وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR UNIVERSITY -ANNABA-
UNIVERSITE BADJI MOKHTAR -ANNABA-

جامعة باجي مختار
- عنابة -

Faculté : Sciences de l'Ingénieur **Année 2008**

Département : Informatique

MEMOIRE

Présentation en vue de l'obtention du diplôme de magister

Identification des Types de Jours Météorologiques et des Dépendances Météo-Paramètres de Pollution par RNA: Application à la Région d'Annaba.

Option

Texte Parole et Image

Par

Soufiane KHEDAIRIA

DIRECTEUR DE MEMOIRE: M. T. KHADIR **Maitre de conférences** **Univ. Annaba**

DEVANT LE JURY

PRESIDENT: Pr. Mokhtar SELLAMI	**Professeur**	**Univ. Annaba**
EXAMINATEURS:		
Dr. Nadir FARAH	**Maitre de conférences**	**Univ. Annaba**
Dr. Labiba SOUICI	**Maitre de conférences**	**Univ. Annaba**
Dr. Hayet MAROUANI.	**Maitre de conférences**	**Univ. Annaba**

Remerciements

ملخص

من أجل تحديد أقسام الطقس اليومية لمنطقة عنابة (شمال شرقي الجزائر)،اقترحنا منهجية تصنيفية غير خاضعة للإشراف -تجميع-. هذه المنهجية تمكن من الوقوف على مجموعة الأ يام المتماثلة وذلك بتقسيمها إلى عدة أصناف و مجموعات . استنادا إلى بيانات تم جمعها من طرف محطة الأرصاد الجوية بمطار عنابة خلال الفترة الممتدة من 1995 إلى 1999. المنهجية المقترحة تقوم على مستويين للتصنيفات: خارطة التنظيم الذاتي nenohoK في المستوى الأول لإستخلاص النماذج الأولية ، التي يتم تجميعها بعد ذلك في الم ستوى الثاني . من أجل تحديد أنسب برنامج لتجميع وحدات خارطة التنظيم الذاتي ، أجرينا مقارنة على أساس الأداء و النتائج بين كل من البرامج التالية: snaem-k و MAP(around medoid gninoititraP)، والتصنيف الهرمي (طريقة وارد) . النتائج المتحصل عليها خضعت ل تحليل كمي على أساس مجموعتين من المعايير (الداخلية والخارجية) ونوعي للتحقق من هذه النتائج وتفسيرها. هذه المنهجية مكنت من استخلاص ست مجموعات تتصل اتصالا مباشرا بالأحوال الجوية النموذجية بالمنطقة.

المنهجية المقترحة للتصنيف على مستويين تم أيضا تقييمها باستخدام قاعدة البيانات التي تم جمعها من طرف محطة سماء صافية ، حيث تم استخدام أصناف الطقس التي تم الحصول عليها 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 المحصل عليها مرضية للغاية و عدة علاقات واستنتاجات تم استخلاصها.

Abstract

A two level approach has been proposed in order to perform a meteorological data classification analysis for the region of Annaba (north east of Algeria) based on data collected from 1995 to 1999. The main objectives of the study presented here have been the characterization of the meteorological conditions in the area using a two stage clustering procedure – first using the Self-Organizing Map (SOM) to produce the prototypes, which are then clustered in the second stage. In order to find out the most suitable algorithm to cluster the SOM units, we have proceeded to a performances comparison between candidate clustering algorithms such as: PAM (Partitioning Around Medoids), K-means, and hierarchical agglomerative clustering (ward's method). 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 six classes, which were related to typical meteorological conditions in the area. The two stage clustering approach was also evaluated by using a meteorological data collected by "SAMASAFIA" network, where the obtained meteorological clusters were used to study the impact of the meteorological parameters (by cluster) on air pollution in this region. A linear analysis based on PCA (principal component analysis), and nonlinear based on neural network model (multi-layer perceptron) were carried out. The results obtained are very satisfactory where several relations and conclusions have been extracted.

Keywords: meteorological day type identification, clustering, impact of meteorological factors to urban air pollution, neural networks.

Résumé

Dans le but de l'identification des types de jours météorologiques pour la région d'Annaba, Nous avons proposé une approche basée sur la classification non supervisée -clustering-. Cette approche permet d'identifier les groupes d'objets similaires (types de jours -clusters-) à partir des données météorologiques captées par la station de l'aéroport d'Annaba pendant la période 1995 à 1999. L'approche proposée est basée sur deux niveaux de classifications : la carte auto-organisatrice de Kohonen (SOM) pour le premier niveau et la classification par partition pour le deuxième niveau. Pour découvrir la méthode la plus appropriée à utiliser pour regrouper les unités de la SOM, nous avons procédé à une comparaison basée les performances des algorithmes de classification automatique les plus utilisés tel que : l'algorithme PAM (Partitioning Around Medoids), K-moyennes, et la classification hiérarchique (méthode de Ward). Une analyse quantitative basée sur deux catégories de critères (internes et externes) et qualitative sont utilisées pour valider et interpréter les résultats obtenues de la classification. Cette approche a permet d'extraire six classes qui sont liés directement aux conditions météorologiques typiques de la région. L'approche de classification à deux niveaux a été également évaluée en utilisant une base de données collectée par la station SAMASAFIA, dont 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 linéaire basée sur l'ACP (analyse en composantes principales), et non linéaire basée sur un model neuronal (perceptron multicouches) ont été effectuées. Les résultats obtenus sont très satisfaisants ou plusieurs relations et conclusions ont été tirées.

Mots clés : classification non supervisée, identification des types de jours météorologiques, influence des paramètres météorologiques sur la pollution atmosphérique.

Table des matières

Listes des figures.

Liste des tableaux

Introduction générale.

Introduction générale

La dégradation de la qualité de l'air a été longtemps perçue en termes de nuisances locales dont les impacts sur la santé humaine sont considérés comme prépondérants. Cette perception se justifie par le fait que les polluants atmosphériques exercent des effets directs et souvent tangibles au niveau local. L'industrie est l'élément moteur de dégradation de la qualité d'air dans la région d'annaba, cette industrialisation a assurément permis de répondre aux besoins des populations et du pays en produits sidérurgiques, engrais azotés, constructions ferroviaires et autres industries de transformations. A l'inverse, elle a suscité une urbanisation démesurée de la ville avec tous ses corollaires et une pollution de l'atmosphère et des sols, suivies de conséquences néfastes sur le biotope et la société. 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 façon significative la qualité de l'air une fois que les sources sont identifiées [1]. Il est extrêmement important de considérer l'effet des conditions météorologiques sur la pollution atmosphérique, puisqu'elles influencent directement les possibilités de dispersion de l'atmosphère. 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 de disperser les polluants [2]. Dans l'étude des conditions météorologiques il est très difficile de représenter les données en termes statistiquement indépendants car la pollution atmosphérique dépend de tous les paramètres et données météorologiques [3].

Une des taches 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 qui permet l'extraction d'un ensemble de prototypes (clusters) représentatifs des modèles météorologiques dans une région d'intérêt. Ceci revient donc à identifier les différents types de jours météorologiques qui caractérisent la région. 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 pour comprendre les implications possibles des changements climatiques, a conduit la recherche pour plus, et meilleur approches de classification météorologiques [4]. En raison de la quantité énorme de données fournie par la station météorologique d'Annaba, des outils efficaces d'analyse sont indispensables pour extraire les dispositifs utiles, fournissant des informations plus simples et plus maniables. Les techniques d'extraction de connaissances sont des méthodes statistiques avancées développées pour cette tâche. La carte auto-organisatrice de kohonen (SOM) est l'une des techniques de datamining les plus connues qui a prouvé son efficacité pour le regroupement des données multidimensionnelles.

La méthode SOM permet à la fois de partitionner l'ensemble de données dans des clusters (représentés par des vecteurs référents) et de projeter les vecteurs prototypes dans un treillis de neurones de dimension réduite (habituellement 2D) qui préserve les contraintes de voisinages de l'espace de données originale. Ainsi, le treillis résultant après un apprentissage non supervisé représente les différentes situations météorologiques, et par conséquent il est bien approprié pour montrer et caractériser les événements météorologiques particuliers au moyen des distributions représentées dans le treillis par les modèles associés. Pour une carte fortement peuplée (composée d'un nombre important de neurones), il est très utile de regrouper les unités similaires pour faciliter l'analyse quantitative de la carte et de données. Pour découvrir la méthode la plus appropriée à utiliser pour regrouper les unités de la SOM, nous avons procédé à une comparaison basée sur les performances des algorithmes de classification automatique les plus utilisés tel que : l'algorithme PAM (Partitioning Around Medoids), K-moyennes, et la classification hiérarchique (méthode de Ward). Une classification basée sur deux niveaux est plus performante que celle d'un regroupement direct (utilisant uniquement k-moyennes ou PAM) pour la classification de données et la réduction du temps de calcul. En effet cette approche à rendu possible le traitement de grandes masses de données dans des temps de calcul très raisonnable. L'exactitude des résultats obtenus par l'approche de classification automatique est vérifiée en utilisant une validation quantitative basée sur deux catégories de critères (internes et externes) et une validation qualitative. Ces indices de validation nous permet également de répondre à certains questions fréquentes tel que :"combien de clusters sont dans l'ensemble de données?","est ce que les résultats du processus de classification ajuste bien l'ensemble de données ?","y a-t-il un meilleur partitionnement pour notre ensemble de données?".

Pour étudier l'influence des paramètres météorologiques sur la pollution atmosphérique dans la région d'Annaba et sa périphérie où les effets de la pollution, notamment atmosphérique, sont gravissimes pour sa population, les clusters météorologiques pour la période 2003-2004 ont été utilisés. L'analyse en composante principale (ACP) a été utilisée pour chercher les relations cachées entre le polluant examiné et les facteurs météorologiques. Cependant, l'ACP s'est montrée insuffisante pour modéliser les relations entre les conditions météorologiques et la pollution atmosphérique du fait que ces relations sont de type non linéaire. Étant donné que les RNAs sont capable de capturer les relations non linéaires qui existent entre les paramètres météorologiques et les niveaux de concentration des polluants, leurs performances ont été trouvées supérieures une fois comparées aux méthodes statistiques telles que la régression linéaire multiple [5], [6] et c'est pour cette raison qu'ils sont considérés dans cette étude. Parmi toutes les architectures neuronales étudiées, celle qui semble mieux adaptée aux problématiques de la modélisation de la qualité de l'air est l'architecture des perceptrons multicouches (PMC).

Dans ce mémoire, nous commencerons par présenter le contexte géographique et météorologique de la région et notamment les paramètres météorologiques utilisés. Dans le deuxième chapitre nous abordons un état de l'art sur la classification des données, nous commençons par quelques définitions de certains outils mathématiques:(partition, hiérarchie,..), et comme la plupart des techniques de classification de données sont basées sur des mesures de proximité, les types et les indices de dissimilarité sont présentés. Les étapes d'une classification automatique, ainsi que les différentes techniques de classification seront présentées en divisant ces techniques en plusieurs familles (hiérarchique, par partition,..) pour chaque technique, nous citons les algorithmes les plus utilisés et nous citons pour chaque technique les points forts et les points faibles. Une des étapes les plus importantes pour l'analyse d'un regroupement est l'évaluation des résultats pour identifier le meilleur partitionnement de l'ensemble de données, pour cette raison nous avons présenté les outils et les critères de validités les plus connus pour évaluer une telle classification. Et nous terminerons ce chapitre par une brève représentation de quelques notions et techniques de l'analyse de données.

Dans le troisième chapitre, nous présentons une méthode de classification automatique basée sur deux niveaux. L'approche proposée consiste à utiliser la carte auto-organisatrice de Kohonen (SOM) pour le premier niveau et la classification par partition (k-moyennes) 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 de l'aéroport d'Annaba durant la période 1995 à 1999.

Dans le quatrième chapitre et en vue d'étudier l'influence des paramètres météorologiques sur la pollution atmosphérique pour la région d'Annaba, les principaux facteurs et sources de la pollution atmosphérique dans la région ont été présentés. Ainsi, nous présentons une évaluation statistique de l'ensemble de données (polluants et paramètres météorologiques) fourni par la station SAMASAFIA pour la période 2003-2004. Pour identifier les principales sources de la pollution de l'air, nous avons utilisé et discuté les résultats de l'analyse en composante principale (ACP). Ce chapitre examinera également un survol rapide sur les réseaux de neurones artificiels, et plus précisément le perceptron multicouche (PMC), ainsi que les notions et les principaux critères utilisés pour l'évaluation de la qualité d'un modèle neuronal. Et nous terminerons ce chapitre par l'évaluation des performances des modèles neuronaux et discuter les résultats obtenus afin d'évaluer les impacts des paramètres météorologiques sur la qualité de l'air dans la région d'Annaba.

Chapitre 1

Météorologie et pollution atmosphé--rique dans la région d'Annaba.

1.1. Profils géographique et climatique

La région d'Annaba est située dans la partie orientale de la côte d'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 méditerranéen 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 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 [7], [8].

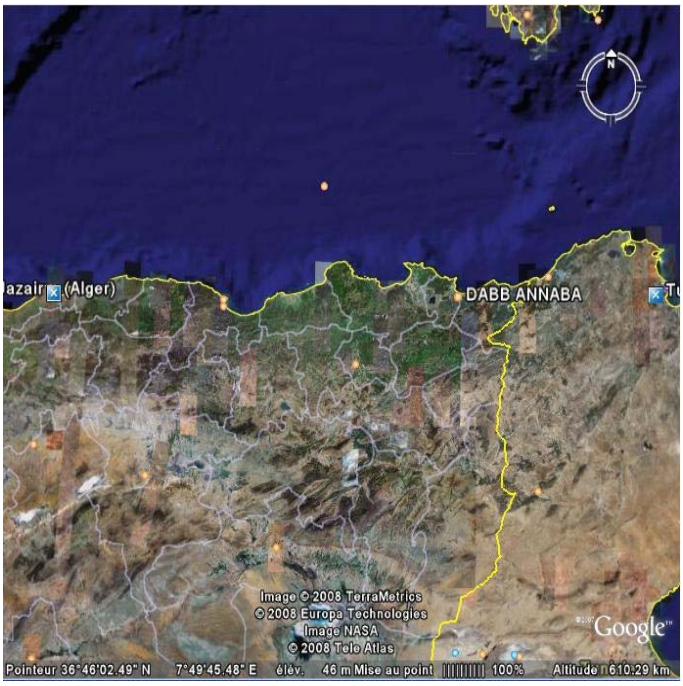


Figure 1.01-La région d'Annaba

Les paramètres météorologiques moyens mensuels de la région d'Annaba sont représentés par le tableau 1.1, (Selon la station météorologique de l'aéroport d'Annaba). Ces paramètres montrent également qu'il est possible d'extraire plusieurs situations météorologiques dont les paramètres sont similaires (classe météorologique). Selon ce tableau, nous pouvons également remarquer que les paramètres météorologiques pour le mois de juin, juillet, et aout sont presque identiques avec une température qui dépasse en moyenne 22 °C, et un faible taux d'humidité. Plusieurs autres classes peuvent être facilement identifiées tel que la classe du mois de décembre, janvier, février, ainsi la classe météorologique représentant le mois d'octobre et le mois de mai.

La topographie de la région d'Annaba est en forme de cuvette favorisant ainsi la stagnation de l'air et la formation d'inversions de températures. Ces situations permettent l'accumulation de polluants et l'élévation des taux de concentration qui en résulte. Les effets des brises de mer, terre, et pente concourent au transport des nuages de polluants. En effet, les nuages de polluants sont entraînés par la brise de terre la nuit vers la mer, et de jour. Ces nuages de polluants retournent sur la ville par effet de brise de mer en longeant la montagne de SERAIDI. Les nuages tournent sur la ville sous une forme de cercle. Les polluants se déposent lentement par gravité et l'on assiste à une pollution affectant les trois récepteurs (mer, terre, air). L'industrie est l'élément moteur de dégradation de la qualité d'air, cette industrialisation a assurément permis de répondre aux besoins des populations et du pays en produits sidérurgiques, engrais azotés, constructions ferroviaires et autres industries de transformations. A l'inverse, elle a suscité une urbanisation démesurée de la ville avec tous ses corollaires et une pollution de l'atmosphère et des sols, suivies de conséquences néfastes sur le biotope et la société [7].

Tableau 1.1--Paramètres météorologiques moyens dans la région d'Annaba.

	Pression (millibar)	*Temperature (°C)*	*Humidité (%)*	*Vitesse du vent (nœuds)*
Janvier	112190	12,285	75,973	6,8733
Février	107140	11,891	77,938	7,5507
Mars	68689	13,318	72,782	7,612
Avril	68354	15,327	72,725	7,5625
Mai	56815	19,144	74,219	7,1205
Juin	47815	22,783	72,092	7,93
Juillet	25268	24,786	71,135	8,4846
Aout	33402	26,158	71,791	8,0258
Septembre	58163	23,476	72,847	7,4396
Octobre	65284	19,931	73,097	6,8879
Novembre	119370	15,521	75,514	7,5688
Décembre	115030	12,925	75,669	7,7968

1.2. Paramètres météorologiques

Les paramètres météorologiques principaux à mesurer pour identifier les types de jours météorologiques d'une région quelconque, ou prédire un climat à un moment donné sont au nombre de sept. Ces paramètres météorologiques peuvent être divisés en trois familles. Le premier sous ensemble regroupe trois paramètres météorologiques qui sont des descriptions thermodynamiques de l'air : la température ; l'humidité et la pression atmosphérique. Le deuxième sous ensemble regroupe deux paramètres météorologiques qui sont des descriptions dynamiques de l'air : la direction du vent et la vitesse du vent. Le troisième sous ensemble regroupe deux paramètres météorologiques qui s'intéressent à la couverture nuageuse : les précipitations et l'ensoleillement. Quatre paramètres météorologiques représentant une description thermodynamique et dynamique de l'air ont été utilisés pour l'extraction des différentes situations météorologiques pour la région d'Annaba : l'humidité relative, la pression atmosphérique, la vitesse du vent et la température [9].

