SOM traite le problème des données manquantes avec élégance. En effet, lors de la recherche du BMU, seulement les données qui ont des valeurs connues sont utilisées dans les

calculs de distance. Ceci implique que les valeurs manquantes sont considérées identiques aux valeurs présentes dans les vecteurs prototypes. Le traitement des données manquantes contenu dans la base de donnes est divisé en deux modes selon la largeur des séquences de donnes manquantes (gap width): pour les séquences de petite largeur, les mesures manquantes ont été remplacées avec des valeurs moyennes en se basant sur les valeurs valides avant et après la mesure manquante. Les séquences de très grande largeur ont été supprimées.

4.3 Approche de classification à deux niveaux (SOM et K-means)

La carte de kohonen est à la fois une méthode de quantification vectorielle et un algorithme de projection de données. La quantification de N échantillons d'apprentissage aux M prototypes réduit l'ensemble de données original à un ensemble plus petit tout en préservant les propriétés originales de l'ensemble de données. Le travail réalisé est basé sur l'utilisation de la librairie somtoolbox, disponible gratuitement (License libre) dans : http://www.cis.hut.fi/projects/somtoolbox. Avant de passer à l'étape d'apprentissage, quelques paramètres d'initialisation de la carte de kohonen doivent être définis: le nombre de neurones, le type de topologie, dimensions de la carte et les vecteurs poids initiales.

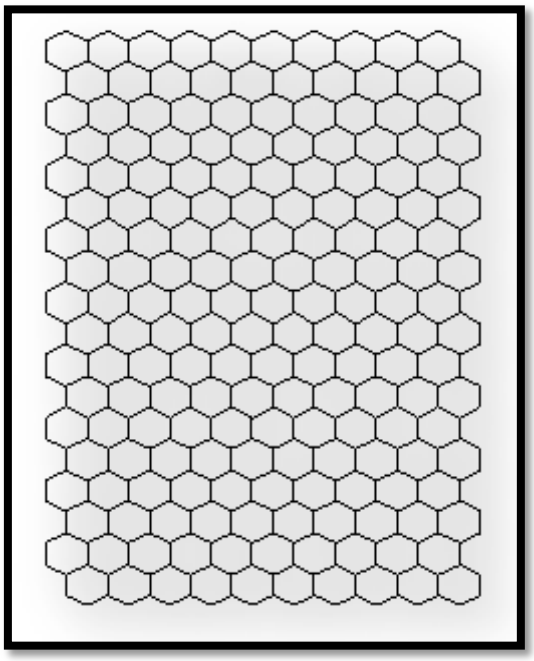


Figure 4. 3–La topologie de la carte de kohonen utilisée pour regrouper les données météorologiques.

La topologie de la carte de kohonen utilisée pour regrouper les données météorologiques est représentée par la figure 4.3 ou les cartes sont reliées aux nœuds hexagonaux adjacents de taille 12×9 pour l'adaptation des situations météorologiques de la région d'Annaba. Une carte

bidimensionnelle permet de fournir une très bonne visualisation des clusters obtenus, tout en s'appuyant sur une propriété très importante pour les cartes de kohonen qui est la fonction de voisinage. Par contre, un réseau unidimensionnel (ou SOM 1-D avec *k* unités de sorties) peut être vu comme une version stochastique de la méthode de regroupement classique K-means, et ses performances sont très semblables à cet algorithme. Pour ces raisons une carte bidimensionnelle, avec une topologie hexagonale et une fonction de voisinage gaussienne a été utilisée pour la classification des paramètres météorologiques de la région d'Annaba. Il n'y a aucune règle explicite permettant de choisir le nombre de neurones d'un réseau de Kohonen (Hautaniemi et al, 2003). Cependant, il existe plusieurs approches permettant de déterminer la taille du SOM. La première approche consiste à prendre un nombre de nœuds approximativement égal au nombre de vecteurs d'entrées. Cette approche semble être utile pour de nombreuses applications lorsque la taille des données en entrée est relativement faible (Kaski 1997). Vesanto et al. (2000) ont utilisés la formule suivante pour déterminer le nombre de neurones:

$$M = \sqrt{(nombre\ d'échantillons}\ .$$

Kuzmanovski et al. (2005) ont utilisé un algorithme génétique pour déterminer le nombre des unités du SOM, tandis que Park et Chung (2006) ont déterminé le nombre de neurones de la carte de kohonen sur la base des conseils d'un expert. Après avoir déterminé le nombre de neurones de la carte de kohonen (on a choisi d'utiliser la méthode de Vesanto et al. (2000)), la taille de la carte doit être déterminée. Fondamentalement, les deux plus grandes valeurs propres des données de l'apprentissage sont calculées et le rapport largeur-longueur de la grille de neurones est déterminé selon ces valeurs propres. Pour déterminer le nombre optimal des unités de la carte de kohonen, plusieurs expériences ont été faites, en changeant le nombre de nœuds et en comparant les performances des résultats obtenus. Lors de l'apprentissage du SOM, il est important de vérifier que la carte de kohonen a été correctement apprise ou non. Pölzlbauer (2004) a présenté plusieurs mesures de qualité pour la carte de kohonen : erreur de quantification, erreur topographique, produit topographique, préservation de voisinage et l'erreur de distorsion du SOM. La mesure de qualité la plus couramment utilisée est l'erreur de quantification ainsi que l'erreur topographique (Pölzlbauer (2004).

4.3.1 Résultats et apprentissage de la carte de kohonen

La base de données utilisée pour l'identification des types de jours météorologiques comprend trois paramètres météorologiques (vitesse du vent, température et humidité relative). Chaque

vecteur météorologique représente une journée (24 h). Pour chaque vecteur, les données sont échantillonnées chaque 3 h pour les trois paramètres météorologiques, ce qui donne un vecteur de 24 dimensions. La carte de kohonen a été linéairement initialisée en premier temps, par la suite un apprentissage séquentiel a été effectué. Différents nombres d'époques ont été considérés pour l'apprentissage du réseau compétitif (100, 200, 250, 350, 500). Les résultats d'application de la méthode SOM sur les données météorologiques sont montrés par la figure 3.3. Les cartes sont reliées aux nœuds hexagonaux adjacents avec des tailles de 12 x 9 en adaptant les situations météorologiques de la région. Une carte de Kohonen à deux dimensions peut être considérée comme une méthode efficace pour le regroupement et la visualisation de données multidimensionnelles, tout en s'appuyant sur sa fonction de voisinage. La carte U-matrix montré sur la Figure 4.4 représente une mesure relative de distance entre les unités colorées du réseau, où la couleur grise (ombre) de l'hexagone indique une mesure de distance entre un nœud quelconque et ses unités voisines. Plus l'ombre est foncé plus la distance est grande, un cluster (classe) qui représente des vecteurs de données similaire peut être vu sur la carte U-matrix comme une zone claire avec des frontières foncées. Plus précisément, Les clusters peuvent être considérés comme des «champs» clairs entre des "montagnes" sombres. Les distances moyennes entre chaque unité et ses unités voisines de la carte peuvent être facilement extraites. Le résultat est une matrice de distance moyenne représentée par la figure 4.4(b) qui peut être vu comme une version moyenne de l'U-matrix. La figure 4.4(b) permet de visualiser la distribution des données de façon très claire. Ainsi, selon la figure 4.4(a) et la figure 4.4(b) quelques régions claires peuvent être regroupées, mais sans frontières claires. Par conséquent, l'utilisation uniquement d'une mesure de distance n'est pas suffisante pour identifier clairement les clusters météorologiques. Les clusters obtenus sur la carte U-matrix sont représentés par des cercles. Selon (Kiang et al, 2004), il est très difficile de regrouper visuellement les neurones de sorties d'un SOM lorsque le réseau de kohonen est fortement peuplé. Dans ce cas, l'identification des frontières de chaque cluster est une tâche très difficile et l'utilisation uniquement de la distance euclidienne pour déterminer les groupes météorologiques n'est pas suffisante. Pour surmonter cette déficience, une combinaison de la mesure de distance et la carte code-couleur a été utilisée. La carte code-couleur du SOM est une méthode basée sur la carte de kohonen pour la visualisation des clusters, selon leurs propriétés (Vesanto, 1999). Les neurones dont les valeurs des paramètres météorologiques sont similaires (proche en termes de distance) évaluent automatiquement des couleurs similaires sur les nœuds de la grille. Les grandes

mesures de distance entre les nœuds du réseau de kohonen sont automatiquement assignées aux différentes couleurs et clusters. L'identification des classes météorologiques est basée sur la sélection des régions de couleur similaire. Dans le cas où les couleurs des nœuds ne sont pas claires pour indiquer les différences entre les clusters, une mesure de distance est alors utilisée pour identifier les frontières de ces clusters. Cependant, selon la carte code-couleur montrée par la figure 4.4(c), il est très difficile d'attribuer quelques régions ambigües à un cluster donné. En conséquence, un deuxième niveau de classification peut être très utile pour enlever l'ambiguïté et valider les résultats de la carte de kohonen.