1.2.1. L'humidité relative

C'est la quantité de vapeur d'eau qui se trouve dans une particule d'air. L'humidité est présente en permanence dans l'atmosphère et même au niveau du Sahara ! La raison est la suivante : les rayons du Soleil réchauffent la surface de la terre et provoque l'évaporation de l'eau des océans ou de certaines réserves d'eau dans le Sahara. A l'inverse, l'humidité peut être absorbée, c'est le processus d'hygroscopique. Il arrive à un moment donné qu'une particule d'air soit saturée en vapeur d'eau mais pas tout le temps ; l'humidité relative est donc la quantité d'eau présente dans une particule d'air sur la quantité d'eau que peut contenir la particule d'air. La mesure de l'humidité relative reste très simple grâce à 2 instruments météorologiques aussi performants l'un que l'autre : -L'hygromètre et le -Le psychromètre.

1.2.2. La température

Dans l'air, il existe des particules d'eau aux propriétés physiques fortes différentes ; si bien que lorsque 2 particules d'eau se rencontrent, il y a interaction (elles ne se mélangent pas) ce qui entraînent des échanges d'énergie très importants qui donnent naissance à la température. Ces transferts d'énergie peuvent avoir lieu grâce à la conduction : transfert de la chaleur d'un point à un autre sans que les propriétés physiques de la particule d'air soient modifiées. La température se mesure, soit en degré Celsius (célèbre astronome et physicien Suédois 1701-

1744) noté °C, soit en degré Kelvin (alias William Thomson, physicien britannique 1824-1907) noté K tel que 1°C = 273,15 K [9].

1.2.3. La pression atmosphérique

La pression atmosphérique est la pression exercée par la colonne d'air se situant au dessus d'une surface. Elle dépend des conditions météorologiques et elle diminue avec l'altitude. Elle est couramment mesurée en hectopascals (hPa) à l'aide d'un baromètre. Le baromètre à colonne de mercure est le plus connu. Une pression qui monte est signe de beau temps, même si la pression est basse en valeur absolue. De même, une pression haute mais en baisse est signe de dégradation de la situation météorologique. La pression moyenne est de l'ordre de 10^5 N (newton) par mètre carré, ce qui était autrefois appelé le bar (son millième est le millibar) et qui vaut 10^5 Pa (pascal), soit 1000 hPa (hectopascal).

1.2.4. La vitesse du vent

Le vent est un mouvement horizontal de l'air sur la surface de la terre. Il naît d'une différence de pression, et se propage perpendiculairement aux isobares, des pressions hautes vers les basses, de façon à réduire les écarts de pression. Le vent peut être défini par sa direction (le plus souvent son origine) et par sa vitesse (en Beaufort, en kilomètre par heure, en mètre par seconde...). On utilise pour mesurer la direction du vent une girouette et pour la vitesse un anémomètre.

1.3. La pollution atmosphérique

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 [10], [11].

Depuis le début du siècle dernier l'accroissement démographique et le développement industriel sont à l'origine d'importantes émissions de gaz et d'aérosols (particules en suspension dans l'air). Les modifications de la constitution de l'atmosphère qui en découlent, peuvent avoir des répercussions aussi bien à l'échelle locale (conséquences sur la santé humaine, les végétaux ou les matériaux) qu'à l'échelle planétaire (modification du climat : effet de serre, diminution de la couche d'ozone stratosphérique).

Les principales émissions anthropiques concernent le dioxyde de soufre (SO2), les oxydes d'azote (NOx), le monoxyde de carbone (CO), les composés organiques volatils (COV), les aérosols...

Certains polluants sont émis directement par une source. C'est le cas notamment du dioxyde de soufre (SO2) et du monoxyde d'azote (NO). Ils sont dits primaires. Les concentrations dans l'air de ces polluants sont maximales à proximité des sources, puis tendent à diminuer au fur et à mesure que l'on s'éloigne de celles-ci du fait de leur dilution dans l'air. Des polluants peuvent évoluer chimiquement après leur émission, se transformer ou produire d'autres composés. Ce sont des polluants dits secondaires. L'ozone, qui se forme à partir des oxydes d'azote et des COV sous l'action du rayonnement solaire, appartient à cette famille [10].

1.3.1. Le monoxyde de carbone (CO)

Le monoxyde de carbone résulte d'une combustion incomplète des combustibles et carburants. Dans l'air ambiant, on le rencontre essentiellement à proximité des voies de circulation routière. Il provoque maux de tête, vertiges. Il est mortel, à forte concentration, en cas d'exposition prolongée en milieu confiné.

1.3.2. L'ozone (O3)

L'ozone provient de la réaction des polluants primaires (issus de l'automobile ou des industries) en présence de rayonnement solaire et d'une température élevée. Il provoque toux, altérations pulmonaires, irritations oculaires.

1.3.3. Le dioxyde d'azote (NO2)

Les oxydes d'azote proviennent des combustions et du trafic automobile. Le dioxyde d'azote provient à 60% des véhicules. Ils affectent les fonctions pulmonaires et favorisent les infections.

1.3.4. Le dioxyde de soufre (SO2)

Le dioxyde de soufre (SO2) provient essentiellement de la combustion des combustibles fossiles contenant du soufre tels que le fuel et le charbon. Il est émis par les industries, le chauffage urbain. Il irrite les muqueuses, la peau et les voies respiratoires supérieures.

1.3.5. Particules en suspension (Particulate Matter PM 10)

Les particules en suspension proviennent du trafic automobile, des chauffages fonctionnant au fioul ou au bois et des activités industrielles. Plus elles sont fines, plus ces poussières pénètrent profondément dans les voies respiratoires. 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 [11].

1.4. Base de données

Les données utilisées dans le cadre de cette étude ont été collecté par la station météorologique de l'aéroport d'Annaba, et la station SAMASAFIA de Sidi Amar.

A/ la station météorologique de l'aéroport

La base de données collectée par cette station contient 04 paramètres météorologiques captés pendant 60 mois (1995- 1999) avec une échéance de 3 heures. Donc chaque individu de la base météorologique (un jour quelconque inclut dans cette période) est caractérisé par 32 paramètres durant les 24 heures, ces paramètres météorologique sont: la pression mesurée en dixièmes de millibar ; la température mesurée en dixièmes de °C ; l'humidité mesurée en centièmes et la vitesse du vent mesurée en nœuds.

B/ la station SAMASAFIA de Sidi-Amar

La base de données collectée par la station SAMASAFIA (structure responsable de la surveillance de la qualité de l'air en Algérie) d'Annaba sur une base de mesure continue de 24 heures pendant la période 2003-2004. Les polluants atmosphériques surveillés en continu inclut 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 de vent, la température et l'humidité relative.

Chapitre 2

Classification non supervisée : état de l’art.

2.1. Introduction

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 tous via le web. De nombreux processus industriels sont également de plus en plus contrôlés par l'informatique. Et de nombreuses mesures effectuées un peut 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 [12].

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. Dans ce travail de magister, nous nous intéressons à la classification non supervisée (automatique) pour l'identification des types de jours météorologiques de la région d'Annaba. La classification automatique a été utilisée avec sucées dans de nombreuses études météorologiques tel que dans [13]. La méthode de l'analyse en composante principale (ACP) et l'algorithme k-moyennes ont été utilisés dans [14] pour déterminer les scénarios météorologiques synoptiques. L'ACP a prouvé son utilité comme une technique puissante d'extraction de connaissances à partir de données [15], cependant sa visualisation n'est pas bien appropriée pour représenter la structure complexe d'un ensemble de données [16]. Récemment, et comme outil alternatif servant à traiter ce problème de complexité des données multidimensionnelles, les cartes auto-organisatrices de Kohonen (SOM) ont été largement utilisées pour le regroupement des données météorologiques [17], [18], [19].

La classification non supervisée est liée à plusieurs domaines de recherche. Elle a été utilisée couramment dans les statistiques [20], la reconnaissance des formes, segmentation et traitement d'images [21], et dans beaucoup d'autres domaines d'application tel que la recherche documentaire, forêt, agriculture, etc. [22], [23], qualité de l'eau [24], [25], identification des types de jours des charges électriques [26].

Le but de ce chapitre est d'introduire les notations et les concepts de base sur lesquels s'appuiera la suite de ce mémoire, et de mettre en évidence la diversité qu'il existe parmi les

différentes méthodes de classification non supervisée. Le lecteur intéressé est invité à consulter l'une des nombreuses références disponibles [27] ; [28]; [29] ; [30] ; [31] pour approfondir son étude.

2.2. Concepts et définitions utiles

La classification est une étape importante pour l'analyse de données. Elle consiste à regrouper les objets d'un ensemble de données en classes homogènes. 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. La classification supervisée (ang. classification) est basée sur un ensemble d'objets *L* (appelé ensemble d'apprentissage) de classes connues, le but étant de découvrir la structure des classes à partir de l'ensemble L afin de pouvoir généraliser cette structure sur un ensemble de données plus large. La classification non supervisée (ang. clustering) consiste à diviser un ensemble de données *D* 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 [32].

Nous commençons par rappeler quelques concepts et définitions formelles essentielles pour comprendre les différentes méthodes et outils de classification avant de présenter quelques approches utilisées en classification automatique.

2.2.1. Qu'est ce qu'une classification

Le concept identification des types de jours météorologiques 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 ce manuscrit. La définition qui suit correspond à la notion de classification dure mais ce qualificatif ne sera plus précisé dans la suite du papier.

Définition 2.2.1 : ***Partition***

Étant donnée un ensemble fini d'objets, on appelle partition toute famille de parties non vides disjointes deux à deux dont l'union forme l'ensemble *I*. On appelle partition de *I*, l'ensemble *P*, $P = \{C_i, i \in I\}$

I : ensemble d'indices, C_i : partie de *I* (ou une classe) possédant les propriétés suivantes:

- $\forall i \in I, C_i \neq \phi$
- $\forall i \in I, \forall j \in I, i \neq j \rightarrow C_i \cap C_j \neq \phi$
- $\bigcup_{i \in I} C_i = I$
- $\forall i \in I, C_i \neq \phi$

Comparaison de partitions :

Soit P et P' deux partitions de I : $P = \{C_i, i \in I\}$ et $P' = \{C'_j, j \in I'\}$

P est plus fine que P' (ou P' est moins fine que P) si et seulement si $\forall i \in I, \exists j \in I', C_i \subset C'_j$

(Chaque élément de P' est une réunion d'élément de P)

P_D est une partition discrète si $P_D = \{\{x\}, x \in I\}$

P_G est dite partition grossière si $P_G = \{I\}$

Définition 2.2.2. ***Recouvrement***

On appelle un recouvrement de I l'ensemble P, $P = \{C_i, i \in I\}$

I : ensemble d'indices.

C_i : partie de I (ou une classe)

Possédant les propriétés : $\forall i \in I, C_i \neq \phi$ et $\bigcup_{i \in I} C_i = I$

Hiérarchies de parties

Une hiérarchie est un ensemble de parties de I issues d'une suite emboîtée de partitions, et qu'on peut représenter par un dendrogramme.

Définition 2.2.3. ***Hiérarchie***

On appelle hiérarchie de parties de I tout sous ensembles, $\mathcal{H}$ de $\mathcal{P}(I)$ tel que :

1. $\phi \notin$ H
2. $\forall x \in I, \{x\} \in H$
3. $I \in H$
4. $\forall H_1 \in H, \forall H_2 \in H \Rightarrow H_1 \cap H_2 = \phi$ ou $H_1 \subset H_2$ ou $H_2 \subset H_1$

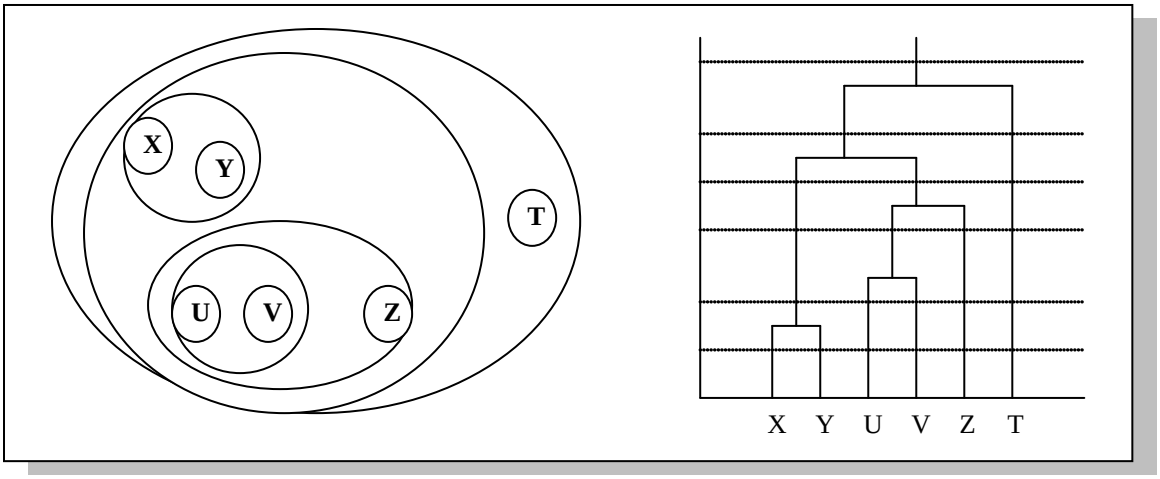


Figure 2.1-Exemples d'une hiérarchie de parties de l'ensemble I.

2.2.2. Notion 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 [30] ; [31] ; [32] :

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

Il est commun de définir le concept de similarité à l'aide de la notion duale de dissimilarité; on dit que deux individus sont similaires s'ils sont proches au sens d'une mesure de dissimilarité. Nous présentons par la suite la définition générale d'une mesure de dissimilarité ainsi que les deux concepts : métrique et ultra métrique.

2.2.2.1. Définitions :

Définition 2.2.2.1. (*Mesure de dissimilarité)*

On appelle indice ou mesure de dissimilarité sur un ensemble I, une application $d : I \times I \to R_+$ qui vérifie les propriétés suivantes pour tout couple $(x, y) \in I \times I$:

$$d(x, y) = d(y, x) \quad \text{(Symétrie)}$$

$$d(x, y) = 0 \Leftrightarrow x = y \quad \text{(Séparabilité)}$$

Définition 2.2.2.2. (*Métrique)*

On appelle métrique sur un ensemble I, une application $d : I \times I \to R_+$ qui vérifie les propriétés suivantes pour tout couple $(x, y) \in I \times I$:

$$d(x, y) = d(y, x) \quad \text{(Symétrie)}$$

$$d(x, y) = 0 \Leftrightarrow x = y \quad \text{(Séparabilité)}$$

$$d(x, y) \leq d(x, y) + d(z, y) \quad \text{(Inégalité triangulaire)}$$

Définition 2.2.2.3. (*Ultramétrique)*

on appelle métrique sur un ensemble I, une application $d : I \times I \to R_+$ qui vérifie les propriétés suivantes pour tout couple $(x, y) \in I \times I$:

$$d(x, y) = d(y, x) \quad \text{(Symétrie)}$$

$$d(x, y) = 0 \Leftrightarrow x = y \quad \text{(Séparabilité)}$$

$$d(x, y) \leq \max\{d(x, y), d(z, y)\} \quad \text{(Inégalité ultramétrique)}$$

L'homogénéité des individus regroupés au sein d'un groupe est souvent évaluée à l'aide d'un critère statistique appelée variance dont la définition est rappelée ci-dessous.

Définition 2.2.2.4. (*Variance)*

On définit la variance $V(C_i)$ d'un groupe d'objets C_i ainsi :

$$V(C_i) = \frac{1}{N_i} \sum_{x_j \in C_I} d^2(x_j - \mu_i) \quad \textbf{(2.1)}$$

Ou N_i et μ_i sont respectivement le nombre d'objets et le centroide du groupe C_i.

Dans le contexte de la classification automatique, on distingue généralement la **variance intra-classe** V_{intra}, que l'on souhaite minimiser, de **la variance inter-classe** V_{inter}, que l'on cherche à maximiser :

$$V_{intra} = \frac{1}{N} \sum_{C_I \in C} N_i \times V(C_i) \quad \textbf{(2.2)}$$

$$V_{inter} = \frac{1}{N} \sum_{C_I \in C} N_i \times (\mu_i - \mu)^2 \quad \textbf{(2.3)}$$

Ou N_i et μ_i sont respectivement le nombre d'objets et le centroïde du groupe C_i, et de manière analogue N et μ désignent respectivement le nombre d'objets et le centroïde de I, la première équation évalue l'homogénéité moyenne des groupes d'une partition et la seconde permet de quantifier la différence entre les groupes. La formule de könig-Huyghens permet de relier la variance intra-classe et inter-classe à la variance totale $V_{totale}=V(I)$

$$V_{totale}=V_{intra}+V_{inter} \quad \textbf{(2.4)}$$

Il existe trois concepts de similarité en classification automatique [24] ; [12] ; [33] :

- La similarité entre objets : à maximiser pour deux objets appartenant au même cluster, et à minimiser pour deux objets appartenant à des clusters différents;
- la similarité entre un objet et un cluster : à maximiser si l'objet est associé au cluster pour une bonne cohésion interne du cluster;
- la similarité entre clusters : à minimiser pour une bonne isolation externe des clusters.

2.2.2.2. Similarité entre objets

La similarité entre objets est estimée par une fonction qui permet de calculer la distance entre ces objets. Une fois cette fonction distance définie, la tache de classification consiste alors à réduire au maximum la distance entre les membres d'un même cluster tout en augmentant au

maximum la distance entre clusters. Ainsi, deux objets proches selon cette distance seront considérés comme similaires, et au contraire, deux objets séparés par une large distance seront considérés comme différents. Le choix de cette mesure de distance entre objets est très important. 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. Les mesures de distance les plus courantes entre deux objets sont les suivantes :

- La distance de Manhattan :

$$dist_1(x_i, x_j) = \left| x_i - x_j \right| = \sum_{d=1}^{M} \left| x_{id} - x_{jd} \right| \quad \textbf{(2.5)}$$

- La distance euclidienne :

$$dist_2(x_i, x_j) = \left\| x_i - x_j \right\|_2 = \sqrt{\sum_{d=1}^{M} \left| x_{id} - x_{jd} \right|^2} \quad \textbf{(2.6)}$$

- La distance de Minkowski, qui généralise les deux précédentes :

$$dist_p(x_i, x_j) = \left\| x_i - x_j \right\|_p = \sqrt[p]{\sum_{d=1}^{M} \left| x_{id} - x_{jd} \right|^p} \quad \textbf{(2.7)}$$

- Le cosinus :

$$Cos(x_i, x_j) = \frac{x_i . x_j}{\left\| x_i \right\| . \left\| x_j \right\|} = \frac{\sum_{d=1}^{M} \left(x_{id} . x_{jd} \right)}{\sqrt{\sum_{d=1}^{M} x_{id}^2 \sum_{d=1}^{M} x_{jd}^2}} \quad \textbf{(2.8)}$$

- La corrélation de Pearson, cosinus de l'écart à la moyenne :

$$Pearson(x_i, x_j) = \frac{\sum_{d=1}^{M} \left(x_{id} - \bar{x}_i \right) \left(x_{jd} - \bar{x}_j \right)}{\sqrt{\sum_{d=1}^{M} \left(x_{id} - \bar{x}_i \right)^2 \sum_{d=1}^{M} \left(x_{jd} - \bar{x}_j \right)^2}} \quad \textbf{(2.9)}$$

avec $\bar{x}_i$ la moyenne des valeurs de l'objet x_i sur l'ensemble des attributs :

$$\bar{x}_i = \frac{\sum_{d=1}^{M} x_{id}}{M} \quad \textbf{(2.10)}$$

2.2.2.3. Cohésion interne d'un cluster

Étant donnée une mesure de distance entre les objets nommés *dist* et choisie parmi celles proposées précédemment ou d'autres, la cohésion interne des clusters peut être définie, et les deux mesures les plus utilisées dans ce cadre étant les suivantes :

1. le radius, distance moyenne entre un objet membre d'un cluster et son centroïde, à minimiser pour maximiser la similarité des observations à l'intérieur du cluster :

$$Raduis(C_k) = \frac{\sum_{i \in D_k} dist(x_i, \mu_k)}{N_k}$$

(2.11)

2. et le diamètre, qui constitue la distance moyenne entre un paire d'objet dans un même cluster, également à minimiser :

$$Diam(C_k) = \frac{\sum_{i \in D_K} \sum_{j \in D_k} dist(x_i, \mu_k)}{N_k \times (N_k - 1)} \quad \textbf{(2.12)}$$

2.2.2.4. Isolation externe d'un cluster

La notion de similarité entre différents clusters est très importante en classification automatique. Ainsi, une première possibilité pour mesurer la similarité ente deux clusters est de calculer la distance ente les objets qui les représentent. Il s'agit alors de maximiser la distance entre les centroïdes, ou bien la distance qui sépare les objets les plus proches de chaque cluster. Cependant, observant qu'en assignant aux *K-1* premiers clusters (pour *k* le nombre de clusters recherchés) les *k-1* objets les moins similaires avec le reste de la base, puis au dernier cluster le reste des objets de la base, la somme pondérée des distances entre les centroïdes est maximisée, une autre solution proposée dans [34] pour pallier à ce problème consiste à maximiser la somme des distances entre chaque cluster et le centroïde de l'ensemble des objets de la base :

$$x_0 = \frac{\sum_{i=1}^{N} x_i}{N} \quad \textbf{(2.13)}$$

Trois autres mesures peuvent également être utilisées dans ce cadre :

- la distance moyenne inter-clusters :

$$Inter(k1, k2) = \frac{\sum_{i \in D_{K1}} \sum_{j \in D_{k2}} dist(x_i, x_j)}{N_{k1} \times N_{k2}} \quad \textbf{(2.14)}$$

- la distance moyenne intra-clusters :

$$Intra(k1,k2) = \frac{\sum_{i\in D_{k1}\cup D_{k2}} \sum_{j\in D_{k1}\cup D_{k2}} dist(x_i, x_j)}{(N_{k1}+N_{k2})(N_{k1}+N_{k2}-1)} = Diam(C_{k1}\cup C_{k2}) \quad \textbf{(2.15)}$$

- où la variance globale entre clusters :

$$Var(k1,k2) = \sum_{i\in D_{k1}\cup D_{k2}} dist(x_i, \mu_{k1,k2}) - \sum_{i\in D_{k1}} dist(x_i, \mu_{k1}) - \sum_{j\in D_{k2}} dist(x_j, \mu_{k2}) \quad \textbf{(2.16)}$$

2.2.3 Les étapes d'une classification automatique

Les étapes de base pour développer un processus de classification automatique sont présentées par la figure 2.2 et peuvent être récapitulées comme suit [31] ; [35] :

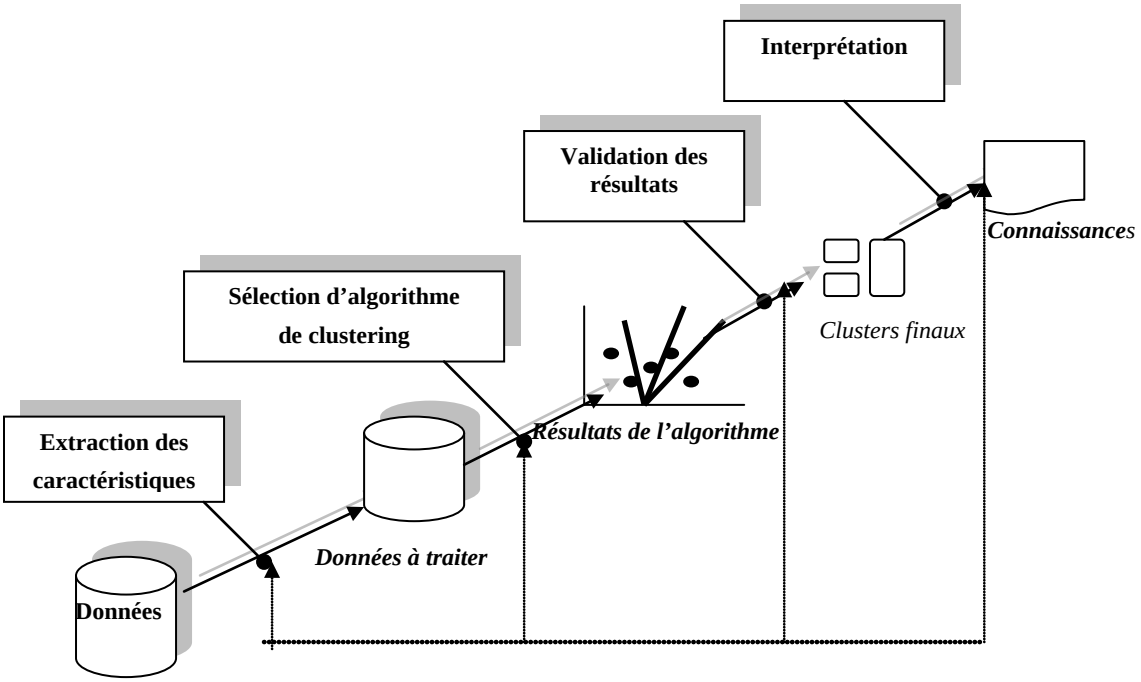


Figure 2.2--Les étapes d'un processus de classification automatique.

Sélection des caractéristiques : le but de cette phase est de sélectionner ou d'extraire un sous ensemble optimal de caractéristiques pertinentes pour un critère fixé auparavant. La sélection de ce sous ensemble de caractéristiques permet d'éliminer les informations non pertinentes et redondantes selon le critère utilisé. Ainsi, le prétraitement des données peut être nécessaire avant de les utiliser dans une tâche de classification.

Algorithme de classification automatique : Cette étape se rapporte au choix d'un algorithme de classification le mieux adapté pour le regroupement de l'ensemble de données. Une mesure de proximité et le critère de regroupement caractérisent principalement un algorithme

regroupement aussi bien que son efficacité pour définir les clusters représentatifs de l'ensemble de données.

- 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 et il n'y a aucun variable qui domine d'autres.
- Critère de regroupement. Dans cette étape, nous devons définir le critère de regroupement, qui 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. L'exactitude des résultats obtenus par les algorithmes de regroupement est vérifiée en utilisant des techniques et des critères bien appropriés. Les algorithmes de regroupement permettent d'extraire des clusters qui ne sont pas connu a priori, indépendamment de ces méthodes, une classification finale d'un ensemble de données exige un certain genre d'évaluation dans la plupart des applications [36].

Interprétation des résultats : cette étape vise à :

- l'évaluation de la qualité d'une classification,
- la description des classes obtenues.

2.2.4. Présentation des méthodes de classification de données

Une diversité de méthodes de classification non supervisée est proposée dans la littérature. 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. Les méthodes de regroupement peuvent être classé selon [31] ; [37] :

- 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).

Ainsi selon la méthode adoptée pour définir les clusters, les algorithmes de regroupement peuvent être largement classés dans les catégories suivantes :

- Algorithmes hiérarchiques,

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.

- 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-moyennes
- 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

2.3. Les méthodes de la classification automatique

2.3.1. L'approche neuromémitique

2.3.1.1. Source historique et principes :

Les cartes auto-organisatrices communément désigné par SOM (pour Self Organising Maps) ont été introduites par T.Kohonen en 1981 en s'inspirant du fonctionnement des systèmes neuronaux en biologie, plus précisément du fait que les zones du cerveau qui gèrent le fonctionnement du corps humain respectent la topologie du système physique. D'un point de vue informatique, on peut traduire cette propriété de la façon suivante : supposons que l'on dispose de données que l'on désire classifier. On cherche un mode de représentation tel que des données voisines soient classées dans la même classe ou dans des classes voisines. Ce type de réseaux de neurones artificiels a largement montré son efficacité pour la classification des données multidimensionnelles, mais malheureusement il a été resté ignoré de nombreuses années malgré son grand intérêt. Le principe des cartes de Kohonen est de projeter un ensemble de données complexe 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 s'est produite tout en conservant la topologie et les métriques les plus importantes des données initiales 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 [38, 39, 40, 41, 42].

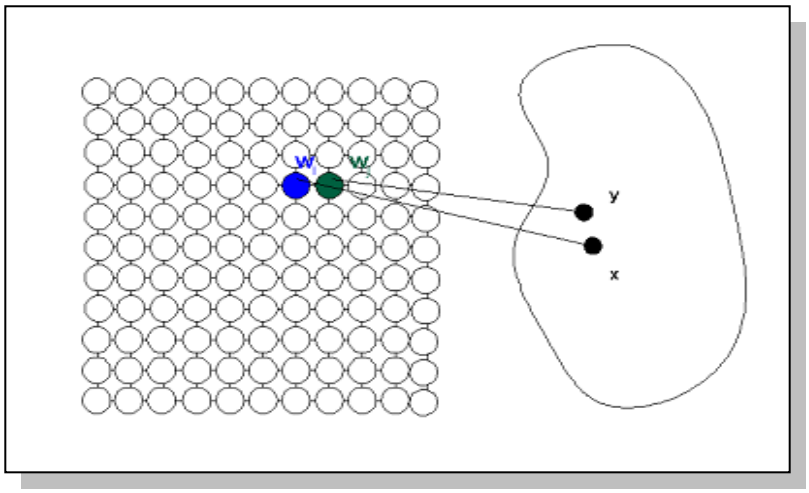


Figure 2.3-Des neurones voisins sur la carte représentent des observations assez "proche" dans l'espace des données.

2.3.1.2. Architecture des cartes de Kohonen

La structure de base d'une carte de kohonen est composée de *M* neurones éparpillé sur une grille régulière de basse dimension, habituellement 1-ou 2 dimensions, les grilles de grandes dimensions sont possibles, mais elles ne sont pas généralement utilisées puisque leur visualisation est problématique [43]. Si la visualisation n'est pas nécessaire, les grilles dont la dimension est supérieure à trois peuvent être bénéficières [44].

La carte de kohonen se compose habituellement de deux couches de neurones une couche d'entrée et une couche de sortie. Dans la couche d'entrée tout individu à classer (dans notre cas, un jour de la semaine) est représenté par un vecteur multidimensionnel (voir section 1.4). Chaque individu va affecter un neurone qui représente le centre du cluster. La couche (topologique) d'adaptation ou la couche de sortie est composée d'un treillis de neurones selon la géométrie prédéfinie [14] ; [19]. 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 référent ou le prototype associé à chaque neurone, il est de même dimension que les vecteurs d'entrées. 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 au sens de la distance euclidienne de l'individu présenté. C'est l'utilisation de la notion de voisinage qui introduit les contraintes topologiques dans la géométrie finale des cartes de kohonen. Les différentes formes géométriques que peuvent avoir une carte de kohonen sont présentées dans la figure 2.5. Ainsi, La structure de base d'une carte de kohonen bidimensionnelle de voisinage rectangulaire avec M=3 (dimension des vecteurs d'entrées) et L=4*3=12 neurones est montrée par la figure 2.4. Un vecteur d'entré $x(t) = [x_1,..x_M]^T$ est appliqué 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 $w_j = [w_{j1},.., w_{jM}]^T$.

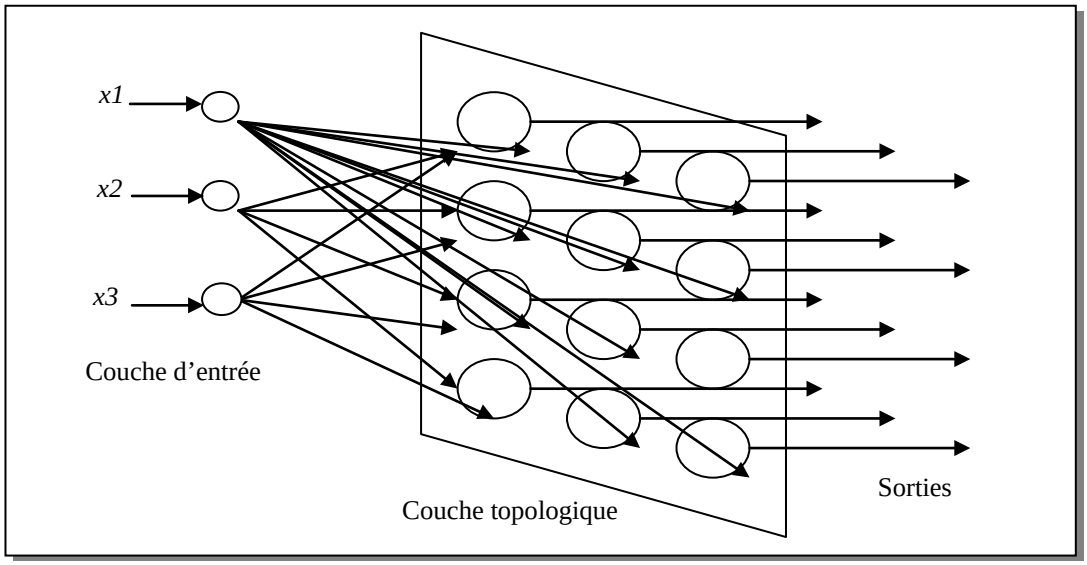


Figure 2.4-Structure d'une carte auto-organisatrice.

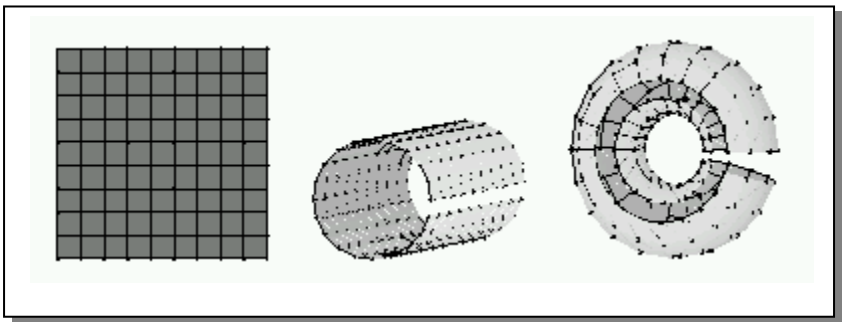


Figure 2.5-Différentes formes de cartes : (a) rectangulaire (la plus utilisée), (b) cylindrique et (c) toroïdale.