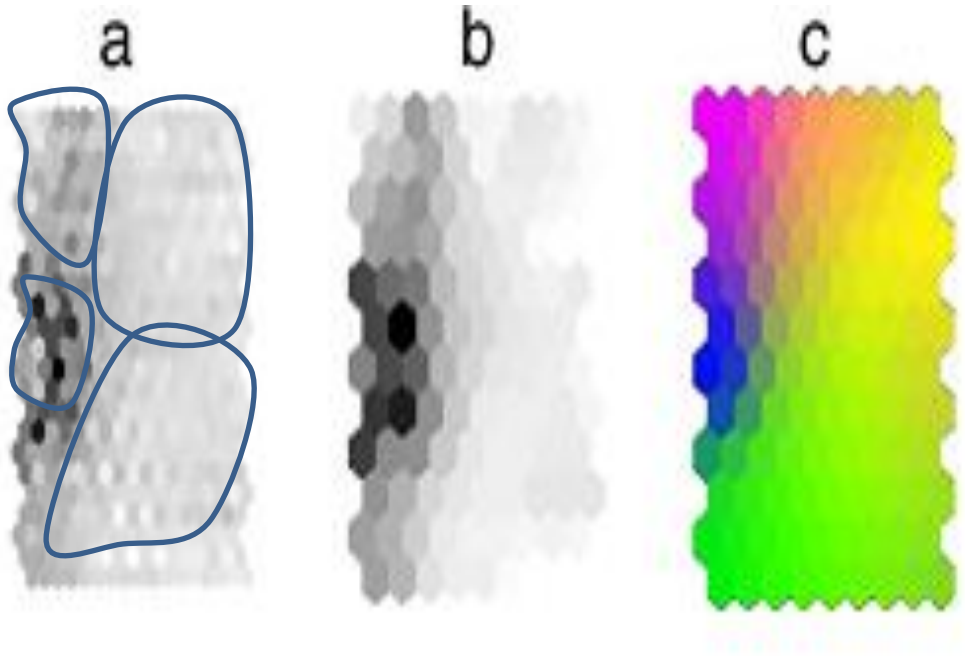


Figure 4. 4--la figure (a)présente l'U-matrix, (b) matrice de distance moyenne, (c) la carte code-couleur.

4.3.2 Affinage des résultats du SOM par K-means

L'algorithme de clustering K-means a été utilisé pour regrouper les unités du SOM pour plusieurs valeurs de *k* (le nombre de groupes dont lequel les données sont partitionnées). En raison du caractère aléatoire du processus de regroupement et du fait que les résultats de ces méthodes dépendent des centres initiaux, de l'ordre de présentation et des propriétés géométriques des données, un nombre relativement élevé d'expériences (50 expériences) ont été exécuté et leurs résultats ont été comparés. La meilleure partition pour chaque (*k*) est sélectionnée selon les indices de validités décrits dans la section 2.8.2. Ainsi, le nombre optimal de groupes parmi les différentes valeurs de *k* est choisi en fonction des indices de validités décrites précédemment. Les résultats de ces indices sont présentés dans la figure 4.5.

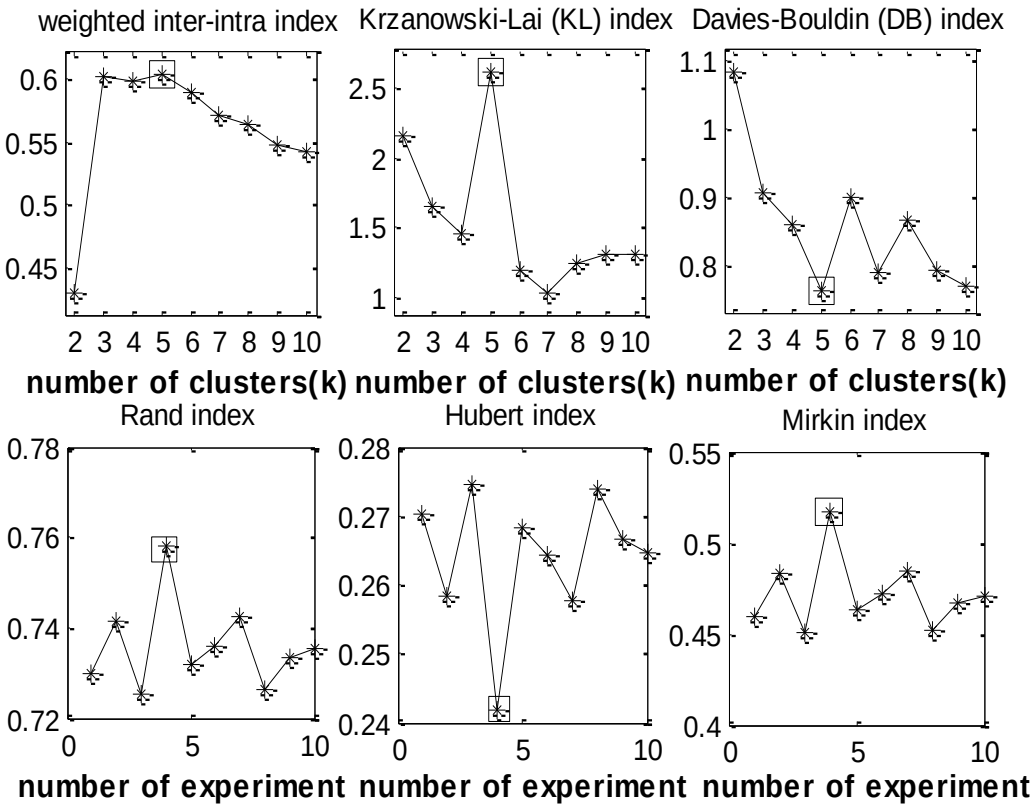


Figure 4. 5--Les indices de validité obtenus pour chaque k clusters.

Selon l'indice Davies-Bouldin, représenté sur la figure 4.5, un pic négatif est remarqué pour k = 5, qui indique le nombre optimal de classes proposée par cet indice. Le même résultat est également proposé par l'indice inter-intra poids où la valeur maximale pour cet indice indique le nombre optimal de clusters. Selon les résultats des différents indices, les clusters obtenus sont bien séparés et homogène. Après avoir identifié le nombre de clusters optimal, le processus de l'identification des clusters météorologiques a été répété 20 fois. Les résultats finaux pourraient ensuite être comparés et l'incertitude du à l'aspect aléatoire lors de l'initialisation et de l'ordre de présentation des données en entrée est éliminé.

L'algorithme d'apprentissage de la carte de Kohonen a permis en premier lieu de regrouper les 500 vecteurs météorologiques de la base de donnes dans les 108 cellules de la carte auto-organisatrice. Le regroupement de ces derniers a permis d'avoir cinq groupes. Les résultats obtenus sont présentés dans la figure 4.6 et les paramètres météorologiques moyennes pour chaque cluster sont présentés dans la figure 4.7. Le groupe C3 (140 vecteurs météorologiques) est caractérisée par une température moyenne qui varie entre 20 et 25 °C pendant la journée et un peu plus bas dans la nuit, ce cluster est caractérisé également par un taux élevé d'humidité pendant la nuit qui se diminue pendant la journée, la vitesse du vent est très faible durant la

nuit et commence à augmenter au cours de la journée, selon la répartition mensuelle des clusters présentés dans le tableau 4.4 ce groupe représente les mois du printemps ainsi que les mois chauds.

Tableau 4. 4–Distribution mensuelle des vecteurs météorologiques pour chaque cluster.

Mois	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Jan	0	26	0	29	0
Fev	1	41	1	5	0
Mar	0	37	5	5	0
Avr	4	29	10	5	0
Mai	1	16	24	9	0
Juin	0	2	22	1	17
Juil	0	0	23	0	24
Aout	16	0	17	0	10
Sept	7	2	16	0	14
Oct	0	1	20	3	5
Nov	0	31	1	8	0
Dec	0	8	0	2	0

Le deuxième cluster météorologique (composé de 194 vecteurs météorologiques) est particulièrement concentré dans les mois d'hiver, ce groupe se caractérise principalement par une faible température et vitesse du vent. Le deuxième cluster est caractérisé également par un taux élevé d'humidité qui dépasse en moyenne 70%. Le premier groupe (composé de 29 vecteurs météorologique distribué principalement dans le mois d'aout et de septembre est caractérisé par une très haute température qui varie Entre [15.2^0, $39,6^0$], avec une grande vitesse de vent une fois comparé à d'autres groupes météorologiques. Le quatrième groupe (composé de 67 vecteurs météorologique distribué principalement dans les mois d'hiver) les paramètres sont similaires au deuxième cluster avec une vitesse de vent élevée qui dépasse 4m/s. Le cinquième groupe (composé de 70 vecteurs météorologique distribué principalement dans les mois d'aout) est presque similaire au premier groupe avec une faible vitesse de vent pendant toute les 24h.

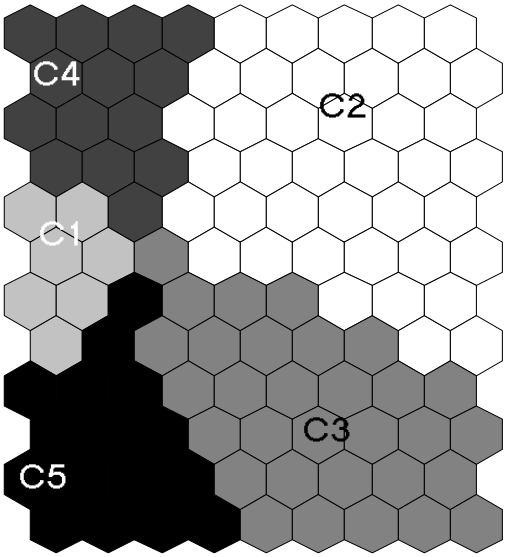


Figure 4. 6–Résultats du deuxième niveau de regroupement.

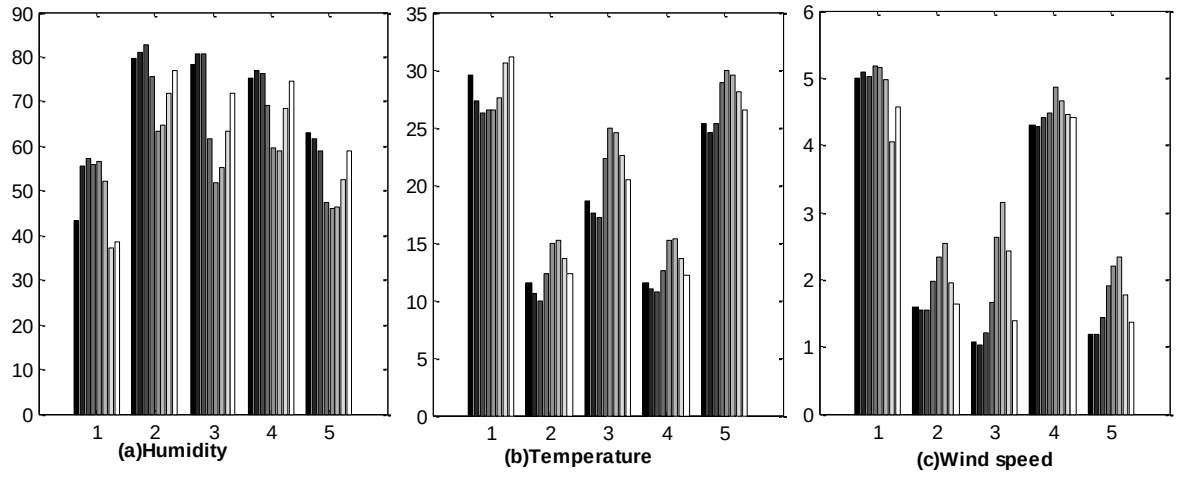


Figure 4. 7– Les paramètres météorologiques moyens pour chaque cluster.

4.4 Carte auto-organisatrice chaotique (CHAO-SOM)

Si l'on demande à une personne au hasard de nous définir ce que signifie le chaos pour elle, elle pourrait dire que cela rime avec «absence de règle». Vient du mot grec "Khaos" signifiant «abime». Le chaos est l'état primordial du monde caractérisé par une confusion de tous les éléments et par l'absence de l'ordre (Talbi, 2010). En 1952, Hodgkin et Huxley (HH) ont présenté un modèle mathématique pour prédire le comportement quantitatif d'un axone isolé d'un calmar géant (Hodgkin, A.L., Huxley, 1952), où les canaux ioniques sont constitués par trois types de portes indépendantes qui s'ouvrent et se ferment en fonction de la valeur du

potentiel de la membrane. Leur idée principale était de considérer la membrane cellulaire et les différents courants ioniques comme un circuit électrique composé de condensateurs, des résistances et des batteries. Le taux d'ouverture de ces portes est décrit par les variables *m*, *n* et *h*. Depuis lors, le modèle est devenu un paradigme pour décrire mathématiquement le fonctionnement des neurones, et de nombreux chercheurs ont étudié sa dynamicité non linéaire. En particulier, il a été découvert qu'avec le changement du courant externe injectée dans la membrane cellulaire, différentes bifurcations peuvent se produire rendant un neurone (HH) oscillatoire, voire chaotique. Le modèle de Hodgkin et Huxley est composé de plusieurs équations différentielles couplées entre elles comme suit

$$\frac{dV}{dt} = I - 120.0m^3h(V - 55.0) \quad \textbf{(4.1)}$$
$$-36.0n^4(C + 72.0)$$
$$-0.24(V + 49.387)$$

$$\frac{dm}{dt} = \frac{0.1(-35-V)}{\exp\left(\frac{-35-V}{10}\right)-1}(1 - m) \quad \textbf{(4.2)}$$
$$-\exp(\frac{-60 - V}{18})m$$

$$\frac{dh}{dt} = 0.07\exp(\frac{-60-V}{20})(1 - h) \quad \textbf{(4.3)}$$
$$-\frac{1}{\exp\left(\frac{-30 - V}{10}\right) + 1}h$$

$$\frac{dn}{dt} = \frac{0.01(-50-V)}{\exp\left(\frac{-50-V}{10}\right)-1}(1 - n) \quad \textbf{(4.4)}$$
$$-0.125\exp(\frac{-60 - V}{80})n$$

Où *V* est le potentiel de membrane, *m* et *h* représentent respectivement le coefficient d'activation de sodium et le coefficient d'inactivation de sodium, et *n* est un coefficient d'activation de potassium. *I* désigne le stimulus actuel.

Prenons une situation qu'un oscillateur neuronal est stimulée en appliquant une force périodique externe. La figure 4.8 montre la série temporelle de la première variable V (t), lorsque l'équation (HH) est stimulée par une force extérieure décrite comme suit :

$$I = I_0 + A\ sin2\pi ft \quad \textbf{(4.5)}$$

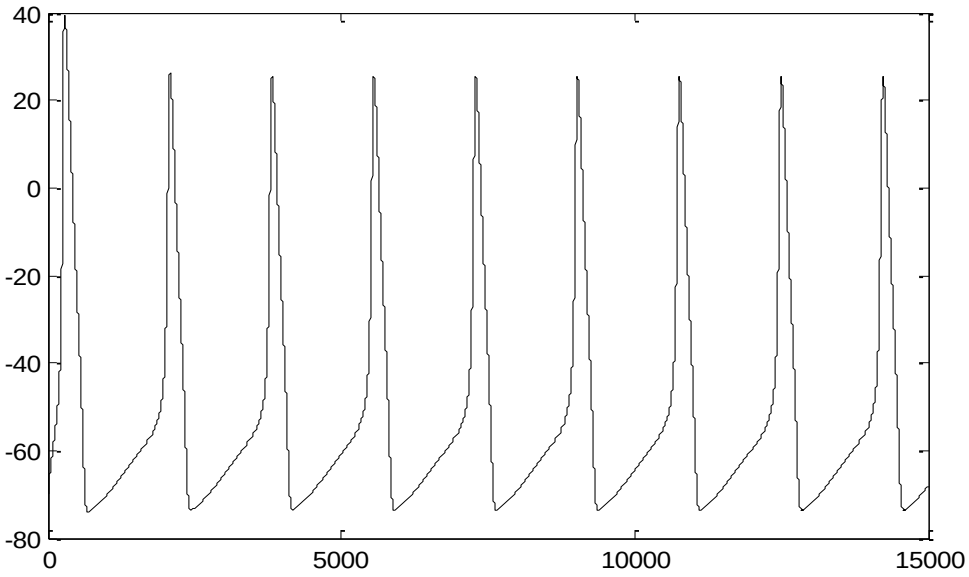


Figure 4. 8--Réponse chaotique de l'équation Hodgkin-Huxley

4.4.1 Algorithme d'apprentissage

L'algorithme d'apprentissage du CHAOSOM est identique à celui du SOM, toutefois, la principale caractéristique du CHAOSOM est l'actualisation du taux d'apprentissage et du coefficient de voisinage au moment des pointes (spikes) générées chaotiquement par l'équation de Hodgkin-Huxley. Le taux d'apprentissage *a (s)* et la fonction de voisinage du neurone $N_{CS}(s)$ de CHAOSOM sont définis comme:

$$a(s) = \frac{1}{\ln(s+2)} \quad \textbf{(4.6)}$$

$$N_{cs}s(s) = \left[-\frac{1}{n_s}s + \frac{m}{6(p+1)} + 1\right] \quad \textbf{(4.7)}$$

Où *p* représente le nombre des pointes de *V(t)* généré par le modèle HH et augmente d'une unité, et *ns* désigne la fréquence décroissante du coefficient de voisinage Ncs (Matsushita et Nishio, 2006).

4.4.2 Approche de classification à deux niveaux utilisant CHAOSOM

Le premier niveau d'abstraction de l'approche de regroupement à deux niveaux est obtenu par la création d'un ensemble de prototypes utilisant un CHAOSOM bidimensionnel. Ces prototypes sont ensuite regroupés dans le second niveau d'abstraction en utilisant CHAOSOM unidimensionnelle. Un SOM unidimensionnelle (ou SOM 1-D avec K unités de sorties) peut être vu comme une forme stochastique de l'algorithme de clustering K-means (Turias et

al, 2006), et ses performances sont très similaires à celles du K-means. Cependant il a été remarqué par (Turias et al, 2006) qu'un SOM unidimensionnel (SOM 1-D) a obtenu de meilleurs résultats que K-means pour le regroupement des données météorologiques.

Il est à noter que les résultats obtenus en appliquant CHAOSOM à deux niveaux pour l'identification des types de jours météorologiques pour la région d'Annaba ont été trouvés presque similaires aux résultats obtenus par l'approche à deux niveaux utilisant SOM et K-means. L'approche CHAOSOM à deux niveaux a été également utilisée pour l'identification des classes des particules en suspension selon les paramètres météorologiques. Les émissions de PM_{10} sont scientifiquement mal connues. En effet, les tailles et natures des particules en suspension sont diverses, ce qui rend la quantification de leur origine et les quantités émises une tâche très difficile. Le regroupement des données permet d'obtenir six groupes ou clusters. Les paramètres de pollution moyenne pour chaque groupe sont présentés dans la figure 4.9. Le cluster C1 est caractérisé par des concentrations élevées de PM_{10} et une faible humidité. Selon (khedairia et khadir, 2012), PM_{10} est négativement corrélée avec l'humidité relative. Ce groupe est également caractérisé par une température élevée ou la plus part des pics de PM_{10} appartient à ce cluster. Le cluster C4 est caractérisé principalement par un taux d'humidité élevé et de faibles concentrations de PM_{10}. L'humidité relative pour ce cluster dépasse une moyenne de 60%, par conséquent plus le taux d'humidité relative est élevé plus le taux d'arrangement de la poussière atmosphérique est élevé aussi. Les paramètres météorologiques moyens ainsi que les concentrations du PM_{10} du cluster C2 sont presque similaires à celles du Cluster C4 avec un faible taux d'humidité. Les clusters C2 et C3 ont des paramètres presque identiques, où la simple différence peut être vue dans le taux d'humidité. Le cluster C5 regroupe les vecteurs dont les concentrations de PM_{10} sont les plus élevés et dépassent en moyenne 60 ($\mu g /m^3$). Ce groupe est également caractérisé par un faible taux humidité, par conséquent ces situations provoquent une mauvaise qualité de l'air.