2.3.1.3. Matérialisation du Voisinage

On peut utiliser deux façons pour repérer une unité sur une grille. La première consiste à numéroter les unités de 1 à *U* (ligne par ligne). Dans la seconde, on affecte à l'unité *u* ses coordonnées cartésiennes sur la carte (*i, j*). Pour la ficelle, ces deux notations sont identiques. 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*. En utilisant les coordonnées cartésiennes, on peut définir la distance *d* par [45] ; [46] :

$d(u, u_o) = max(|i_u - i_{uo}|, |j_u - j_{uo}|)$ pour une grille

$d(u, u_o) = max(|i_u - i_{uo}|) = \max(|u - u_o|)$ pour une ficelle

Dans les deux cas, $V_r(u_o) = \{u \in \{1,\ldots, U\} / d(u,u_o) <= r\}$

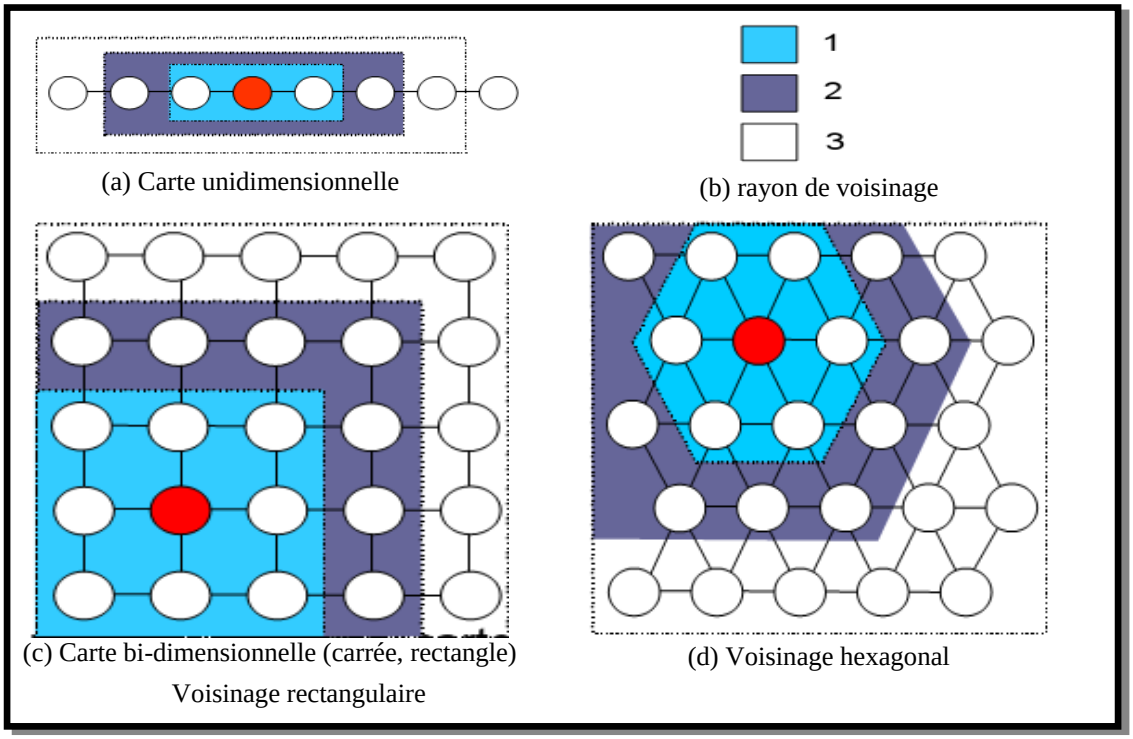


Figure 2. 6-Différentes topologies et voisinages des cartes de kohonen.

Exemples

Dans le cas d'une grille 5x5, les unités qui constituent le voisinage de rayon 2 de l'unité 17 repérée par le bipoint (4,2) vérifient la relation :

$max(|i_u - i_{17}|, |j_u - j_{17}|) = max(|i_u - 4|, |j - 2|) <= 2$

La figure 2.6(c) représente une telle grille où l'unité 17 correspond à la case rouge et ses voisines aux cases bleues, c'est-à-dire les unités 6, 7, 8, 9, 11, 12, 13, 14, 16, 18, 19, 21, 22, 23, 24.

La figure 2.6(d) représente une grille hexagonal où l'unité 7 correspond à la case rouge et ses voisines aux cases bleues, c'est-à-dire les unités 1, 2, 3, 4, 6, 8, 9, 11, 12, 13, 14, 16, 17, 18.

2.3.1.4. Fonctions de voisinage

La fonction de voisinage détermine à quel degré les neurones sont reliés entre eux [47]. La fonction de voisinage la plus simple est la fonction bulle : elle est constante sur tout le voisinage du neurone gagnant et zéro ailleurs. Une autre fonction de voisinage très utilisée est la fonction gaussienne $\exp(-\frac{\|r_c - r_i\|^2}{2\delta^2(t)})$, où r_c est l'emplacement de l'unité *c* sur la grille de la carte et *sigma (t)* est le rayon de voisinage au temps *t*. les quatre fonctions de voisinage les plus connus sont montrées dans la figure 2.7.

La fonction de voisinage et le nombre de neurones déterminent la granularité des résultats. Plus la région où la fonction de voisinage a des valeurs élevées est grande, plus la carte est rigide. Plus la carte est grande, plus elle devient flexible. Cet effet détermine les possibilités d'exactitude et de généralisation d'une SOM.

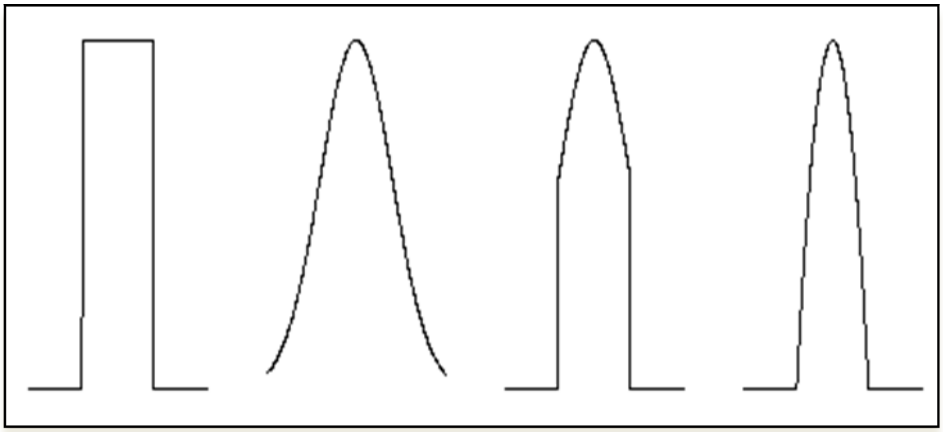


Figure 2.7-Fonctions de voisinage : bulle, gaussienne, coupe gaussienne et epanechicov.

2.3.1.5. Algorithme d'apprentissage

Avant de procéder à l'apprentissage, des valeurs initiales sont attribuées aux vecteurs prototypes. La SOM est très robuste en ce qui concerne l'initialisation, mais une bonne initialisation permet à l'algorithme d'apprentissage de converger plus rapidement à une bonne solution. En général une des trois procédures d'initialisation suivantes est utilisée [20] ; [14] ; [48] :

- Initialisation linéaire, l'initialisation est faite en calculant d'abord les vecteurs propres de l'ensemble de données, les vecteurs poids de la carte sont alors initialisés le long des *mdim* plus grands vecteurs propres de la matrice de covariance de données d'apprentissage, où *mdim* est la dimension de la grille de la carte, en général 2.
- Initialisation aléatoire, où les vecteurs poids sont initialisés avec de petites valeurs aléatoires.
- échantillons aléatoires tirés de l'initialisation linéaire des données d'entrée.

Après avoir choisir l'architecture de la carte, vient donc l'étape d'apprentissage ou encore l'estimation des paramètres du modèle (les poids synaptiques appelés vecteurs référents). Cette phase se fait d'une manière itérative. Chaque itération se compose de deux étapes : une étape de compétition entre les neurones qui détermine la région du treillis que l'on va ajuster, et une étape d'adaptation des poids de la zone sélectionnée à l'observation projetée. Le principe des deux phases est illustré sur la figure 2.8.

Étape de compétition :

Si on prend *X* un espace de taille *M* dans lequel est répartie une distribution de points, chaque point est représenté par un vecteur $x \in X$. Soit $\mathcal{M}$ une grille de neurones de dimension *M*. le vecteur poids synaptique *w* de chaque neurone possède une représentation dans l'espace d'entrée, il peut être vu comme un vecteur de références. De cette façon, à un instant *k* ou est présenté un vecteur prototype x_k tiré de la distribution de l'espace d'entrée, tous les neurones de la grille sont mis en compétition. Cette compétition revient à chercher un neurone vainqueur, c'est à dire celui qui se rapproche le plus du vecteur d'entrée. En d'autre termes, parmi tous les neurones de la carte, le neurone vainqueur d'indice $i_k=i$ (x_k) 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 BMU (Best Matching Unit), est obtenu par :

$$i(x_k) = \arg\min_{j \in M} \left\| x_k - w_{j,k} \right\| \qquad \textbf{(2.17)}$$

Le neurone vainqueur, pour un stimulus est également dénommé 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 envisagé.

Étape d'adaptation :

Le processus de compétition permet de déterminer la façon d'ajuster les poids des neurones de la carte. Ainsi, 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) :

$$w_{j,k+1} = w_{j,k} + \Delta w_{j,k} \qquad \textbf{(2.18)}$$

$$= w_{j,k} + \eta_k h_{j,i(x_k),k}(x_k - w_{j,k}) \qquad \textbf{(2.19)}$$

avec : $$\Delta w_{j,k} = \eta_k h_{j,i(x_k),k}(x_k - w_{j,k}).$$

Dans les expressions (3) et (4) η_k représente le coefficient d'apprentissage et, $h_{j,i(x_k),k}$ est une fonction de voisinage. L'adaptation des poids de chaque neurone est donc fonction de la position d'un neurone dans la grille m par rapport au neurone gagnant.

Le coefficient d'apprentissage pour rendre l'apprentissage plus performant, est généralement fonction du temps et du type :

$$\eta_k = \eta_i \left(\frac{\eta_f}{\eta_i} \right)^{\frac{k}{k_{\max}}} \qquad \textbf{(2.20)}$$

Ou η_i représente la valeur initiale du coefficient, η_f sa valeur finale et $k_{\max}$ détermine la durée de l'apprentissage. Au cours de l'apprentissage, la taille du voisinage du BMU, qui détermine la zone active, décroît avec le temps. L'évolution temporelle du coefficient d'apprentissage est illustrée sur la figure 2.8.

La fonction de voisinage généralement adoptée est la fonction gaussienne (invariante selon la translation c'est-à-dire qu'elle ne dépend pas du neurone vainqueur) cette fonction est centrée sur le neurone déclaré vainqueur après la phase de compétition qui à suivi la présentation d'un vecteur d'entrée.

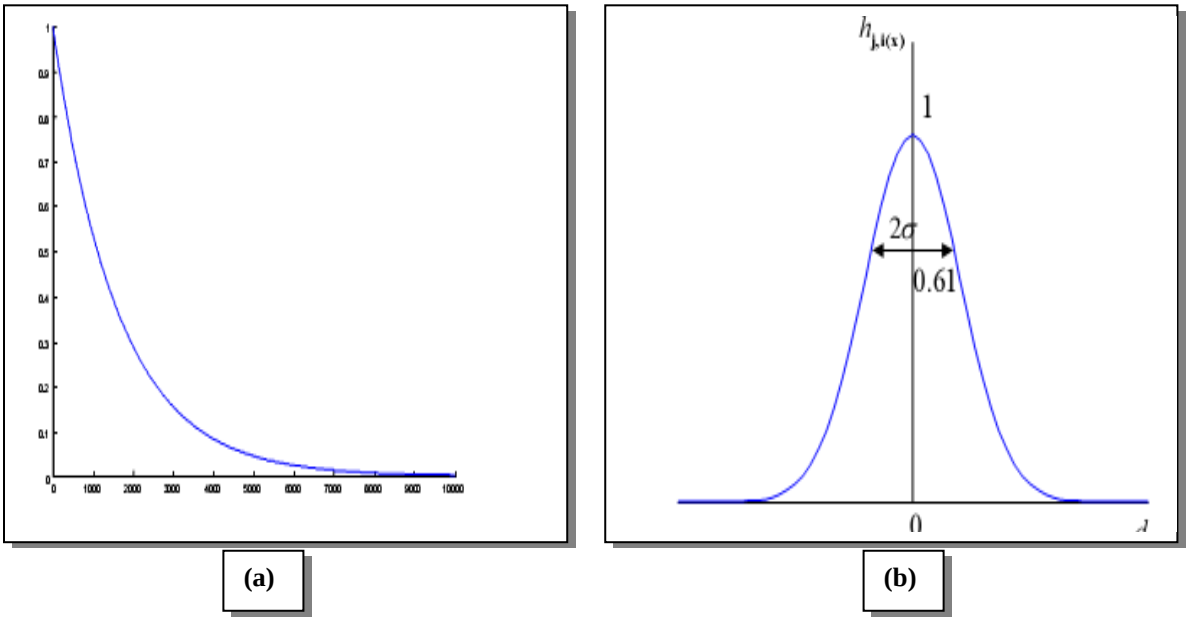


Figure 2.8--Évolution des paramètres d'une carte de kohonen au cours de l'apprentissage. (a) l'évolution du cœfficient d'apprentissage au cours de l'apprentissage. (b) L'allure de la fonction de voisinage pour un rayon donné (s=0.61).

La modification appliquée dans 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 présenté, et moins probable si le vecteur d'entrée est totalement différent du vecteur précédent. Comme nous avons dit précédemment, la fonction de voisinage tient compte de la distance par rapport à la position du neurone vainqueur pour pondérer la correction des poids synaptiques delta du neurone j à l'instant k, soit $d_{j,i}$ la distance entre le neurone vainqueur d'indice i et un neurone voisin d'indice j. Cette distance ne se calcule pas dans l'espace des entrées mais dans l'espace topologique de la carte :

$$d_{j,i}^2 = \|j - i\|^2 \quad \textbf{(2.21)}$$

La fonction de voisinage $h_{j,i(x_k),k}$ s'écrit alors :

$$h_{j,i(x_k),k} = \exp\left(-\frac{d_\mu^2}{2\delta^2}\right) \quad \textbf{(2.22)}$$

Où δ est u rayon de voisinage. Ce rayon δ peut être dépendant du temps selon l'expression suivante :

$$\delta_k = \delta_i \left(\frac{\delta_f}{\delta_i}\right)^{\frac{k}{k_{\max}}} \quad \textbf{(2.23)}$$

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.

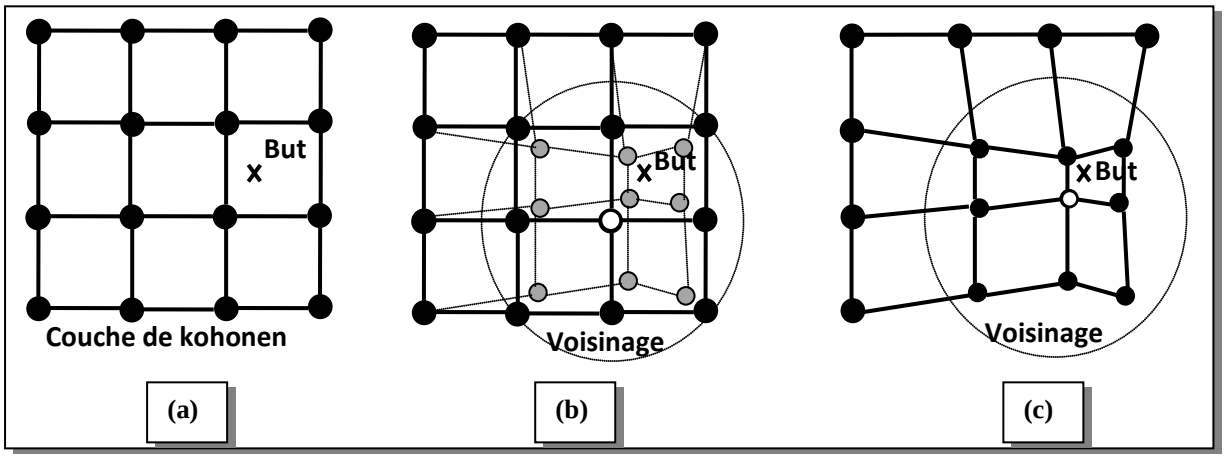


Figure 2.9-Illustration de l'apprentissage de la méthode SOM : (a) État initial, (b) état à l'étape k, (c) état à l'étape k+1.

Les cartes de Kohonen ont prouvé leur utilité pour la classification des bases de données multidimensionnelles traitant des problèmes non linéaire. Ces cartes sont capables d'extraire les propriétés statistiques des paramètres météorologiques présentes dans la base d'entrée et c'est la raison pour laquelle ce type de réseau a été choisi pour la présente étude. Afin d'obtenir de bon résultat, un apprentissage du réseau par des données statiquement représentatives de la totalité des données doit être achevé. Dans cette application les propriétés statistiques des données météorologiques ne sont pas claires donc la base de données entière est nécessaire pour une bonne modélisation.

2.3.1.6. Paramètres d'apprentissage

Un certain nombre de paramètres doivent être décidés avant la phase d'apprentissage : la taille de la carte (le nombre d'unités qui forment la carte) et ça forme, la fonction de voisinage, le rayon de voisinage, le taux et la longueur (le nombre d'époques) d'apprentissage.

On utilisant la librairie somtoolbox[1], l'utilisateur peut librement indiquer tous ces paramètres, mais pour réduire au minimum son effort, des valeurs par défaut sont fournis également à ces paramètres, ces valeurs par défaut sont [49] :

- le nombre des unités (neurone) de la carte peut être défini approximativement par l'heuristique: $m = 5\sqrt{n}$ 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 voisinage correspond au rapport entre deux plus grandes valeurs propres de la matrice de covariance des données.
- la fonction de voisinage par défaut est 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, aussi bien 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 : une époque correspond à un passage par les données. 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.