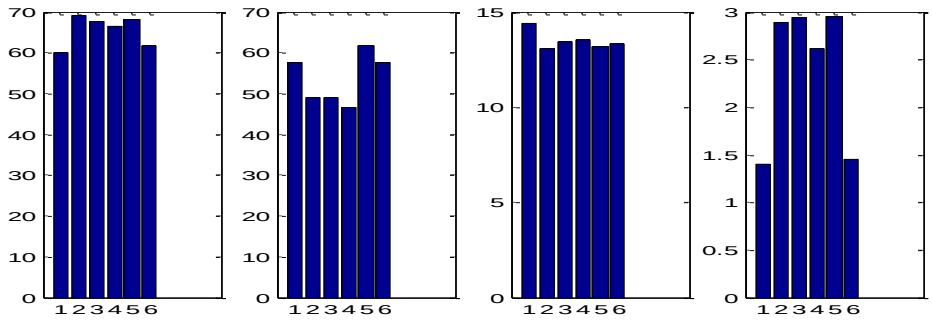


Figure 4. 9– les valeurs moyennes des paramètres pour chaque cluster.

4.5 Impact des clusters météorologiques sur les concentrations de polluants atmosphériques dans la région d'Annaba

L'air que nous respirons n'est jamais totalement pur. Si l'azote et l'oxygène représentent environ 99% de la composition totale de l'air, on trouve dans le 1% restant une grande variété de composés plus ou moins agressifs pour l'homme et son environnement (Samasafia, 2007). La qualité de l'air est affectée non seulement par l'émission des polluants mais également par les paramètres météorologiques. L'identification des sources de la pollution atmosphérique est une étape importante pour le développement des stratégies de contrôle de la qualité de l'air. Les stratégies de réduction peuvent améliorer de manière significative la qualité de l'air une fois que les sources sont identifiées (Gupta et al, 2008). Il est extrêmement important de considérer l'effet des conditions météorologiques sur la pollution atmosphérique, puisqu'elles influencent directement la capacité de dispersion de l'air polluant. Certains graves épisodes de pollution dans l'environnement urbain ne sont pas habituellement attribués aux augmentations soudaines de l'émission des polluants mais à certaines conditions météorologiques qui diminuent la capacité de l'atmosphère à disperser les polluants (Ziomas et al, 1995).

L'étude des relations entre les paramètres des bases de données volumineuses telles que la pollution atmosphérique et les paramètres météorologiques peut fournir des informations importantes concernant la nature des dépendances des données. Pendant les dernières décennies beaucoup d'efforts ont été consacrés à étudier les relations entre la pollution atmosphérique et les paramètres météorologiques (Kostas et al, 2007; Kermani et al, 2003). Par conséquent plusieurs méthodologies, déterministes et statistiques, ont été proposées. Ces méthodologies sont souvent basées sur les modèles de la régression linéaire ou non linéaire. Dans (Levy et al, 2006), les relations entre les concentrations des particules ultrafines (PM2.5) et des hydrocarbures aromatiques polycycliques d'une part et le volume de trafic, la direction du vent et la distance de la route d'autre part, ont été étudiés en utilisant les modèles de la régression linéaire à effets mixtes. Ainsi, la régression multiple a été utilisée par (Hien et al, 2002) pour étudier l'effet des paramètres météorologiques captés pendant l'hiver et les périodes d'été sur les concentrations des particules ultrafines PM2.5 et PM2.5-10. L'analyse en composantes principales (ACP) a été utilisée dans de nombreuses études telles que (Kostas et al, 2007; Statheropoulos, 1998) pour identifier les relations cachées entre le polluant examiné et les facteurs qui favorisent sa formation. D'autre part, les méthodes non linéaires ont été aussi largement utilisées pour la modélisation de la qualité de l'air (Kostas et al, 2007; Gardner et Dorling, 1999; Kolehmainen et al, 2001). Vu que les réseaux de neurones

artificiels (RNAs) sont capables de capturer les relations non linéaires qui existent entre les paramètres météorologiques et les niveaux de concentration des polluants, leurs performances ont été trouvés supérieures une fois comparée aux méthodes statistiques telles que la régression multiple (Gardner et Dorling, 1999; Kolehmainen et al, 2001).

Dans ces dernières années, l'utilisation de la classification automatique pour bien élucider les dépendances météo paramètres de pollution a été proliférée. L'intérêt accru pour ce procédé est attribué à son utilité à résoudre une grande partie de problèmes climatologiques, Le souci de comprendre les impacts de la météorologie, particulièrement comprendre les implications possibles des changements climatiques, a conduit la recherche pour plus, et meilleur approches de classification météorologiques. Cette technique a été utilisée avec sucées dans de nombreuses études tel que dans (Eder et al, 1994). La méthode de l'analyse en composante principale (ACP) et l'algorithme K-means ont été utilisés dans (Davis et al, 1998). Notre objectif à travers ce chapitre est d'étudier l'influence des paramètres météorologiques sur la pollution atmosphérique dans la région d'Annaba. Dans le but de bien éclaircir ces dépendances nous avons proposé d'utiliser les clusters météorologiques de la région.

4.5.1 Réseaux de neurones artificiels (RNA) pour identifier l'impact des paramètres météorologiques sur la pollution atmosphériques

En raison de la non-linéarité inhérente des concentrations des polluants et les interactions complexes entre les variables météorologiques et les polluants, le développement de modèles non linéaires, tels que les réseaux de neurones artificiels, est largement utilisée dans la littérature. Selon Kukkonen et al. (2003), le Perceptron Multicouches (PMC) est le type de réseaux de neurones le plus approprié dans la modélisation des problèmes de qualité de l'air. En effet, les modèles de PMC sont largement utilisés dans la modélisation et la prédiction des concentrations des polluants atmosphériques, car ils peuvent capturer efficacement les relations fortement non linéaire entre les variables, et leurs performances sont meilleure que d'autres modèles linéaires (Gardner et Dorling, 1999; Kolehmainen et al, 2001). En outre, les PMC peuvent être formés pour approximer toute fonction mesurable (hautement non-linéaire) sans aucunes hypothèses préalables concernant la distribution de données.

Dans notre approche nous utilisons les clusters météorologiques pour étudier l'influence des paramètres météorologique sur la pollution atmosphérique pour la région d'Annaba. Pour

chaque vecteur météorologique qui appartient à un cluster quelconque on associe les observations horaires correspondantes pour le polluant considéré. Il en résulte donc une nouvelle base de données composée de quatre variables (03 paramètres météorologiques et le polluant considéré). Par conséquent, pour chaque polluant dans chaque cluster météorologique on a créé une nouvelle base de données qui sera utilisée pour mesurer les dépendances entre le polluant considéré et les paramètres météorologiques. Le Tableau 4.5 présente le nombre des lignes de données utilisés pour la création des modèles neuronaux pour chaque polluant. Selon le tableau 4.5, le nombre de lignes de données utilisés pour la création des modèles neuronaux pour O_3 et PM_{10} dans le premier cluster (C1) est insuffisant. Sachant que les vecteurs météorologiques de C1 sont très proches de ceux du cluster C5, on a décidé de fusionner le cluster C1 et le cluster C5 en un seul cluster (C1+C5) qui contient à la fois les vecteurs météorologiques de C1 et C5. La figure 4.10 présente la distribution des nouveaux clusters après fusionnement des clusters C1et C5.

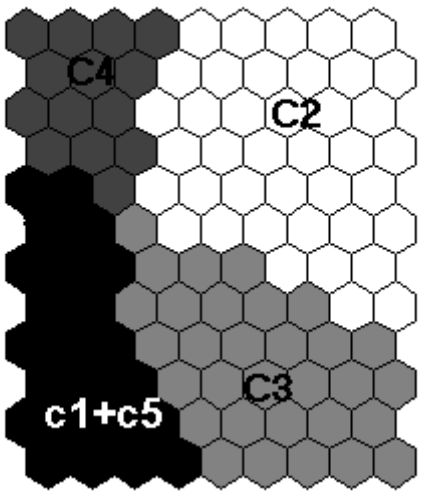


Figure 4. 10--Fusionnement des clusters C1 et C5 en (C1+C5).

L'approche proposée est basée sur la création d'un PMC pour chaque polluant dans chaque cluster météorologique de la région d'Annaba. L'architecture proposée pour chaque modèle neuronal est le résultat de toute une phase de comparaison et de test des déférentes topologies possibles. Cette architecture est décrite en trois couches, une couche pour l'entrée du réseau, unique et commune à tous les neurones, recevant les trois paramètres météorologiques, une couche pour les sorties, et une couche cachée. Les neurones dans la couche d'entrée reçoivent trois signaux d'entrée représentant la température ambiante, l'humidité relative et la vitesse de vent; par conséquent trois neurones ont été utilisés pour l'entrée du PMC. D'autre part, la

couche de sortie est composée d'un seul neurone qui représente à chaque fois le niveau de concentration du polluant étudié. Il n'y a aucune règle directe et précise permettant de déterminer le nombre de couches cachées à utiliser ni le nombre exact de neurones à inclure dans chaque couche cachée.

Tableau 4. 5– Le nombre de lignes de données utilisés pour l'apprentissage et la validation dans chaque cluster météorologique.

	CO	***NO***	***NO_x***	***NO_2***	***O_3***	***PM_{10}***	***SO_2***
C1	672	672	672	672	24	144	672
C2	844	2338	2338	2338	3141	4651	1485
C3	982	2510	2510	2510	2275	3298	1067
C4	824	1270	1270	1270	552	1608	1055
C5	1243	1612	1612	1612	335	802	1245
C1+C5	1915	2284	2284	2284	359	946	1917

Le nombre de neurones dans la couche cachée de chaque modèle neuronal a été estimé en utilisant une approche d'optimisation topologique ou le réseau est initialement construit avec un nombre de neurones réduit, puis le réseau est examiné pour évaluer ses performances en faisant apprendre le PMC par l'ensemble de données relative au polluant étudié (80% de données pour l'apprentissage et 20% pour la validation) tout en modifiant à chaque fois le nombre de neurone dans la couche cachée. Pour chaque PMC résultant d'un nombre de neurone dans la couche cachée, l'erreur de validation et l'index de l'accord (IA) ont été calculés. On retient donc le nombre de neurones pour le PMC dont l'erreur de validation est minimum et l'index de l'accord est maximum. Le tableau 4.6 présente les performances des différents PMC obtenus en modifiant le nombre neurones dans la couche cachée pour modéliser l'ozone dans le deuxième cluster météorologique. Selon ce tableau, le PMC le plus adéquat est obtenu pour un nombre de neurone dans la couche cachée égale à quinze.

Tableau 4. 6--Les performances des différents PMC, selon le nombre de neurone.

Neurones	*Nombre d'époch*	*Erreur de Validation*	*IA*
09	350	0.632	0.786
11	350	0.643	0.750
13	350	0.650	0.711
14	350	0.651	0.692
15	350	0.639	0.802
16	350	0.639	0.761
18	350	0.626	0.767
20	350	0.628	0.757

4.5.2 Apprentissage et résultats

La période considérée (2 ans) pour la collection de données est un peu limitée, et le manque d'horaire continu de données nous a conduits à utiliser la validation croisée. L'ensemble de données est divisé en cinq sous-ensembles. Il en résulte cinq modèles à identifier en utilisant cinq sous-ensembles de données. A chaque fois un nouveau sous ensemble différent est sélectionné pour l'opération de validation, le reste des sous-ensembles, concaténés, est utilisé pour l'apprentissage. Afin de résoudre le problème de sur apprentissage, nous avons implémenté une méthode d'apprentissage supervisé basée sur l'erreur de validation. Cette méthode vérifie l'erreur de validation à chaque itération de l'apprentissage et la compare à sa valeur précédente pour déterminer le moment ou cette valeur a commencé de s'accroitre. A ce moment, le réseau est sauvegardé. Pour être sûre qu'il s'agit d'un minimum global (ou au moins un minimum local décent), l'apprentissage n'est pas interrompu et devrait être effectuée pendant un nombre d'itération suffisent pour être sure que l'erreur de validation ne décroitra pas. Les indicateurs statistiques sont alors calculés sur la base de chaque modèle pour fournir une description numérique et évaluer la qualité du modèle neuronal obtenu. A la fin de la validation croisée cinq modèles neuronaux sont obtenus. On peut alors choisir le meilleur PMC ou le plus approprié. Les résultats obtenus sont représentés au tableau 4.7, Les indicateurs statistiques décrits dans l'annexe sont utilisés pour l'évaluation des performances

de chaque modèle neuronal obtenu. Les valeurs dans ce tableau sont pour l'algorithme d'apprentissage le plus approprié et pour l'architecture à une couche cachée donnant la structure la plus appropriée pour le problème étudié. Les performances de chaque modèle neuronal ont été évaluées à l'aide de l'indice de l'accord (IA) (Les indicateurs numériques de performance d'un modèle neuronal sont discutés dans l'annexe).

Tableau 4. 7--Indicateurs statistiques pour les modèles neuronaux obtenus pour chaque polluant dans chaque cluster (de C2 jusqu'à (C1+C5)).

		CO	***NO***	***NO_x***	***NO_2***	***O_3***	***PM_{10}***	***SO_2***
C2	***IA***	0.75	0.68	0.62	0.74	0.85	0.77	0.80
	RMSE	0.13	0.04	0.05	0.05	0.11	0.03	0.07
	MAE	0.06	0.01	0.03	0.04	0.09	0.02	0.03
	R	0.62	0.58	0.50	0.61	0.76	0.67	0.70
C3	***IA***	0.69	0.84	0.76	0.77	0.76	0.50	0.62
	RMSE	0.10	0.02	0.04	0.08	0.10	0.09	0.24
	MAE	0.05	0.01	0.02	0.05	0.06	0.06	0.16
	R	0.54	0.74	0.66	0.66	0.62	0.39	0.43
C4	***IA***	0.79	0.68	0.65	0.60	0.85	0.69	0.79
	RMSE	0.06	0.06	0.07	0.13	0.13	0.08	0.10
	MAE	0.03	0.02	0.04	0.08	0.10	0.05	0.08
	R	0.68	0.51	0.54	0.45	0.74	0.57	0.66
C1+C5	***IA***	0.63	0.76	0.67	0.66	0.89	0.85	0.58
	RMSE	0.08	0.04	0.06	0.05	0.16	0.06	0.12
	MAE	0.03	0.02	0.04	0.03	0.11	0.04	0.09
	R	0.45	0.65	0.55	0.54	0.80	0.74	0.37

L'intervalle des valeurs de l'indice IA pour l'ensemble des polluants dans le deuxième cluster météorologique est de 62% pour NOx à 85% pour l'ozone, avec une valeur moyenne de 74% pour l'ensemble des polluants. L'intervalle de valeurs de l'indice IA pour le troisième et le quatrième cluster est respectivement 50,1% à 84,2% pour le troisième et 60,3% à 84,6% pour

le quatrième cluster météorologique. Les mesures de performances de l'IA pour le cluster (C1+C5) varient de 58% pour le SO_2 à 89% pour l'ozone. Il est donc clair que les résultats de l'indice (IA) sont très satisfaisants pour l'ensemble des polluants considérés pour chaque cluster météorologique, car il atteint en moyen 72.25%. Ces résultats montrent également que les niveaux de concentration des polluants sont influencés par les paramètres météorologiques. Un autre indicateur statistique très utilisé pour évaluer les performances des modèles neuronaux est l'erreur RMSE. Les valeurs de l'RMSE indiquent un meilleur ajustement si elles sont plus proches de zéro. L'intervalle des valeurs de l'RMSE pour le deuxième cluster météorologique est de 0.03 (g/m^3) pour les PM_{10} à 0.13 (mg/m^3) pour le CO. En effet les valeurs de l'erreur RMSE sont proches pour l'ensemble des polluants dans tous les clusters. Ces résultats sont très satisfaisants et montrent également l'efficacité de l'approche considérée.

Tableau 4. 8--Les indicateurs statistiques moyens pour tous les polluants dans chaque cluster.

	C 2	***C 3***	***C 4***	***C1+ C5***
IA	0,742	0,707	0,721	0.72
RMSE	0,069	0,096	0,089	0.08
MAE	0,038	0,059	0,055	0.05
R	0,634	0,578	0,593	0.59

Tableau 4. 9--Les valeurs moyennes des indicateurs statistiques pour chaque polluant.

	CO	***NO***	***NO_x***	***NO_2***	***O_3***	***PM_{10}***	***SO_2***
IA	0,72	0,74	0,68	0,69	0,84	0,70	0,70
RMSE	0,09	0,04	0,06	0,08	0,13	0,07	0,13
MAE	0,04	0,02	0,03	0,05	0,09	0,04	0,09
R	0,57	0,62	0,56	0,57	0,73	0,59	0,54

Le pouvoir de généralisation des modèles neuronaux a été également testé par le coefficient de corrélation. En se basant sur les résultats obtenus, les coefficients de corrélation moyens pour tous les polluants dans chaque cluster qui sont représentés par le tableau 4.8 montrent

également que tous les clusters sont proches en termes de corrélation moyenne dont l'intervalle des mesures varie de 0,578 pour le troisième cluster à 0,634 pour le deuxième cluster météorologique. Selon le tableau 4.9 qui montre également les valeurs moyennes des indicateurs statistiques pour chaque polluant, l'ozone (O_3) est le polluant le plus corrélé avec les paramètres météorologiques. Les résultats obtenus (en terme de corrélation) prouvent qu'approximativement 60% de la variation des variables dépendantes (concentration des polluants) peuvent être expliquée par les variables indépendantes (paramètres météorologiques). D'une manière générale, les résultats de l'indicateurs MAE montrent également que l'approche proposée est très utile pour étudier l'influence des paramètres météorologique sur la pollution atmosphérique en se basant sur les clusters météorologiques.