Par défaut, l'apprentissage est divisé en deux phases, La première phase est exécutée en utilisant un plus grand rayon de voisinage et taux d'apprentissage que la deuxième phase. Ainsi, elle est également plus courte que la deuxième phase.

2.3.1.7. Visualisation

La carte de kohonen peut être efficacement utilisée pour la visualisation de données cela est due principalement à sa capacité de rapprocher la densité de probabilité de l'ensemble de données et de les représenter dans deux dimensions [50] ; [51]. Par la suite, plusieurs méthodes de visualisation du réseau de kohonen sont présentées. Le neurone gagnant pour

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

chaque échantillon de la base est sélectionné sur la carte, et il est par la suite marqué par l'étiquette correspondante.

La méthode de la matrice des distances unifiées (u-matrice) de Ultsch [52] montre les distances entre les unités voisines et ainsi visualise la structure des clusters de la carte. Notons que l'u-matrice contient beaucoup plus d'hexagones que la matrice des composants. C'est parce que les distances entre les unités de la carte sont montrées, et non pas seulement évalués dans les unités de la carte. Les valeurs élevées sur l'U-matrice signifient une grande mesure de distance entre les unités voisines de la carte, et indiquent ainsi les frontières du cluster. Les clusters sont des régions en général uniformes de valeurs faibles de distance. (Se référer à la barre de couleur pour voir quelles couleurs signifient des valeurs élevées). La carte de la base de données IRIS montré par la figure 2.10 semble y avoir deux clusters.

Les cartes de distribution (« distribution maps » ou « component planes »), issues de la carte de kohonen. Sa visualisation peut être considérée comme une version découpée en tranches de la carte, où chaque « plane » montre la distribution du vecteur poids d'un composant. En utilisant ces distributions, différentes dépendances entre les paramètres de la base de données peuvent être étudiées. Par exemple, [53] ont utilisé ce genre de visualisation pour étudier les variations des paramètres de conception d'un circuit VLSI.

Les cartes de distributions de la base Iris sont montrées par la figure 2.10, ou les noms des composants sont inclus comme titres des figures secondaires. Notons également que les valeurs des variables ont été dénormalisées à la gamme originale. Les cartes de distributions ('PetalL ', 'PetalW ', 'SepalL 'et 'SepalW ') montrent quel genre de valeurs ont les vecteurs prototypes des unités de la carte. Les valeurs sont indiquées avec des couleurs, et la barre de couleur à la droite montre ce que signifient ces couleurs.

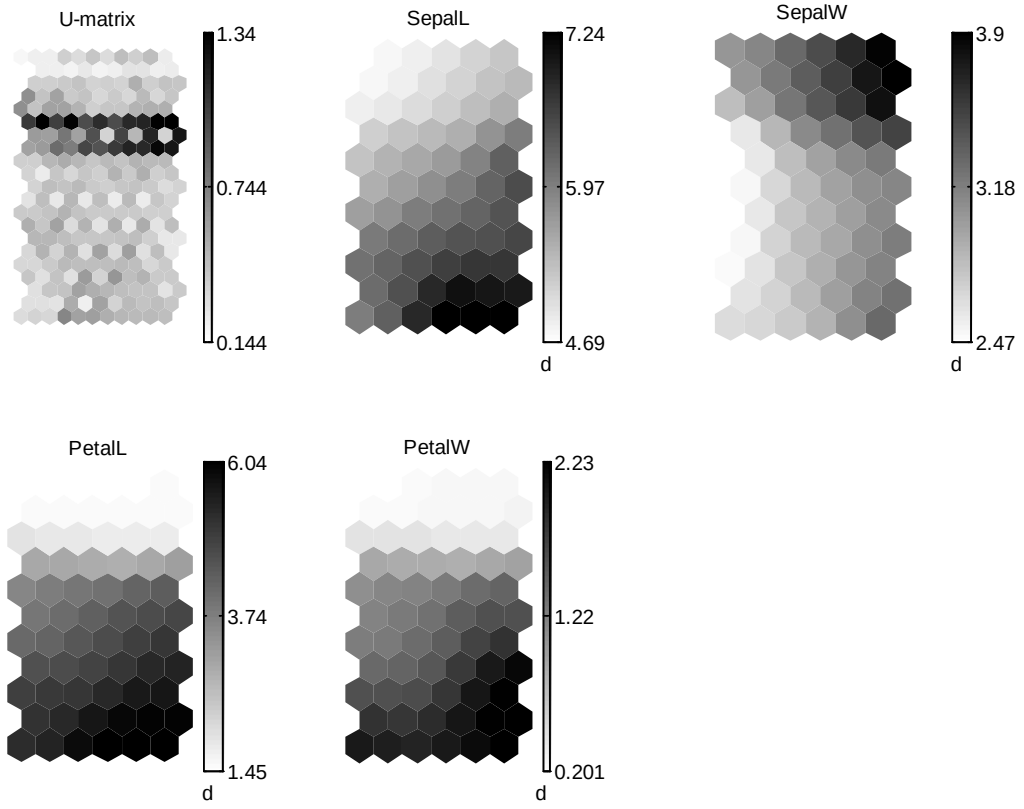


Figure 2.10--Différentes visualisation de la SOM : u-matrice et Les cartes de distribution.

2.3.2. Quelques approches classiques

Les deux approches principales de classification automatique de données (faire une partition) sont les méthodes hiérarchiques et partitives. Les méthodes hiérarchiques peuvent être encore divisées en algorithmes agglomératifs et divisifs, correspondant aux stratégies ascendantes et descendantes [28] ; [30] ; [31].

2.3.2.1. La classification par partition

Les algorithmes de classification par partition divisent directement un ensemble d'individus en k classes dont chaque classe doit contenir au moins un individu et que chaque individu doit appartenir à une classe unique contrairement à la classification dite floue qui n'impose pas cette condition. Pour ce faire, étant donné le nombre k de classes requises, ces algorithmes génèrent une partition initiale, puis cherchent à l'améliorer en réattribuant les individus d'une classe à l'autre. Ces algorithmes cherchent donc des maximums locaux en optimisant une fonction objective qui traduit que les objets doivent être similaire au sein d'une même classe, et dissimilaire d'une classe à l'autre. Contrairement aux méthodes hiérarchiques qui ne révisent pas les classes (intermédiaire) une fois qu'elles sont construites, les algorithmes de

classification par partition améliorent graduellement les clusters avec des données appropriées, ce qui permet la possibilité qu'une pauvre partition initiale pourrait être corrigée ultérieurement. La plupart des méthodes supposent que le nombre de groupes cherché doit être prédéfini a priori par l'utilisateur, bien que certaines méthodes permettent à ce nombre d'être changé pendant l'analyse, il peut également faire partie d'une fonction d'erreur [54]. 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.

Si le nombre de clusters cherché est inconnu, 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 $\sqrt{N}$, ou N est le nombre d'individus dans l'ensemble de données. La méthode la plus classique et qui reste très utilisée et celle des k-moyennes et ses nombreuses variations qui en découlent. Les classes sont représentées par leur centre, qui correspond 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 des méthodes hiérarchiques. Dans sa version la plus classique, l'algorithme consiste à sélectionner aléatoirement k individus initiales qui représentent les centroïdes initiaux. Un individu est assigné au cluster pour lequel la distance entre l'individu et le centroïde est minimale. Les centroïdes sont alors recalculés et on passe à l'itération suivante [31]. La version classique de cet algorithme présente l'avantage d'être très simple. Sa complexité algorithmique est également intéressante. De plus les classes obtenues sont facilement interprétables et représentées naturellement par les centroïdes. Cependant, ses inconvénients sont nombreux. Tout d'abord, il nécessite évidement de définir une moyenne entre les données dont le calcul de ces moyennes est très sensible aux données aberrantes. Ainsi, il est nécessaire de définir le nombre de clusters k, et le résultat de la classification est très dépendant du choix des centroïdes initiaux.

Pour résoudre notamment le problème de sensibilité aux données aberrantes lors de l'initialisation de l'algorithme des k-moyennes, 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 principale différence avec les k-moyennes se situe au niveau du choix du

représentant d'une classe. Dans les k-médoïdes chaque classe est représentée par un de ses membres appelé médoïdes et non plus par un centroïdes. La représentation par des k-médoïdes a deux avantages. D'abord, elle ne présente aucune limitation sur les types d'attributs manipulés, et en second lieu, le choix des médoïdes est dicté par l'endroit d'une fraction prédominante des points à l'intérieur d'un cluster et, par conséquence, il est moins sensible aux données aberrantes. Cependant, l'absence de centroide qui résume les données se fait au détriment de la complexité, puisqu'il apparaît d'après le descriptif précèdent de l'algorithme que chaque itération est en $O(k(n-k)^2)$. De plus comme pour les k-moyenne, il est nécessaire de spécifier le nombre de clusters k. afin de résoudre partiellement le problème des temps élevés de calcul avec les algorithmes de type PAM, l'algorithme CLARANS (Clustering Large Applications bases upon randomised search) est très souvent utilisé.

2.3.2.1.1. La méthode de k-moyennes

K-moyennes est la méthode de quantification vectorielle la plus connue, elle permet de rassembler en classes un ensemble de points de l'espace des observations sans que l'on dispose de connaissances a priori des propriétés particulières sur ces classes ; k-moyennes converge rapidement à un minimum local de son fonction de coût [55], [56], [57], la méthode détermine l'ensemble des vecteurs référents *W*; et la fonction d'affectation *X*, en minimisant la fonction de coût [39] :

$$l(w,x) = \sum_{x_i} \left\| z_i - w_{x(z_i)} \right\|^2 = \sum_{c} \sum_{x_i \in P_C \cap d} \left\| z_i - w_c \right\|^2 \qquad \textbf{(2.24)}$$

L'expression :

$$I_c = \sum_{z_i \in P_C \cap A} \left\| z_i - w_c \right\|^2$$

représente l'inertie locale, par rapport au référent $w_{\mathbf{c}}$, des observations de l'ensemble d'apprentissage *A* qui lui sont affectées; ces observations appartiennent donc au sous-ensemble P_c . L'inertie I_c représente l'erreur de quantification obtenue quand on décide de remplacer les observations de P_c par le référent $w_{\mathbf{c}}$ qui les représente. La quantité I (*W, X*) représente la somme des inerties locales I_c :

$$I(W,X) = \sum_{c} I_c = \sum_{c} \sum_{\substack{x_i \in A \\ X(x_i)=c}} \left\| z_i - w_c \right\|^2 \qquad \textbf{(2.25)}$$

L'algorithme des k-moyennes procède d'une manière itérative, chaque itération comportant deux phases. La première phase minimise la quantité I (*W, X*) : en supposant les valeurs des référents fixées aux valeurs calculées précédemment, elle calcule une valeur de la fonction *X*.

La seconde phase suppose que la fonction d'affectation est fixée à la valeur qui vient d'être calculée ; elle minimise alors la fonction $I(W, X)$ par rapport aux paramètres W. En procédant ainsi en deux phases, on fait décroître la valeur de $I(W, X)$ à chaque itération.

Une itération se résume donc de la manière suivante :

• *Phase d'affectation*. Il s'agit, dans cette phase, de minimiser la fonction $I(W, X)$ par rapport à la fonction d'affectation X ; à cette étape, l'ensemble W des référents est fixé. La minimisation s'obtient en affectant chaque observation z_i au référent w_c à l'aide de la fonction d'affectation X :

$$X(z) = \arg\min_{r} \|z - w_r\|^2 \qquad \textbf{(2.26)}$$

Où r varie de 1 à p.

• *Phase de minimisation*. La seconde phase de l'itération fait décroître à nouveau la quantité $I(W, X)$ en fonction de l'ensemble des référents W; la fonction d'affectation X utilisée à la phase précédente est fixée. Le minimum global pour la fonction $I(W, X)$ est atteint pour :

$$\frac{\delta I}{\delta W} = \left[\frac{\delta I}{\delta w_1}, \frac{\delta I}{\delta w_2}, \ldots \frac{\delta I}{\delta w_p} \right]^T = 0 \qquad \textbf{(2.27)}$$

Les p nouveau référents sont alors définit par :

$$w_c = \frac{\sum_{z_i \in P_c \cap A} z_i}{n_c} \qquad \textbf{(2.28)}$$

2.3.2.1.2. La méthode k-médoïdes

Dans les méthodes k-médoïdes un cluster est représenté par un de ses points. Nous avons déjà mentionné que c'est une solution facile puisqu'elle couvre n'importe quels types de données et que les médoïdes ne sont pas sensible aux données aberrantes. Après avoir sélectionner les médoïdes, les clusters sont définis comme le sous-ensemble de points respectivement le plus proche au médoïdes, et la fonction objective est définie par la distance moyenne ou une autre mesure de dissimilarité entre un point et son médoïdes. La médoïdes d'un groupe est un individu présent dans le groupe possédant la dissimilarité moyenne la plus faible avec les autres individus du groupe [31].

Deux versions récentes des méthodes k-médoïdes sont l'algorithme PAM (Partitioning Around Medoids) et l'algorithme CLARA (Clustering LARge Applications) [47]. 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 les méthodes k-moyennes en présence du bruit cependant cette méthode est d'une complexité de calcul élevé pour chaque itération, en effet cette méthode est très efficace dans le traitement de petites bases de données et très coûteuse en cas de bases de données volumineuses, ce qui a conduit les chercheurs à proposer d'autres algorithmes tel que CLARA pour traiter les données multidimensionnelles.

CLARA est un algorithme introduit par [58] 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 obtenu est le meilleur parmi les différents résultats. L'inconvénient principal de cette méthode est que les paramètres d'échantillonnage sont choisis expérimentalement.

Pour classer les données de grande taille une méthode appelée CLARANS a été proposée par [59]. Cette méthode propose d'utiliser une recherche stochastique basée sur différents paramètres permettant de borner le nombre d'itération de la méthode, ainsi que sur l'échantillonnage aléatoire. Étant donné le nombre *k* de clusters recherchés, une solution consiste en un ensemble de *k* médoïdes, objets représentatifs des clusters, auxquels sont associés l'ensemble des objets en fonction de leur proximité avec ces médoïdes [12]. Les étapes principales de la méthode sont les suivantes :

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 $O(k.n^2)$.

2.3.2.2. La classification hiérarchique

Historiquement, elles furent les premières développées, principalement en raison de la simplicité des calculs. L'avènement des puissants ordinateurs leur a fait perdre une certaine popularité au profit des méthodes non hiérarchiques. Les classifications hiérarchiques se présentent comme la succession de partitions emboîtées et peuvent être représentées graphiquement à l'aide de dendrogrammes tels que celui de la figure 2.12. On distingue deux types d'approches de classification hiérarchique : les méthodes descendantes –divisives- et les méthodes ascendantes –agglomeratives- selon la façon dont le dendrogramme hiérarchique est formé [60], [61].

La **C**lassification **A**scendante Hiérarchique (CAH) permet de construire une hiérarchie entière des objets sous la forme d'un arbre dans un ordre ascendant. Cette méthode considère les singletons (les classes formées uniquement d'un seul individu) et procède 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 tous les individus. Cette classification génère un arbre que l'on peut couper à différents niveaux selon la mesure de distance sélectionnée pour obtenir un nombre de classes plus au moins grand. Différentes mesures de distances interclasses peuvent être utilisées : la distance euclidienne, la distance inférieure (qui favorise la création de classes de faible inertie) ou la distance supérieure.

La classification descendante hiérarchique considère tous les individus comme une seule classe et procède par division successive jusqu'à l'obtention des classes formées uniquement d'un seul individu. Cette méthode est très coûteuse pour être utilisées sur les volumes de données manipulés aujourd'hui. En effet, la division d'une partie à N élément nécessite l'évaluation des $(2^{N-1} - 1)$ divisions possibles.

2.3.2.2.1. La classification hiérarchique ascendante

Le principe de construction des méthodes ascendantes est d'élaborer, pas à pas, une suite de partitions emboîtées depuis la partition la plus fine (formée des n singletons {x}, i = 1, 2, ..., n) jusqu'à la partition la plus grossière ({X}). On commence par agréger les deux individus les plus proches, il ne reste donc plus que $n - 1$ objets (les deux premiers individus regroupés sont considérés comme un nouvel élément), et on itère cette opération jusqu'à ce que tous les

éléments aient été traités. Une telle recherche s'appuie sur une mesure de distance entre individus qui quantifie l'hétérogénéité d'une partie basée sur les distances entre individus qui sont dedans et une mesure de dissimilarité entre deux parties basée sur la distance entre un individu de l'un et un individu de l'autre classe. La distance euclidienne est généralement utilisée pour les points individuels. Il n'y a aucun critère connu permettant de distinguer le type de distance le plus adéquat à utiliser pour telle base de données. Après avoir regrouper les deux individus les plus proches au sens de la dissimilarité de départ on peut regrouper soit des individus, soit un individu et une classe, soit, un peu plus tard, deux classes. Parmi les variations les plus utilisées de la classification hiérarchique basé sur les différentes critères d'agrégation [31] ; [60] ; [61] :

3.3.1. Lien simple ou *single linkage* ou saut minimum ou *nearest neighbour*

Le lien simple ou lien du saut minimum ou celui du plus proche voisin consiste à choisir la plus petite distance séparant deux éléments appartenant à deux groupements disjoints différents. L'agrégation selon le saut minimal est, peut-être, le premier algorithme de classification automatique mis en œuvre pour la description de données multidimensionnelles [56]. La distance entre deux parties est définit par :

$$D(A,B) = \min_{a\in A, b\in B}(d(a,b)) \quad \textbf{(2.29)}$$

2.3.2. Lien complet ou *complete linkage* ou agrégation par le diamètre

La CAH à lien complet (*complete linkage*) ou lien d'agrégation par le diamètre ou *furthest neighbour* est exactement la même procédure que la précédente pour une nouvelle distance entre parties dérivée de la distance entre individus. La distance entre deux parties est définie par :

$$D(A,B) = \max_{a\in A, b\in B}(d(a,b)) \quad \textbf{(2.30)}$$

2.3.3. Lien moyen ou *average linkage* ou UGPMA ou *group average*

La dissimilarité entre clusters est calculée en utilisant des valeurs moyennes. La distance moyenne est calculée de la distance entre chaque point d'un cluster et tous autres points dans un autre cluster. Les deux clusters dont la distance moyenne est la plus basse sont joints ensemble pour former un nouveau cluster.