4.6 Conclusion

Dans ce chapitre, deux approches de classification non supervisée à deux niveaux ont été développées afin de visualiser et d'identifier les types de jours météorologiques de la région de Annaba. Ainsi, les conclusions suivantes peuvent être tirées:

- L'approche de classification non supervisée à deux niveaux utilisant SOM et K-means semble être plus efficace qu'une approche de clustering directe utilisant uniquement SOM ou K-means selon les résultats obtenus.
- Cinq classes météorologiques différentes ont été identifiées avec des frontières claires, et par conséquent, les paramètres météorologiques de chaque groupe peuvent être facilement interprétés.
- Le nombre approprié de clusters et la qualité des résultats de clustering ont été validés à l'aide des indices de validité de clustering.
- Les clusters météorologiques obtenus ont été utilisés pour étudier l'influence des paramètres météorologiques sur la pollution atmosphérique pour la région d'Annaba.

Bien que la période considérée (2 ans) pour la collection de données est un peu limitée, et le manque d'horaire continu, il s'avère qu'un réseau de neurone artificiel avec une seule couche cachée basée sur l'algorithme de rétro propagation du gradient est très efficace pour modéliser le comportement non linéaire des émissions des polluants dans chaque cluster météorologique. Selon les indicateurs statistiques utilisés pour évaluer les performances des modèles neuronaux obtenus, les résultats obtenus sont très satisfaisants.

Conclusion générale

Conclusion générale

Dans ce travail nous nous sommes intéressés aux méthodes de classification non supervisée pour l'identification des types de jours météorologiques pour la région d'Annaba. En raison des quantités de données fournies par la station Samasafia, des outils efficaces d'analyse sont indispensables pour extraire les caractéristiques utiles, fournissant des informations plus simples et plus efficaces. L'algorithme SOM a prouvé son utilité pour la classification des bases de données multidimensionnelles traitant des problèmes non linéaire. Cependant, l'identification des frontières de chaque cluster est une tâche très difficile dans le cas des données complexes. La classification à deux niveaux de a été utilisée dans le cadre de cette étude dans deux approches différentes basées sur l'utilisation des cartes de kohonen. La première approche consiste à utiliser la carte auto-organisatrice de Kohonen (SOM) pour le premier niveau et la classification par partition (K-means) dans le deuxième niveau. Cette approche a été également utilisée pour l'identification des types de jours météorologiques pour la région d'Annaba à partir des données météorologiques captées par la station Samasafia d'Annaba durant la période 2003 à 2004. La deuxième approche de classification à deux niveaux consiste à utiliser une carte CHAOSOM à deux dimensions dans le premier niveau pour l'apprentissage des données et dans le deuxième niveau une carte CHAOSOM unidimensionnel est utilisée pour regrouper les vecteurs prototypes générés par le premier niveau. Le nombre optimal de clusters a été sélectionné en utilisant deux catégories de critères de validation de la classification automatique (critères interne et externe). Ainsi, cinq clusters météorologiques différents (C1-C5) ont été détectés avec des frontières claires, donc les paramètres météorologiques de chaque cluster peuvent être facilement interprétés.

En vue d'étudier l'influence des paramètres météorologique sur la pollution atmosphérique au niveau de la région d'Annaba. Les clusters météorologiques obtenus on utilisant la base de données collectée par la station Samasafia ont été utilisé. Plusieurs rapports linéaires entre les trois paramètres météorologiques et les polluants considérés ont été identifiés et plusieurs significations et conclusions ont été tirées. Ainsi, il a été remarqué que les paramètres météorologiques étudiés sont non linéairement corrélés avec les niveaux de concentration des polluants considérés. Pour modéliser les relations non linéaires qui existent entre les paramètres météorologiques et les niveaux de concentration des polluants au niveau de chaque cluster météorologique, nous avons utilisé un modèle neuronal (PMC). Selon les indicateurs statistiques utilisés pour évaluer les performances des modèles neuronaux obtenus, les résultats sont très satisfaisants. Les résultats obtenus peuvent être utilisés dans des travaux

futurs pour concevoir un système de prédiction des paramètres atmosphériques pour chaque classe météorologique, qui peut améliorer les résultats d'un système global unique pour tous les modèles de la base de données. Ainsi, il serait intéressent d'utiliser les différentes topologies des cartes de kohonen (rectangulaire, hexagonal, toroïdal, etc…) afin d'examiner l'applicabilité de la méthode et son topologie dans le domaine de la classification des paramètres météorologiques. Il serait également intéressant d'utiliser les résultats obtenus pour faire une étude épidémiologiques permettant de faire une corrélation entre les niveaux de pollution atteints et les maladies respiratoires.

Références

Aguilera, P. A., Frenich, A. G., Torres, J. A., Castro, H., Vidal, J. L., & Canton, M. (2001). Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. Water Research, 35(17), 4053-4062.

Aihara, K., Takabe, T., & Toyoda, M. (1990). Chaotic neural networks. *Physics letters A*, *144*(6), 333-340.

Alioua, A., Maizi, N., Maizi, L., & Tahar, A. (2008). Caractérisation de la pollution par le NO2 à l'aide d'un couplage de technique biologique et physico-chimique dans la région d'Annaba (Algérie). Pollution atmosphérique, 50(200), 325-332.

Arabie, P., Hubert, L. J., & De Soete, G. (Eds.). (1996). Clustering and classification. World Scientific., pages 5-63.

Berkhin P. (2002). Survey of clustering data mining techniques". Technical report, Accrue Software, San Jose, CA.

Blaise K. Y., Théophile, L. Alain A. P., Johannet A., (2007). Optimization of Multi-Layers Perceptrons Models With Algorithms of First and Second Order. "Application to The Modeling of Rainfall-Rainoff Relation in Bandamma Blanc Catchment (North Of Ivory Coast)", *European Journal of Scientific Research*, 17(3), 313-328, 2007.

Blansché, A. (2006). Classification non supervisée avec pondération d'attributs par des méthodes évolutionnaires (Thèse de doctorat). Univ. Louis Pasteur.

Boinee, P. (2006). *Insights into machine learning: data clustering and classification algorithms for astrophysical experiments* (Doctoral dissertation, Ph. D thesis, Dept of Maths and Computer Science, 2006, Univ of Udine–Italy).

Boubou, M. (2007). *Contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions* (Doctoral dissertation, Université Claude Bernard-Lyon I).

Bradley, P. S., & Fayyad, U. M. (1998, July). Refining Initial Points for K-Means Clustering. In *ICML* (Vol. 98, pp. 91-99).

Bradley, P. S., Fayyad, U. M., & Reina, C. (1998, August). Scaling Clustering Algorithms to Large Databases. In *KDD* (pp. 9-15).

Bridgman, H. A., Davies, T. D., Jickells, T., Hunova, I., Tovey, K., Bridges, K., & Surapipith, V. (2002). Air pollution in the Krusne Hory region, Czech Republic during the 1990s. *Atmospheric Environment*, *36*(21), 3375-3389.

Candillier, L. (2006). *Contextualisation, visualisation et évaluation en apprentissage non supervisé* (Doctoral dissertation, Université Charles de Gaulle-Lille III).

Chaffai H., & Mourdi, W. (2011). Etat de la pollution atmosphérique dans la région d'Annaba et son impact sur l'eau et l'environnement, ScienceLib Editions Mersenne : Vol.3, N ° 11080.

Chaloulakou, A., Kassomenos, P., Spyrellis, N., Demokritou, P., & Koutrakis, P. (2003). Measurements of PM10 and PM2.5 particule concentrations in Athens, Greece. *Atmospheric Environment*, *37*(5), 649-660.

Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S., & Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica*, *12*(1), 241-262.

Chessel D., Thioulouse J. & A.B. Dufour, "introduction à la classification hiérarchique", Rapport technique, 2004.

Comrie, A. C., & Diem, J. E. (1999). Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona. Atmospheric Environment, 33(30), 5023-5036.

Dave, R. N., (1996). Validating Fuzzy Partitions Obtained Throughc-Shells Clustering", *Pattern Recognition Letters*, vol. 17, pp. 613–623.

Davis, J. M., Eder, B. K., Nychka, D., & Yang, Q. (1998). Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. *Atmospheric Environment*, *32*(14-15), 2505-2520.

Dean., E. J. (2010) computer aided identification of biological specimens using self organizing maps, Magister science, univ pretoria, south africa.

Derradji, F., Kherici, N., Djorfi, S., Romeo, M., & Caruba, R. (2005). Etude de l'influence de la pollution de l'oued Seybouse sur l'aquifère d'Annaba (Algérie Nord-orientale) par le chrome et le cuivre. *Houille blanche*, (1), 73-80.

Dimitriadou, E., Dolničar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, *67*(1), 137-159.

Dreyfus, G., Martinez, J. M., Samuelides, M., Gordon, M. B., Badran, F., Thiria, S., & Hérault, L. (2004). Réseaux de neurones: méthodologies et applications. édition EYROLLES Avril 2004.

Eder, B. K., Davis, J. M., & Bloomfield, P. (1994). An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. Journal of Applied Meteorology, 33(10), 1182-1199.

El Golli A. (2004). Conan-Guez Brieuc, and Fabrice Rossi, "A Self Organizing Map for dissimilarity data", In proc of IFCS'.

Elminir, H., Abdel-Galil, H., (2006). Estimation of air pollutant concentrations from meteorological parameters using artificial neural network. Electr. Eng. 57 (2), 105–110.

Engelbrecht, A. P. (2007). Computational intelligence: an introduction. Wiley. com.

Ganti, V., Gehrke, J., & Ramakrishnan, R. (1999, August). CACTUS—clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 73-83).

Gardner, M. W., & Dorling, S. R. (1999). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Environment*, *33*(5), 709-719.

Grabmeier, J., & Rudolph, A. (2002). Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, *6*(4), 303-360.

Green, D. G., & Klomp, N. I. (1998). Environmental informatics-a new paradigm for coping with complexity in nature.

Guérif, S. (2006). Réduction de dimension en apprentissage numérique non supervisé. *Doctoral dissertation, Université Paris*, *13*.

Gupta, A. K., Karar, K., Ayoob, S., & John, K. (2008). Spatio-temporal characteristics of gaseous and particulate pollutants in an urban region of Kolkata, India. Atmospheric Research, 87(2), 103-115.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2-3), 107-145.

Halkidi, M., Vazirgiannis, M., & Batistakis, Y. (2000). Quality scheme assessment in the clustering process. In *Principles of Data Mining and Knowledge Discovery* (pp. 265-276). Springer Berlin Heidelberg.

Hargreaves, P. R., Leidi, A., Grubb, H. J., Howe, M. T., & Mugglestone, M. A. (2000). Local and seasonal variations in atmospheric nitrogen dioxide levels at Rothamsted, UK, and relationships with meteorological conditions. *Atmospheric Environment*, *34*(6), 843-853.

Hasan, M. A., Chaoji, V., Salem, S., & Zaki, M. J. (2009). Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, *30*(11), 994-1002.

Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., ... & Kallioniemi, O. P. (2003). Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps. *Machine Learning*, *52*(1-2), 45-66.

Hewitson, B. C., & Crane, R. G. (2002). Self-organizing maps: applications to synoptic climatology. Climate Research, 22(1), 13-26.

Hien, P. D., Bac, V. T., Tham, H. C., Nhan, D. D., & Vinh, L. D. (2002). Influence of meteorological conditions on PM2.5 and PM2.5−10 concentrations during the monsoon season in Hanoi, Vietnam. *Atmospheric Environment*, *36*(21), 3473-3484.

Hilty. Lorenz M., Page Bernd, Hrebıcek J. (2006). Environmental informatics, Environmental Modelling & Software vol. 21 pp : 1517--1518.

Himberg, J. (2000). A SOM based cluster visualization and its application for false coloring. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks* (Vol. 3, pp. 587-592).

Hodgkin, A.L., Huxley, A.F.: A quantitative de-scription of membrane current and its application to con-duction and excitation in nerve. Journal of Physiology 117 (1952) 500–544.

Huang, G. H., & Chang, N. B. (2003). The perspectives of environmental informatics and systems analysis. *Journal of Environmental Informatics*, *1*(1), 1-7.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient K-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(7), 881-892.

Karatzas, K. D., & Kaltsatos, S. (2007). Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. *Simulation Modelling Practice and Theory*, *15*(10), 1310-1319.

Kaski, S. (1997). Data exploration using self-organizing maps. In *ACTA POLYTECHNICA SCANDINAVICA: MATHEMATICS, COMPUTING AND MANAGEMENT IN ENGINEERING SERIES NO. 82*.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. *Neural computing surveys*, *1*(3&4), 1-176.

Kaufman L. and P.J. Rousseeuw. (1990). Finding groups in data. *John Wiley and Sons*, Inc., 1990.

Kaufman, L., & Rousseeuw, P. J. (1990). Clustering Large Applications (Program CLARA). *Finding Groups in Data: An Introduction to Cluster Analysis*, 126-163.

Kermani, M., Naddafi, K., Shariat, M., & Mesbah, A. S. (2003). Chemical composition of TSP and PM10 and their relations with meteorological parameters in the ambient air of Shariati Hospital District. *Iranian J Publ Health*, *32*(4), 68-72.

Khadir, M. T., Fay, D., & Boughrira, A. (2006). Day type identification for Algerian electricity load using Kohonen maps. *Transaction on Engineering, Computing and Technology*, *15*, 296-301.

Khadir, M.T., Khedairia, S., Benabbas, F., (2010). Kohonen Maps Combined to K-means in a Two Level Strategy for Time Series Clustering Application to Meteorological and Electricity Load data, self organizing maps, INTECH.

Khedairia S., Khadir M. T. (2008c). Comparison of Clustering Methods in a Two Stage Meteorological Day Type Identification Approach for the Region of Annaba -Algeria-, 2nd International Conference on Electrical Engineering design and Technologies, Hammamet-Tunisia, 8-10 Nov. 2008.

Khedairia, S., & Khadir, M. T. (2012). Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria. Atmospheric Research, 113, 89-101.

Khedairia, S., Khadir, M.T., (2008a). L'analyse en Composantes Principales (ACP) et Carte Auto-organisatrice de Kohonen pour L'identification des types de jours météorologiques de la région d'Annaba. Proc of the 10th Maghrebian Conference on Information Technologies (MCSEAI'08), Oran, Alegria, pp. 18–23.

Khedairia, S., Khadir, M.T., (2008b). Self-Organizing Map and K-Means for Meteorological Day Type Identification for the Region of Annaba -Algeria-. Proceedings of the IEEE Conference on computer information systems and industrial management applications (CISIM), Ostrava, The Czech Republic, pp. 91–96.

Khodja L. (1997). Contribution à la Classification Floue non Supervisée, thèse de doctorat. Univ de Savoie, 1997.

Kiang, M. Y., Hu, M. Y., & Fisher, D. M. (2006). An extended self-organizing map network for market segmentation—a telecommunication example. Decision Support Systems, 42(1), 36-47.

Kohonen, T.: Self-organising maps. In: IEEE. Volume 78. (1990) 1464–1480

Kolehmainen M., H. Martikainen, J. Ruuskanen, "Neural networks and periodic components used in air quality forecasting", Atmospheric Environment, vol. 35, pp. 815–825, 2001.

Kolehmainen, M. T. (2004). *Data exploration with self-organizing maps in environmental informatics and bioinformatics*. Helsinki University of Technology.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., & Cawley, G. (2003). Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, *37*(32), 4539-4550.

Kuzmanovski, I., Trpkovska, M., & Šoptrajanov, B. (2005). Optimization of supervised self-organizing maps with genetic algorithms for classification of urinary calculi. Journal of molecular structure, 744, 833-838.

LABROCHE N. (2012), Méthodes d'apprentissage automatique pour l'analyse des interactions utilisateurs, thèse de doctorat, Université Pierre et Marie Curie.

Laitinen, N., Rantanen, J., Laine, S., Antikainen, O., Räsänen, E., Airaksinen, S., & Yliruusi, J. (2002). Visualization of particle size and shape distributions using self-organizing maps. *Chemometrics and intelligent laboratory systems*, *62*(1), 47-60.

Lallahem, S. (2002). Structure Et Modélisation Hydrodynamique des Eaux Souterraines : Application à l'aquifère Crayeux de la Bordure Nord du Bassin De Paris, thèse de doctorat, université de Lille.

Lemaire V., (2006). Cartes auto-organisatrices pour l'analyse de données, proc. COnférence en Recherche d'Information et Application (CORIA).

Levy, J. I., Bennett, D. H., Melly, S. J., & Spengler, J. D. (2003). Influence of traffic patterns on particulate matter and polycyclic aromatic hydrocarbon concentrations in Roxbury, Massachusetts. *Journal of Exposure Science and Environmental Epidemiology*, *13*(5), 364-371.

Liu, Y., & Weisberg, R. H. (2011). A review of Self-Organizing Map applications in meteorology and oceanography. *Self-Organizing Maps-Applications and Novel Algorithm*, 253-272.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010, December). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 911-916). IEEE.

Ludwig, B., Boiffin, J., & Auzet, A. V. (1995). Hydrological structure and erosion damage caused by concentrated flow in cultivated catchments. *Catena*, *25*(1), 227-252.

Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). Neural Networks, IEEE Transactions on, 7(1), 16-29.

Matsushita, H., Nishio, Y.: Chaosom; collaboration between chaos and self-organizing map. In: Proceedings of RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'06). (2006) 305–308.

Mebirouk Hayet, Mebirouk-Bendir Fatiha, (2007). Principaux acteurs de la pollution dans l'agglomération de annaba. Effets et développements", Colloque International sur l'Eau et l'Environnement, alger.

Ng, R. T. and Han, J. 5 (1994) Efficient and effective clustering methods for spatial data mining", In bocca, J., Jarke, M., and Zaniolo, C., editors, 20th international conference on very large data bases, pages 144-155.

Noui N., Boukhemis., K., (2011). Annaba face aux risques urbains et technologiques : Quel avenir ? (Cas de Seybouse), In Proc. Intervention sur les Tissus Existants pour une Ville Durable.

Oja, M., Kaski, S., & Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. *Neural computing surveys*, *3*(1), 1-156.

Oleg M. Pokrovsky, Roger H. F. Kwok, C. N. Ng, "Fuzzy logic approach for description of meteorological impacts on urban air pollution species: a Hong Kong case study", *Computers & Geosciences*, vol. 28, pp. 119-127, 2002.

Pal, N.R., Pal, S., (2002), Computational Intelligence for Pattern Recognition, International Journal of Pattern Recognition and Artificial intelligence, vol 16(7), pp:773--779.

Park, Y. S., & Chung, Y. J. (2006). Hazard rating of pine trees from a forest insect pest using artificial neural networks. Forest ecology and management, 222(1), 222-233.

Pölzlbauer, G. (2004). Application of self-organizing maps to a political dataset. *Master Thesis, Vienna University of Technology*.