Average linkage, le lien moyen ou lien d'agrégation UGPMA (Unweighted Pair Group Method of Agregation) définit la distance entre deux parties par :

$$D(A,B) = \underset{a\in A, b\in B}{moyenne}(d(a,b)) \quad \textbf{(2.31)}$$

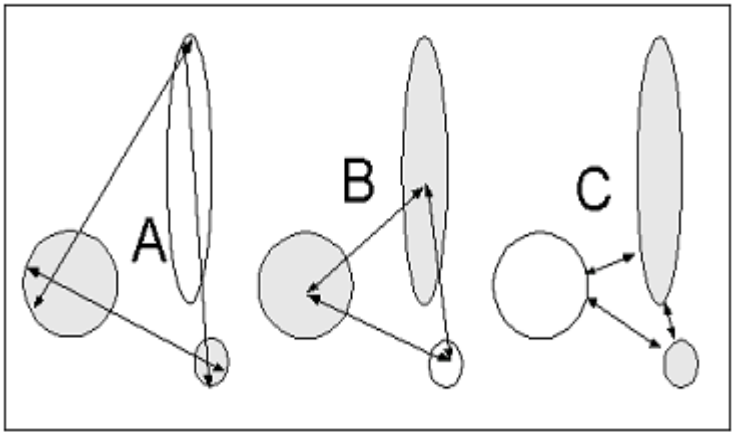


Figure 2.11-CAH et distances entre parties. A lien du diamètre. B lien moyen. C lien du saut minimum. (la distance minimum).

Le critère de Ward

Le critère de Ward est la mesure de similarité la plus connue, il consiste à faire des regroupements 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 [57]. Ce critère est définit par [31] :

$$D(C_1, C_2) = \frac{n_{C1} n_{C2}}{n_{C1} + n_{C2}} d^2(g_{C1}, g_{C2}) \qquad \textbf{(2.32)}$$

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.

$$D(C_1, C_2) = d(g_{C1}, g_{C2})$$

La difficulté du choix du critère d'agrégation réside dans le fait que ces critères peuvent déboucher sur des résultats différents. Selon la plus parts des références le critère le plus couramment utilisé est celui du Ward.

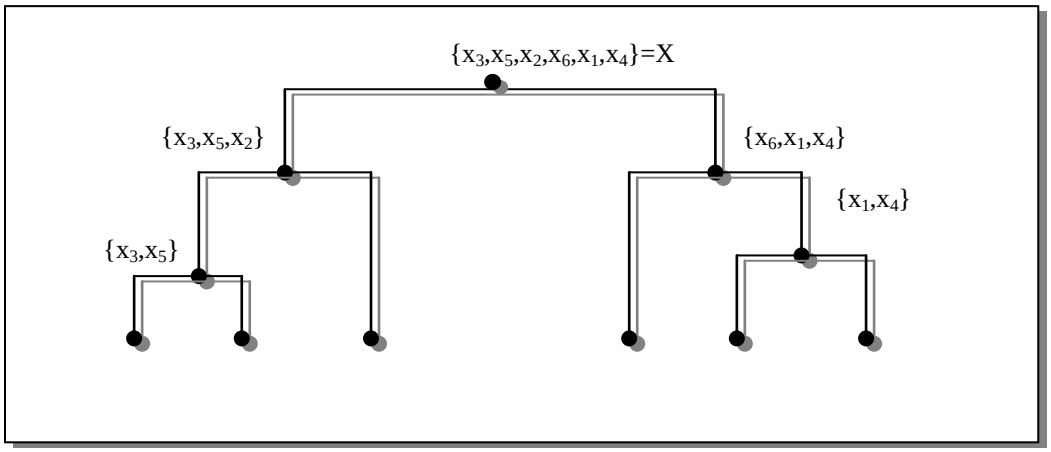


Figure 2.12--Partitions emboîtées d'un ensemble X à 6 éléments. En coupant l'arbre par une droite horizontale, on obtient une partition d'autant plus fine que la section est proche des éléments terminaux.

La figure 2.13 fournit une illustration géométrique des partitions emboîtées de la figure 2.12 dans le cas où les éléments à classer sont situés dans un espace bidimensionnel. Pour des dimensions plus élevées, c'est la représentation en arbre (figure 2.12) qui est généralement utilisée.

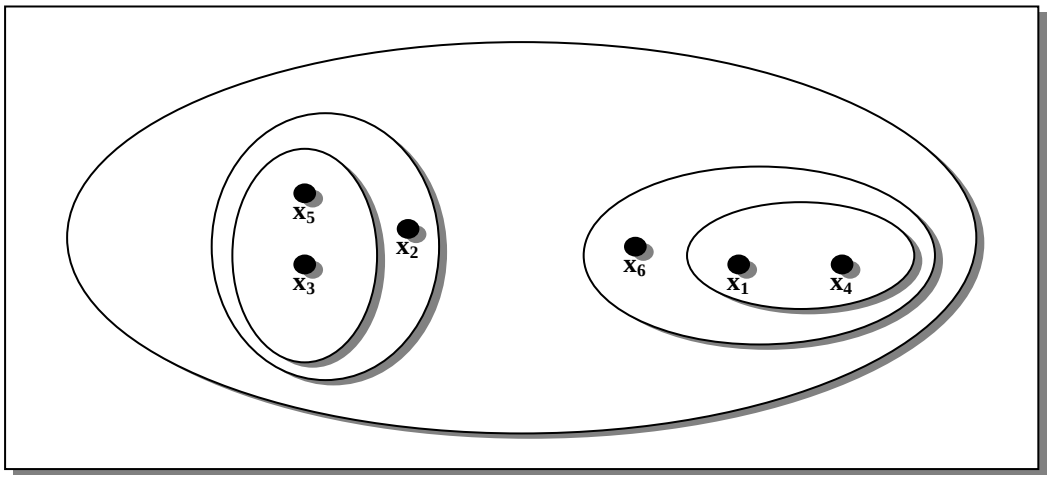


Figure 2.13--Représentation géométrique des partitions emboîtées de la figure 2.12. (Cas particulier de données planes).

La classification ascendante 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 qui sont basés sur le principe de la CAH 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).

2.3.2.2.2. La classification descendante hiérarchique

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'a obtenir une partition formée uniquement de singletons. A l'inverse de la classification ascendante hiérarchique, à chaque étape de l'algorithme il y a deux processus à faire [28] ; [31] :

1- chercher une classe à scinder.
2- choisir une méthode d'affectation des objets aux sous classes.

Pour un sous ensemble de *n* individus, la hiérarchie est construite en *n-1* étapes. Dans la première étape, les données sont divisées en deux classes au moyen de la dissimilarité. Dans chaqu'une des étapes suivantes, la classe avec le diamètre le plus grand se divise de la même façon. Après *n-1* divisons, tous les individus sont bien séparés. La dissimilarité moyenne entre l'individu *x* qui appartient à la classe *C* qui contient *n* individus et tous les autres individus de la classe *C* est définie par:

$$d_x = \frac{1}{n-1} \sum_{x \in C, y \neq x} d(x, y) \qquad \textbf{(2.33)}$$

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 a proposé 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 nombreuses inconvénient la classification descendante présente quelques avantages par rapport à la plus part des 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 dans une direction non réaliste

2.3.3. Comparaison des algorithmes de la classification automatique

La classification automatique apparaît comme un outil très efficace dans beaucoup d'applications. En effet l'approche de regroupement est adressée par les chercheurs de nombreuses disciplines. Cependant, cette classification demeure un problème difficile, qui combine des concepts de divers domaines scientifiques (tel que les bases de données, apprentissage artificiel, reconnaissance des formes, statistiques). Ainsi, les différences dans les prétentions et contexte parmi différentes communautés de recherches ont causé un certain nombre de méthodologies et d'algorithmes de regroupement différents [31] ; [35] ; [35].

Les méthodes partitives sont meilleures par rapport aux méthodes hiérarchiques dans le sens qu'elles ne dépendent pas des clusters précédemment trouvés. D'autre part, les méthodes

partitives font des estimations implicites sur la forme de clusters. Par exemple, k-moyennes essais pour trouver les clusters sphériques.

Les algorithmes partitifs sont principalement applicables aux données numériques. Cependant, il y a quelques variantes de l'algorithme K-moyennes tel que K-mode, qui manipule des données catégorielles. K-Mode est basé sur la méthode K-moyennes 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. Une autre caractéristique des algorithmes partitive est qu'ils ne peuvent pas manipuler des données bruités et ils ne sont pas bien appropriés pour découvrir les clusters avec des formes non convexes. Le résultat du processus du regroupement est l'ensemble des points représentatifs des clusters découvert. Ces points peuvent être les centres ou les médoïdes des clusters selon l'algorithme de regroupement (l'objet le plus centralement localisé dans un cluster). En ce qui concerne les critères de regroupement, l'objectif des algorithmes partitifs est de réduire au minimum la distance des objets dans un cluster du point représentatif de ce cluster. Ainsi, K-moyennes vise à minimiser la distance des objets appartenant à un cluster au centre de ce cluster (medoïdes pour PAM.). CLARA et CLARANS sont basés sur le critère regroupement de la PAM. Cependant, ces algorithmes considèrent des échantillons de données sur lesquels un regroupement est déjà appliqué et par conséquent ils peuvent traiter de plus grande base de données que la PAM.

La méthode des cartes auto-organisatrices peut être vue comme une extension de l'algorithme des k-moyennes : comme lui, il minimise une fonction de coût convenablement choisie. Cette fonction de coût doit tenir compte, d'une part, de 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-moyennes 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 classification de Kohonen a des analogies avec certaines méthodes présentées précédemment et des propriétés qui permettent d'envisager d'éventuels couplages avec elles. La classification 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 trop extravagant (valeur erronée ou 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 de l'initialisation. 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 [62], ou carte de Kohonen – classification hiérarchique.

Les algorithmes de la classification hiérarchique créent une décomposition hiérarchique de la base de données représentée par un dendrogramme. Ils sont plus efficaces en manipulant le bruit que les algorithmes partitifs. Cependant, cette méthode est très coûteuse en raison de sa complexité temporelle (typiquement, complexité $O(n^2)$, où n est le nombre de points dans l'ensemble de données).

2.4. Évaluation et critères de validités

Une des étapes les plus importantes dans l'analyse de clusters est l'évaluation des résultats pour identifier le meilleur partitionnement de l'ensemble de données. En effet c'est le principal sujet pour la validation des clusters. Par la suite nous discuterons les concepts fondamentaux et les différentes approches de la validité des clusters proposées en littérature. Les méthodes de la classification non supervisée devront chercher des groupes dont les membres sont proches (avoir un degré élevé de similarité) et sont bien séparés. Un autre problème auquel nous devons faire face dans le regroupement est de décider le nombre optimal des groupes qui représente mieux les données, donc il est naturel de s'interroger sur la validité et la qualité de la partition obtenue. Dans la plupart des algorithmes de classification des évaluations expérimentales des données à 2D sont utilisés pour que l'utilisateur puisse vérifier visuellement la validité des résultats (c.-à-d., à quel point l'algorithme de classification a découvert les clusters des données d'entrées). Il est clair qu'une visualisation d'un ensemble de données permet une vérification cruciale des résultats d'une classification automatique. Dans le cas des données multidimensionnelles (plus de trois dimensions), la visualisation efficace des données serait difficile [35]. D'ailleurs la perception des clusters à l'aide des outils disponibles de visualisation est une tâche difficile pour les être humains qui ne sont pas accoutumés aux espaces multidimensionnels. Les divers algorithmes de classification se comportent de manière différent selon :

- Les caractéristiques des données d'entrée (géométrie et densité de distribution des clusters),
- Les valeurs des paramètres d'entrées.

Par exemple, si on prend l'ensemble de données montré par la figure 2.14 (a). Il est évident que nous pouvons découvrir trois clusters [35]. Cependant, si on considère un algorithme de classification non supervisée (par exemple K-moyennes) avec certaines valeurs de paramètre (le nombre de clusters par exemple), afin de partitionner l'ensemble de données en quatre clusters, le résultat du processus de partitionnement est l'ensemble de groupes présenté par la figure 2b. Dans cet exemple l'algorithme K-moyennes a trouvé les meilleurs quatre clusters dans lesquels l'ensemble des données peut être divisé. Cependant, ce n'est pas la division 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-d., un regroupement) qui ajuste le mieux que possible les partitions inhérentes de l'ensemble de données.

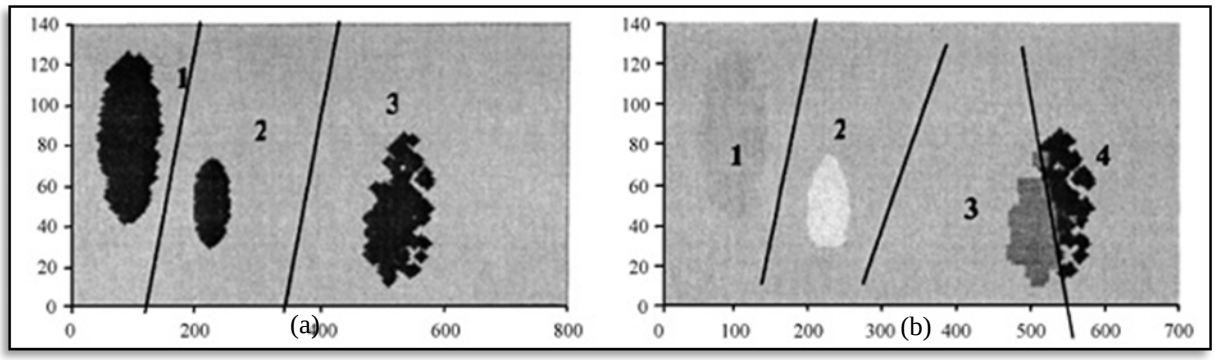


Figure 2.14--(a) ensemble de données qui se compose de trois clusters, (b) le résultat de regroupement des données par k-moyennes (avec comme paramètre d'entrée : le nombre de clusters recherchée=4).

Il est évident que le regroupement représenté par la figure 2.14 (2b) n'est pas convenable pour l'ensemble de données c.-à-d., la classification automatique représentée dans la figure 2b n'ajuste pas bien la base de données. Le regroupement optimal pour cet ensemble de données est un schéma avec trois clusters. Par conséquent, si les valeurs assignées 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 mène 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 [55] ; [63] ; [64] ; [66].

Par la suite, nous discuterons les concepts fondamentaux de la validité de regroupement et nous présentons les critères les plus importants dans le contexte d'évaluation et de validité des résultats d'une classification automatique.

2.4.1. Concepts fondamentaux de la validité des clusters

La procédure d'évaluation des résultats d'un algorithme de classification automatique est connue sous le nom "validité des clusters" ou (cluster validity). D'une façon générale, il y a trois approches pour étudier la validité des clusters [67]. La première approche est basée sur des critères externes. Ce qui implique que nous évaluons les résultats d'un algorithme de regroupement basé sur une structure pré-spécifiée, cette structure est imposée à un ensemble de données et reflète notre intuition au sujet de la structure du regroupement des données. La deuxième approche est basée sur des critères internes. Nous pouvons évaluer les résultats d'un algorithme de regroupement en termes de quantités des vecteurs de l'ensemble de données (ex : matrice de proximité). La troisième approche de validité des clusters est basée sur les critères relatifs. L'idée basique pour cette dernière approche est l'évaluation d'une structure de regroupement en le comparant à d'autres regroupements, résultant par le même algorithme mais avec des paramètres différents. Deux critères sont généralement utilisés pour l'évaluation et le choix du regroupement optimal :

Compacité : les membres de chaque cluster doivent être proches que possible l'un à l'autre. Une mesure commune de la compacité est la variance, qui doit être réduit au minimum.

Séparabilité : les clusters doivent être largement séparés. Il y a trois approches communes mesurant la distance entre deux clusters différents : distance entre les membres les plus proches des clusters (*Single linkage*), distance entre les membres les plus éloignés (*Complete linkage*) et distance entre les centres des clusters (*Comparison of centroids*).

Les deux premières catégories d'indices de validité des clusters sont basées sur des testes statistiques et leurs inconvénient principal est le coût de calcul élevé. D'ailleurs, les index liés à ces deux approches visent à mesurer le degré auquel un ensemble de données confirme un schéma de regroupement apriori. D'autre part, la troisième approche vise à identifier le meilleur regroupement pour un algorithme de classification défini pour certains paramètres. Un certain nombre d'index de validité ont été proposés et définis en littérature pour chacune des approches décrite ci-dessus [63] ; [68].

2.4.1.1. Erreur quadratique moyenne

L'erreur quadratique moyenne -*Mean Squared Error, MSE*- est une mesure de compacité très répandue, elle est notamment équivalente à la fonction de coût de l'algorithme de k-moyenne décrite précédemment [30] :

$$MSE = \frac{1}{N} \times \sum_{i=1}^{N} \sum_{j=1}^{K} c_{ij} \times \left\| x_i - w_j \right\|^2 \quad \textbf{(2.34)}$$

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

2.4.1.2. 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 [70]. Cet indice favorise les groupes hypersphériques et il est donc particulièrement bien adapté pour une utilisation avec la méthode des k-moyennes [30].

$$I_{DB} = \frac{1}{K} \sum_{K=1}^{K} \max_{j \neq K} \left\{ \frac{S_c(C_K) + S_c(C_j)}{D_{ce}(C_K, C_j)} \right\} \quad \textbf{(2.35)}$$

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:

$$S_c(C_i) = \frac{1}{N_i} \sum_{i=1}^{N_i} \left\| x - w_i \right\| \quad \textbf{(2.36)}$$

$$D_{ce}(C_i, C_j) = \left\| w_i - w_j \right\| \quad \textbf{(2.37)}$$

2.4.1.3. 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 étroit. 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 [58] ; [65].

$$a(i) = \frac{\sum_{j \in A, j \neq i} d(i, j)}{|A| - 1} \quad \textbf{(2.38)}$$

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

$$d(i, C) = \frac{1}{|C|} \sum_{J \in C} d(i, j) \quad \textbf{(2.39)}$$

et

$$b(i) = \min_{C \neq A} \{ d(i, C) \ \} \tag{2.40}$$

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

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.41}$$

ce qui est 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 :

$$cluster\_silhouette = \frac{\sum_{i=1}^{|A|} s(i)}{|A|} \tag{2.42}$$

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

$$general\_silhouette = \frac{1}{c} \sum_{j=1}^{c} cluster\_silhouette_i \tag{2.43}$$

Le nombre optimal de clusters est une valeur *q* dans laquelle la valeur de la silhouette générale est la plus grande. Plus la valeur de silhouette est grande, plus la qualité du cluster est meilleure.

2.4.1.4. Homogénéité et séparation

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

$$H_{ave} = \frac{1}{N} \sum_{i} d(i, C(i)), \tag{2.44}$$

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 [66].

$$S_{ave} = \frac{1}{\sum_{i \neq j} N_{Ci} N_{Cj}} \sum_{i \neq j} N_{Ci} N_{Cj} D(C_i, C_j) \tag{2.45}$$

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 du variance intra-clusters et la variance inter-clusters est une constante.

2.4.1.5. La méthode évolution de système

Cette méthode analyse l'ensemble de données comme un système pseudo thermodynamique. Pour avoir une bonne séparation des groupes, l'énergie de la partition E_p (k) dénote la distance de frontière entre deux groupes les plus étroits (appelés groupes jumeaux) parmi les k groupes, alors que le fusionnement de l'énergie E_m (k) dénote la distance moyenne entre les éléments dans la région de frontière pour plus de détail voir [70].

2.4.1.6. Indice inter-intra poids

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 [71].

2.4.1.7. Indices propres aux cartes auto-organisatrices

Plusieurs indices de qualité ont été développés pour les cartes auto-organisatrices, nous n'introduisons ici que les plus utilisés [30].

2.4.1.7.1. Erreur de quantification

Les cartes auto-organisatrices font partie des méthodes de quantification vectorielle, donc il est naturel de les évaluer à l'aide de l'erreur de quantification moyenne qui est définit ainsi :

$$Q_{err} = \frac{1}{N} \times \sum_{i=1}^{N} \left\| x_i - w_{b(i)} \right\| \qquad \textbf{(2.46)}$$

Ou *b(i)* est l'indice du prototype le plus proche de l'observation x_i.

2.4.1.7.2. Taux d'erreur topologiques

Les cartes de kohonen sont aussi une méthode de projection de données sur un espace de faible dimension, donc le taux d'erreurs topologique permet de quantifier la topologie locale des données par la carte. On considère qu'il y a une erreur topologique chaque fois que les deux prototypes les plus proches d'une observation ne sont pas voisins sur la carte. Le taux d'erreur topologique est définit par :

$$T_{err} = 1 - \frac{1}{N} \times \sum_{i=1}^{N} \frac{1}{N(b(i))} \left(\arg\min_{j \neq i} \left\| x - w_j \right\| \right) \qquad \textbf{(2.47)}$$

Où $\frac{1}{N(b(i))}$ est la fonction indicatrice de l'ensemble des voisins du prototype le plus proche de l'observation x_i.

2.4.1.7.3. Mesure de distorsion

La mesure de distorsion permet de prendre en considération à la fois la qualité de la quantification et la conservation de la topologie locale et elle s'apparente à l'erreur quadratique floue ou les degrés d'appartenance seraient remplacés par la fonction de voisinage :

$$distortion = \sum_{i=1}^{N} \sum_{J} h_{b(i)j} \times \left\| x - w_j \right\|^2 \quad \textbf{(2.48)}$$

Où $h_{b(i)j}$ est la fonction de voisinage.

2.5. Analyse de données

L'évolution des produits logiciels et des matériels informatique a permis le stockage et la gestion de grandes bases de données. L'accroissement exponentiel des volumes de données a rendu difficile la gestion et l'extraction d'informations utiles sans traitement au préalable. Les méthodes d'analyse de données se sont alors développées. Elles représentent un ensemble de méthodes pour la synthèse, l'extraction des connaissances à partir de grandes bases de données généralement à travers une représentation graphique particulière de ces données.

Plusieurs méthodes et outils ont été développés depuis les années soixante favorisés par les quantités gigantesques des bases de données [31].

2.5.1. Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode d'analyse de données ancienne et très utilisée en statistique et dans la réduction de dimension des données multidimensionnelles par projection [23]. L'ACP est largement utilisée dans les applications météorologiques tel que [72] ; [73] ; [74] ; [75].

L'idée fondamentale de l'ACP est la suivante : considérant le nuage de *N* points en *P* dimensions (dans cet espace, 1 point = 1 individu), on cherche trouver le plan (donc, une représentation bidimensionnelle que l'on va pouvoir voir et comprendre) dans lequel la projection des points du nuage est la moins déformée possible [76]. L'objectif principal de l'ACP est de construire l'espace euclidien à *P* dimensions le plus caractéristique et le plus

économique pour représenter ces points. L'ACP vise donc de passer d'un espace de données à un espace de caractéristiques (feature space) ou espace factoriel. La matrice de transformation est construite autour des vecteurs propres de la matrice de corrélation d'entrée, ces vecteurs sont ordonnés selon leurs valeurs propres. Si l'on considère l'ensemble de données compilé dans une matrice $A = X_N^P$, ou les *N* individus sont décrits par *p* variables, prenons $\vec{x}_i$ comme le moyen de *P* variables dans la matrice *A*. La matrice de covariance est donnée par :

$$\phi_{JK} = \frac{1}{n}\sum_{i=1}(x_{ji} - \bar{x}_j)(x_{ik} - \bar{x}_k),\ j \neq k \qquad \textbf{(2.49)}$$

ou j=1, 2…, n, k=1, 2,…, p.

L'ACP est une méthode linéaire. Donc seulement les corrélations linéaires sont détectées et exploitables. Les ACP ont été étendues pour attaquer le cas ou les relations entre les caractères sont non linéaires. Il y a au moins trois classes de méthodes :

- méthodes de projection non linéaire (par exemple, l'analyse en composante curviligne) [77];
- Réseaux de neurones non supervisés (carte de kohonen) ;
- ACP noyau reposant sur les machines à vecteurs de supports.

2.5.2. Analyse factorielle des correspondances

L'analyse factorielle des correspondances a été proposée dans les années '60 par Benzecri. Dans le cas d'une AFC les données sont transformées afin de mettre en évidence la répartition relative de l'individu par rapport aux variables et d'établir les corrélations entre les profils obtenus. En effet L'analyse factorielle des correspondances (AFC) est une ACP pour étudier les tableaux de contingence : on considère des individus décrits par deux caractères nominaux. On construit le tableau de contingence dans lequel les lignes représentent les valeurs prises par l'un des caractères, les colonnes représentent les valeurs prises par l'autre caractère. A l'intersection, on trouve le nombre d'individus pour lesquels on a conjonction de ces deux valeurs de caractères [31] ; [76].

L'AFC a été étendu à plus de deux caractères, ce qui a donne l'analyse factorielle des correspondances multiples (AFM). Cette méthode est très utilisée pour l'analyse d'enquêtes auprès de l'opinion (sondage).

2.5.3. Analyse factorielle des correspondances multiples (AFM)

L'analyse des correspondances peut se généraliser de plusieurs façons au cas ou plus de deux ensembles sont mis en correspondance. L'analyse des correspondances multiples permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple privilégié [31]. Les lignes de ces tableaux sont en général des individus ou observations et les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions. L'analyse des correspondances multiples est une analyse des correspondances simple appliquée non plus à une table de contingence, mais à un tableau disjonctif complet. l'AFM consiste alors à réaliser une analyse factorielle (AFC ou ACP) soit à partir de la concaténation des différents tableaux, soit à partir d'un des tableaux, les éléments des autres tableaux étant alors considérés comme des éléments supplémentaires.

2.5.4. Analyse de données en utilisant les cartes de Kohonen

Les projections sur des plans produits par les analyses factorielles ne sont pas adaptées à des bases de données qui nécessitent une représentation dans des espaces de dimension trop grande (supérieure à 4). Cela pose des problèmes dans la pratique de visualisation et de synthèse. Ces bases de données multidimensionnelles ont poussé les chercheurs à améliorer les méthodes existantes en les complétant. Tel que très bien détaillé dans [45], le couplage analyse factorielle –classification n'optimise pas la représentation de la classification et inversement la classification n'optimise pas l'explication de la représentation. Alors que c'est ce que fait l'algorithme de Kohonen – où la représentation et la classification sont jumelées et il apporte ainsi un plus par rapport aux autres méthodes dans le traitement de grande masse de données. De plus cette méthode offre à la fois une grande facilité d'interprétation, une grande souplesse d'utilisation – on peut choisir la distance et donc l'adapter à un cadre d'étude assez large – et permet des représentations graphiques de nombreuses sortes de bases de données en respectant leur topologie. Ces représentations peuvent rendre cette méthode aussi attractive et souvent mieux adaptée que les méthodes linéaires.

Chapitre 3

Identification des types de jours météorologique : Approche proposée.

3.1. Approche proposée

Dans le but de l'identification des types de jours météorologiques pour la région d'Annaba, Nous avons proposé une approche basée sur la classification automatique -clustering-. Cette approche permet d'identifier les groupes d'objets similaires (types de jours -clusters-) à partir des données météorologiques collectées par la station de l'aéroport d'Annaba. L'approche proposée est basée sur deux niveaux de classification : la carte auto-organisatrice de Kohonen (SOM) est utilisée comme un premier niveau de classification et les méthodes de classification par partition dans le deuxième niveau. La base de données utilisée pour l'identification des situations météorologiques dans la région d'Annaba est volumineuse, ce qui a conduit à la génération d'un grand nombre de vecteurs référents (U=180 vecteurs prototypes). Un niveau de classification plus grossier peut être également révélateur. Un niveau élevé est intéressant, car il fournit une qualité d'analyse plus fine et comprime moins l'information que si l'on résume l'ensemble des individus par les représentants d'un petit nombre de classes. Par contre, une classification qui cumule 2 niveaux de regroupement permet d'avoir une analyse plus fine que celle qui a un grand nombre de classes. Elle est de ce fait plus efficace qu'une classification à un seul niveau de regroupement ou que deux qui ne s'emboîtent pas [54]. L'approche proposée est présentée par la figure 3.1.

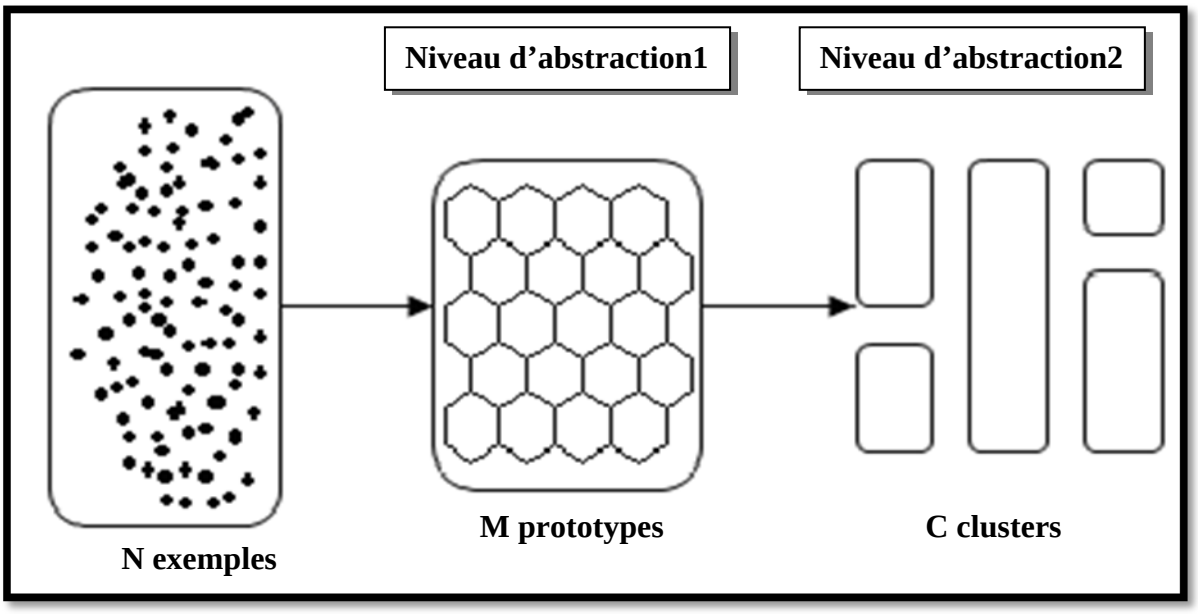


Figure 3. 1-Le premier niveau d'abstraction est obtenu par la création d'un ensemble de vecteurs prototypes en utilisant SOM. Le regroupement des résultats du SOM permet de créer le deuxième niveau d'abstraction.

Pour découvrir la méthode la plus appropriée à utiliser pour regrouper les unités de la SOM, nous avons procédé à une comparaison basée sur les performances des algorithmes de classification les plus utilisés tel que : l'algorithme PAM (Partitioning Around Medoids), K-moyennes, et la classification hiérarchique (méthode de Ward).

Le premier avantage de l'approche proposée est la diminution du coût de calcul. La classification des données météorologiques uniquement par k-moyennes 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 intraitable et très lourd, pour cette raison, il est nécessaire de regrouper un ensemble de prototypes plutôt que regrouper directement les données. Considérant un regroupement de N échantillons de données en utilisant k-moyennes. 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 pré-sélectionné de cluster. Si on utilise un ensemble de prototypes comme étape intermédiaire, la complexité total est proportionnel à $NM + \sum_k MK$, ou M est le nombre de prototypes [54]. 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 cluster 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.

3.2. Extraction des caractéristiques par l'ACP

L'analyse en composantes principales (ACP) est une méthode d'analyse de données largement utilisée pour l'extraction des caractéristiques météorologiques [74]. Les résultats de projection des données météorologiques en utilisant l'ACP-VARIMAX sont présentés par la figure 3.2. Seulement les composantes principales (CPs) associées à des valeurs propres supérieurs à l'unité (>1) ont été extraites. Huit CPs dont les valeurs propres sont supérieures à

1 comptent pour 99% de la variation des données. La dimension mathématique des données peut donc être réduite de 32 à 8 en utilisant l'ACP. La figure 3.3 montre les deux premières composantes principales et leurs paramètres météorologiques.

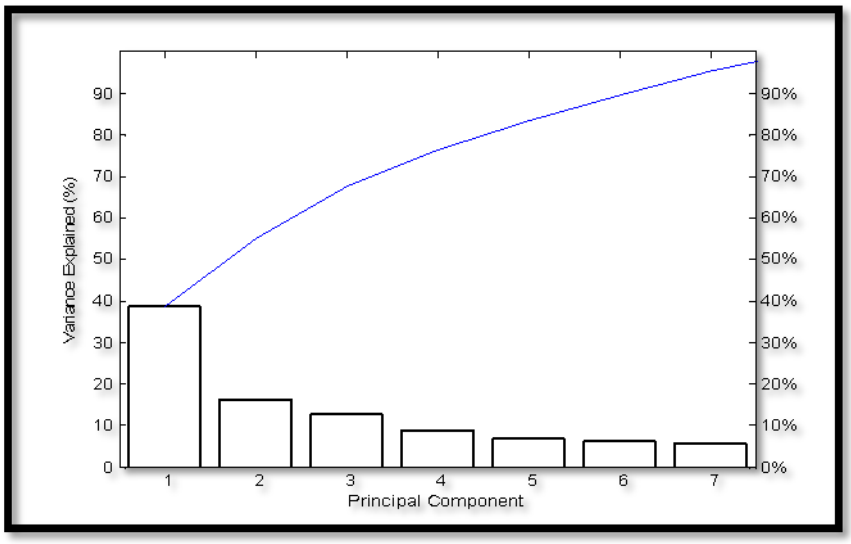


Figure 3. 2-Les sept premières composantes principales et leurs variances.