Recknagel, F. (Ed.), (2002). Ecological informatics: understanding ecology by biologically-inspired computation". *Springer*, Berlin398.

Reljin, I. S., Reljin, B. D., & Jovanović, G. (2003). Clustering and mapping spatial-temporal datasets using SOM neural networks. *Journal of Automatic Control*, *13*(1), 55-60.

Réseau de surveillance de la qualité de l'air, Bilan annuel sur la qualité de l'air pour l'année 2004, « Samasafia», 59p.

Réseau de surveillance de la qualité de l'air, Bilan annuel sur la qualité de l'air pour l'année 2006, « Samasafia», 72p.

Réseau de surveillance de la qualité de l'air, Bilan annuel sur la qualité de l'air pour l'année 2007, « Samasafia», 62p.

Reusch, D. B., Alley, R. B., & Hewitson, B. C. (2007). North Atlantic climate variability from a self-organizing map perspective. *Journal of Geophysical Research: Atmospheres (1984–2012)*, *112*(D2).

Rousset P. (1999). Applications des algorithmes d'auto-organisation à la classification et à la prévision", thése de doctorat, Univ. Paris I.

Shamir R, Sharan R. (2002). Algorithmic approaches to clustering gene expression data**.** In *Current Topics in Computational Biology*. Edited by Tao J, Ying X, Michael QZ. MIT Press.

Sheridan, S. C. (2002). The redevelopment of a weather-type classification scheme for North America. *International Journal of Climatology*, *22*(1), 51-68.

Simula, O., Vesanto, J., Alhoniemi, E., & Hollmén, J. (1999). Analysis and modeling of complex systems using the self-organizing map. *Neuro-Fuzzy Techniques for Intelligent Information Systems*, 3-22.

Singh G., Kumar V. (2013). An Efficient Clustering and Distance Based Approach for Outlier Detection, International Journal of Computer Trends and Technology (IJCTT) – Vol.4, 2067-2072.

Singh, S. S., & Chauhan, N. C. (2011, May). K-means v/s K-medoids: A Comparative Study. In *National Conference on Recent Trends in Engineering & Technology*.

Statheropoulos, M., Vassiliadis, N., & Pappa, A. (1998). Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment*, *32*(6), 1087-1095.

Strehl A.. (2002). Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining. Ph.D thesis, The University of Texas at Austin, May.

Suwardi A., Takenori K. & Shuhei K. (2007). Principal Component Analysis and Self Organizing Map for Visualizing and Classifying Fire Risks in Forest regions, agricultural information research" *Atmospheric Environment*, 16(2), 44-51.

Takahashi G., Suzuki. T., Kawamura H. (2011). Detection of Outliers in Meteorological Observation Data, Journal of Quality, 18(5), 393—405.

Talbi I., Système Dynamique non Linéaires et Phénomènes de Chaos, Mémoire de magister. Univ Constantine, 2010.

Tison, J., Park, Y. S., Coste, M., Wasson, J. G., Ector, L., Rimet, F., & Delmas, F. (2005). Typology of diatom communities and the influence of hydro-ecoregions: a study on the French hydrosystem scale. *Water Research*, *39*(14), 3177-3188.

Tombros, A. (2002). The effectiveness of query based hierarchic clustering of documents for information retrieval (Doctoral dissertation, University of Glasgow).

Tryba, V., Metzen, S., & Goser K. (1989). Designing basic integrated circuits by self-organizing feature maps", *In Neuro-Nimes '89. International Workshop. Neural Net-works and their Applications*, 225- 235.

Turalıoğlu, F., Nuhoğlu, A., & Bayraktar, H. (2005). Impacts of some meteorological parameters on SO2and TSP concentrations in Erzurum, Turkey. *Chemosphere*, *59*(11), 1633-1642.

Turias, I. J., Gonzalez, F. J., Martín, M. L., & Galindo, P. L. (2006). A competitive neural network approach for meteorological situation clustering. *Atmospheric Environment*, *40*(3), 532-541.

Ultsch A. and H. Siemon. (1990). Kohonen's self-organizing feature maps for exploratory data analysis. *In Proc. INNC'90, Int. Neural Network Conf.*, pages 305-308, Dordrecht, Netherlands.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent data analysis*, *3*(2), 111-126.

Vesanto, J. (2000, April). Neural network tool for data mining: SOM toolbox. In *Proceedings of symposium on tool environments and development methods for intelligent systems (TOOLMET2000)* (pp. 184-196).

Vesanto, J. (2002). *Data exploration process based on the self-organizing map*. Helsinki University of Technology.

Vesanto, J., & Ahola, J. (1999, June). Hunting for correlations in data using the self-organizing map. In *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99)* (pp. 279-285).

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on*, *11*(3), 586-600.

Wang, W., & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy sets and systems*, *158*(19), 2095-2117.

Willmott, C.,Ackleson, S.,Davis, R., Fuddema, J., Klink, K., Legates, D.,O'Donnell, J., and Rowe, C. (1985). Statistics for the evaluation and comparison of models, *Journal of Geophysical Research*, 90, 8995-9005.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association*, *77*(380), 841-847.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, *16*(3), 645-678.

Yamada, T., Aihara, K., & Kotani, M. (1993, October). Chaotic neural networks and the traveling salesman problem. In *Neural Networks, 1993. IJCNN'93-Nagoya. Proceedings of 1993 International Joint Conference on* (Vol. 2, pp. 1549-1552). IEEE.

Yu, T. Y., & Chang, L. F. W. (2001). Delineation of air-quality basins utilizing multivariate statistical methods in Taiwan. *Atmospheric Environment*, *35*(18), 3155-3166.

Ziomas, I., Melas, D., Zerefos, C., Bais, A., (1995). Forecasting peak pollutant levels from meteorological variables. Atmos. Environ. 29, 3703–3711.

Annexe

Evaluation de la qualité d'un modèle neuronal

L'évaluation de la qualité d'un modèle est une partie importante dans le processus de développement d'un modèle neuronal. Cette évaluation peut être réalisée par des méthodes numériques ou visuelles. Les méthodes visuelles permettent d'obtenir une vue intuitive de la performance d'un modèle neuronal, tandis que les méthodes numériques fournissent une terre plus solide pour comparer et augmenter les performances des modèles d'une manière scientifique. Les méthodes visuelles incluent les graphes simples du série-temporelle (prévue-observé), les histogrammes observé-prévue et les figures de dispersion de données (observé-prévue) (Lallahem, 2002; Kolehmainen, 2004).

L'utilisation des indicateurs de performance pour évaluer et comparer les modèles neuronaux a été initialement discutée par willmot et al dans (Willmott et al, 1985). Ainsi, leurs recommandations ont été suivies par plusieurs chercheurs (Lallahem, 2002; Kolehmainen, 2004; Blaise et al, 2007). Les indicateurs numériques de performance sont discutés en détail ci-dessous.

- *L'erreur moyenne absolue* (en anglais :Mean Absolute Error(MAE)) :c'est l'indicateur le plus simple des mesures numériques d'évaluation de la qualité des modèles. c'est simplement la moyenne des erreurs absolues pris sur l'ensemble des données prévus. cet indicateur est calculé selon l'équation :

$$\frac{|p_1 - a_1| + \ldots |p_n - a_n|}{n}$$

  Où n est le nombre de points. L'avantage de MAE est qu'il est moins sensible que l'erreur carrée aux valeurs aberrantes.

- *la Racine Carrée de l'Erreur Quadratique Moyenne* (en anglais Root Mean Squared Error (RMSE)) est l'un des indicateurs les plus communs utilisés avec les réseaux de neurones. L'erreur RMSE est calculée selon l'équation :

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

L'erreur RMSE donne une indication quantitative sur l'erreur de simulation obtenue pendant la phase de modélisation. Cette erreur mesure la déviation de prévision et/ou de simulation de la valeur réelle mesurée. Les valeurs idéales pour RMSE et MAE sont 0.

- l'index de l'accord(en anglais : Index of Agreement (IA)) est une mesure relative limitée à l'intervalle [0,1]. cet indicateur représente une évaluation globale de l'accord entre les niveaux de concentrations modélisés et les concentrations réelles. Il est idéal pour faire des comparaisons entre les modèles, il est calculé selon l'équation :

$$IA = 1 - \frac{\sum_i |p_i - a_i|^2}{\sum_i (|p_i - \bar{a}| + |a_i - \bar{a}|)^2}$$

- Le coefficient de corrélation: le coefficient de corrélation noté *R* est tout simplement la racine carrée du coefficient de déterminant; son signe (±) donne le sens de la relation. Il est calculé selon la formule:

$$R = \frac{S_{PA}}{\sqrt{S_p S_A}}$$

$$\text{tel que : } S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1}$$

$$S_p = \frac{\sum_i (p_i - \bar{p})^2}{n-1}$$

$$S_A = \frac{\sum_i (a_i - \bar{a})}{n-1}$$

- L'erreur relative absolue (en anglais : Relative absolute error (RAE)), ce critère permet d'évaluer la prévision d'un modèle par rapport au modèle marche aléatoire.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \bar{a}| + \ldots + |a_n - \bar{a}|}$$

En général, le critère IA est la meilleure mesure opérationnelle, c.-à-d. s'il n'est pas bon alors il est peu probable que le modèle puisse être utilisé dans la pratique. Si les indicateurs numériques sont bons, le coefficient de corrélation devrait avoir une valeur élevée, d'autre part, une valeur élevée du coefficient de corrélation implique habituellement des valeurs basses pour RMSE et élevées pour IA, mais ne le garantit pas.