La première composante principale (CP1) exprime 38,707 % de la variance totale, cette composante est principalement caractérisée par une température élevée durant toute la journée avec une basse mesure d'humidité dans la matinée et une vitesse de vent un peu élevée dans la nuit. (CP2) exprime 16,354 % de la variance, caractérisée par une pression élevée durant les 24h et d'un taux d'humidité un peu élevé. (CP3) est également caractérisée par des grandes mesures d'humidité durant la première demi journée. CP4 est positivement corrélé avec la vitesse du vent pour toute la matinée. CP6 reflète l'influence de grandes mesures de la vitesse du vent durant la nuit. Le reste des composantes principales (CP7, CP8) sont positivement corrélées avec l'humidité.

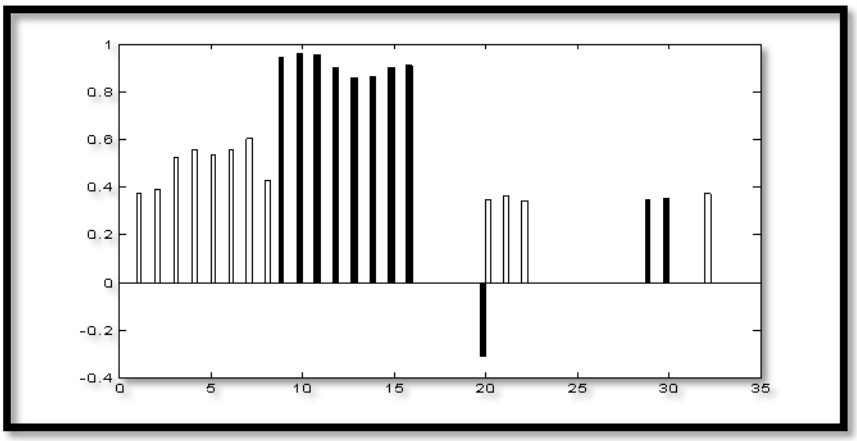


Figure 3.3-Paramètres météorologiques des deux premières composantes principales.

Les deux premières CPs expriment une variance cumulée uniquement de 55.061 % de l'ensemble des axes principaux, ce pourcentage faible (<80%) nous amène à conclure que l'ACP n'est pas bien appropriée pour expliquer la distribution des clusters météorologiques, ceci est principalement due au comportement non linéaire des données météorologiques.

Huit classes météorologiques ont été extraites par l'ACP, ces classes représentent les huit composantes les plus appropriées en terme de variance exprimée, cependant ces classes ne sont pas bien appropriées pour représenter les huit clusters météorologiques. Par le fait que l'ACP n'est pas une transformation discriminante, ce qui rend l'interprétation des clusters une tache difficile. La carte de kohonen peut être vue comme une extension non linéaire de l'ACP, cette méthode a bien prouvé son efficacité dans la classification et la visualisation des bases de données multidimensionnelles.

3.3. Identification des types de jours météorologiques par SOM

3.3.1. Prétraitement de donnés

La base de données collectée par la station météorologique de l'aéroport d'Annaba a été utilisé pour l'identification des types de jours météorologiques pour la région d'Annaba. Une phase de prétraitement des données est nécessaire pour préparer les données à la phase suivante en éliminant le bruit, corrigeant les erreurs et en normalisant les données. En effet les données manquantes et les données aberrantes constituent les principaux problèmes qu'on doit faire face. Les données aberrantes sont principalement dues au fonctionnement incorrect des instruments de mesure ou de la méthodologie incorrecte pour la collection et l'analyse. Pour traiter le problème des données aberrantes, les valeurs maximum et minimum peuvent être considérées comme données aberrantes et elles ont été examinées soigneusement, parce qu'elles peuvent causer une déformation dans le modèle obtenu. La présence des données manquantes peut également infirmer l'analyse statistique, et présenter des composants systématiques des erreurs au sujet de l'évaluation des paramètres du modèle.

L'approche de classification à deux niveaux discutée précédemment permet de traiter les données manquantes et les données aberrantes de façon très efficace, en effet la carte de kohonen permet de générer des prototypes (vecteurs référents) qui sont des moyennes locale de données et donc moins sensibles aux variations aléatoires que les données originales. De plus l'algorithme d'apprentissage de la SOM traite le problème des données manquantes avec élégance l'ors 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 présentent dans les vecteurs prototypes.

3.3.2. Topologie de la carte de kohonen

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 initiales de données. Le travail réalisé est basé sur l'utilisation de la librairie somtoolbox, disponible gratuitement dans : http://www.cis.hut.fi/projects/somtoolbox.

La topologie de la carte de kohonen utilisée pour regrouper les données météorologiques est représentée par la figure 3.4 ou les cartes sont reliées aux nœuds hexagonaux adjacents de taille 18×10 par 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 et unique 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-moyennes, 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. Il n'y a aucune règle explicite permettant de choisir le nombre de nœuds d'un réseau de Kohonen, mais le principe est que la taille devrait permettre la détection facile de la structure d'une carte de kohonen [78]. Le nombre total des unités de la carte de kohonen est estimé on utilisant la formule heuristique $M=5\sqrt{N}$. Après que le nombre d'unités de la carte ait été déterminé, la taille de la carte est déterminée. Fondamentalement, les deux plus grandes valeurs propres des données de l'apprentissage sont calculées et le rapport largeur-longueur de la grille de la carte est mis à ce rapport. 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 vérifiant les performances de chaque solution.

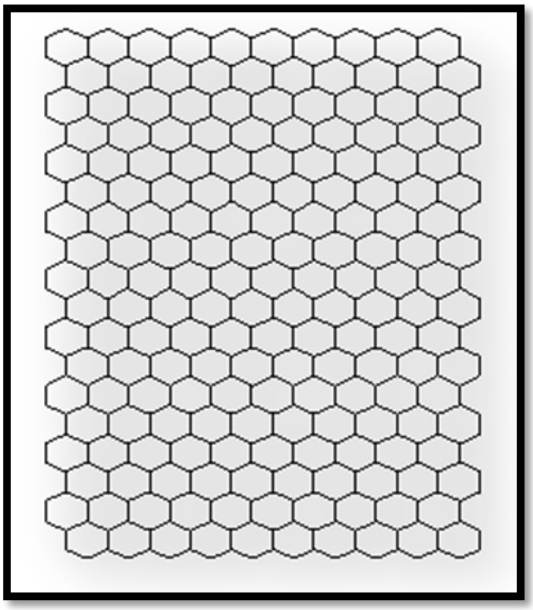
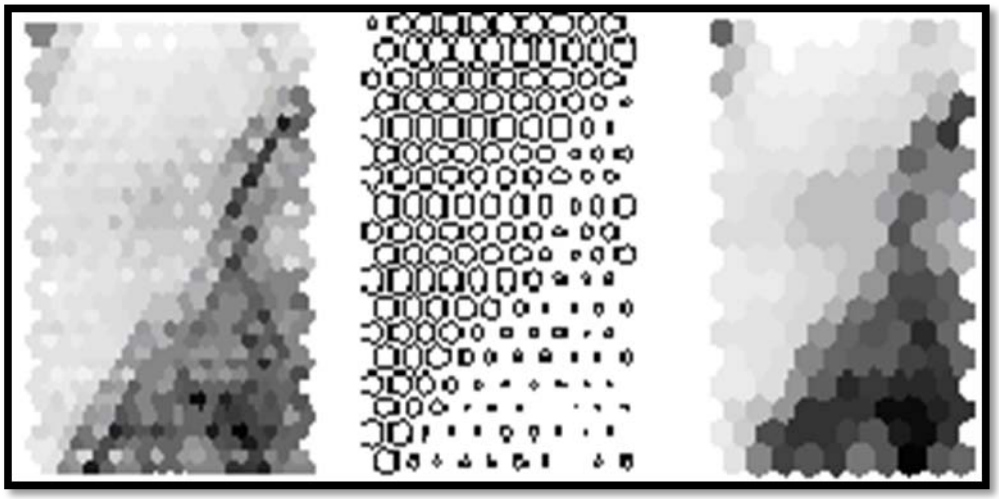


Figure 3.4--La topologie de la carte de kohonen utilisée pour regrouper les données météorologiques.

3.3.3. Apprentissage et résultats

La base de données météorologique récoltée pendant 60 mois a été utilisé pour l'apprentissage de la carte de kohonen dont la topologie est déterminée dans la section précédente. Cette carte a été linéairement initialisée en premier temps, par la suite un apprentissage de type séquentiel a été effectué. Différents nombres d'époques ont été considérés pour l'apprentissage du réseau compétitive (100, 500, 750, 1000, 5000).Cependant, un apprentissage de plus de 500 époques a mené à des changements très légers dans les vecteurs poids du réseau. Les résultats d'application de la méthode SOM sur les données météorologiques sont montrés par la figure 3.3. La figure 3.3(a) fournit une visualisation de l'U-matrice qui représente une mesure relative de distance entre les nœuds colorés du réseau. Les points blancs indiquent les emplacements des unités de la carte et les hexagones entre eux indiquent les valeurs réelles de l'U-matrice. La couleur grise (l'ombre) de l'hexagone dénote la distance à l'unité voisine de la carte. Plus l'ombre est foncé plus la distance est grande, un cluster (classe) qui représente des vecteurs de données similaire peut être vu en tant que zone claire avec des frontières foncées.

La figure 3.3(b) et la figure 3.3(c) représentent la distance moyenne de chaque unité de la carte à ses voisins et peuvent être vu comme une version moyenne de l'U-matrice, pour la figure 3.3(c) la taille de chaque unité de la carte est proportionnelle avec la distance moyenne de ses voisins.



(a) (b) (c)

Figure 3.5--La figure (a) représente l'U-matrice, (b) carte de des distances moyennes, (c) la carte des distances moyennes et représentation de taille des neurones.

La carte U-matrice permet de représenter la distribution des données de façon très claire. Cependant, cette carte indique la situation où l'utilisation uniquement de la mesure de distance n'est pas fiable pour déterminer les groupes représentatifs. Les clusters obtenus de la carte U-matrice sont représentés par des cercles.

Comme il a été montré par [79], il est difficile de regrouper visuellement les neurones de sorties d'une SOM lorsque le réseau de kohonen est fortement peuplé. Dans ce cas, la décision s'avère difficile et l'utilisation uniquement de la distance euclidienne pour identifier les classes météorologiques n'est pas fiable. Face à ce problème, une combinaison de la mesure de distance et de la carte code-couleur de la méthode SOM est encore utilisée pour visualiser les situations météorologiques de la région d'Annaba. Comme montré par la figure 3.6, les neurones dont les valeurs des paramètres météorologiques sont similaires évaluent automatiquement des couleurs similaires (en termes de distance) sur les nœuds de la grille. Les grandes mesures de distance des nœuds du réseau de kohonen sont automatiquement assignées aux différentes couleurs et clusters. Pour sélectionner une classe météorologique, nous identifions d'abord les régions des clusters basés sur la décoloration des nœuds. Dans les situations ou les couleurs des nœuds sont peu claires pour indiquer les différences des clusters, les mesures de distance sont alors utilisées comme moyen pour vérifier les groupes sélectionnés. Bien que ces deux méthodes aient été utilisées pour identifier les clusters, il a été très difficile d'attribuer quelques régions de nœuds à un groupe donné. Le problème que nous avons rencontré est la sélection des frontières des groupes pour cela un deuxième niveau de classification automatique par k-moyennes est très utile pour enlever l'ambiguïté et identifier les clusters météorologiques.

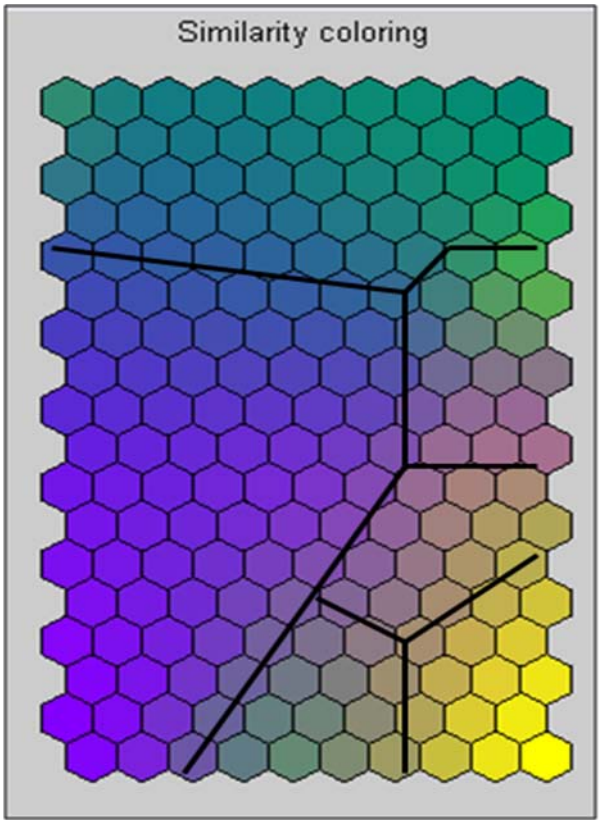


Figure 3. 6--La carte code-couleur.

3.3.4. Affinage des types de jour par la méthode k-moyennes

Dans la deuxième étape du processus de classification à deux niveaux, plusieurs algorithmes de classification automatique ont été utilisés. Dans le but de découvrir la méthode la plus appropriée à utiliser pour regrouper les unités de la SOM, nous avons procédé à une comparaison basée sur les critères de performances des algorithmes de classification les plus utilisés tel que : l'algorithme PAM (Partitioning Around Medoids), K-moyennes, et la classification hiérarchique (méthode de Ward). Les indices de performance : Davies-Bouldin et silhouette ont été utilisés pour l'évaluation de la qualité de la classification et la comparaison des performances des méthodes de regroupement utilisées. Ainsi, le meilleur algorithme est alors utilisé pour regrouper les unités de la SOM. Selon la figure 3.7, nous pouvons constater que k-moyennes génère les plus faibles valeurs pour l'indice Davies-Bouldin dans k= {4, 5, 6, 7} parmi les méthodes de classification considérées, et sa valeur pour k=6 est la plus faible, tel que *k* est le nombre de groupes (regroupement à k clusters). Cet algorithme génère aussi les plus grandes valeurs de silhouette et sa valeur pour k=6 est la plus grande parmi toutes les valeurs, on peut donc conclure que k-moyennes est l'algorithme le plus approprié pour regrouper les nœuds de la carte de kohonen.

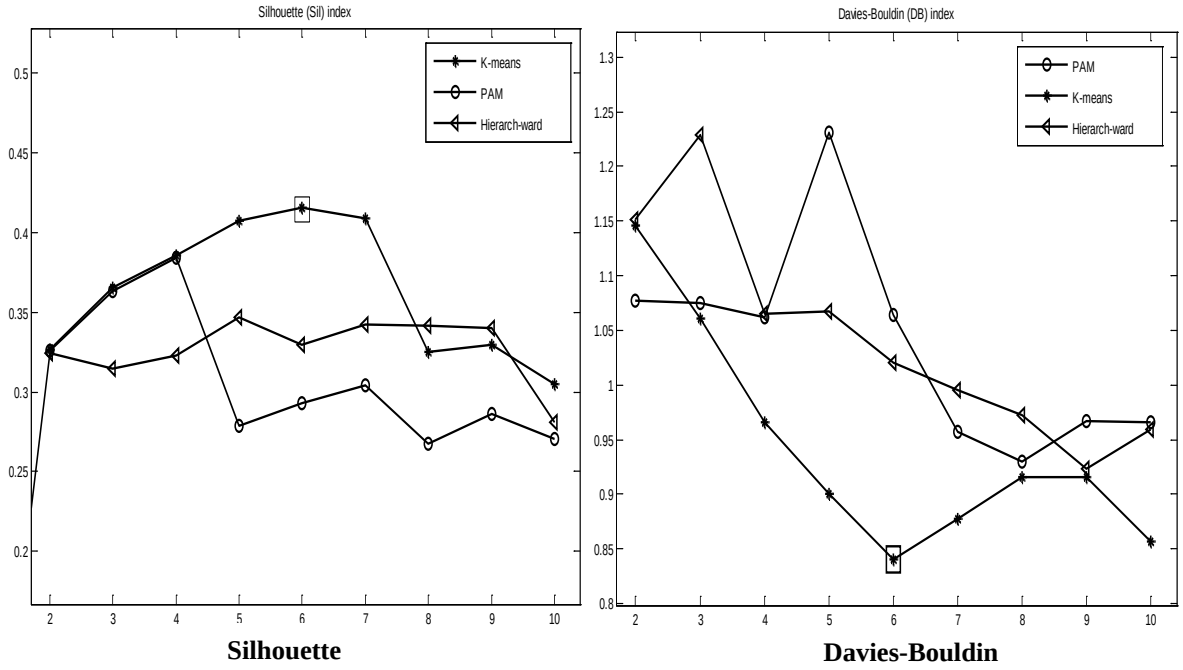


Figure 3.7-Validation quantitative et comparaison des résultats.

L'algorithme des k-moyenne est sensible aux paramètres d'initialisations, pour cette raison il a été exécuté 100 fois pour chaque *k*. Le meilleur regroupement obtenu parmi différentes valeurs de *k* est choisi selon les indices de validité. Selon les deux indices de Davies-Bouldin et silhouette discutés précédemment, le nombre optimal de clusters est égale à six. La méthode évolution de système dont les valeurs d'application sont représentées par la figure 3.8, indique aussi qu'une partition dont le nombre de clusters est égale à six est optimale. Et c'est le même résultat obtenu à partir de l'indice inter-intra poids présenté par la figure 3.9, ainsi d'après les indices de performance représentés par la figure 3.9, les clusters météorologiques obtenus sont bien séparés et homogènes.

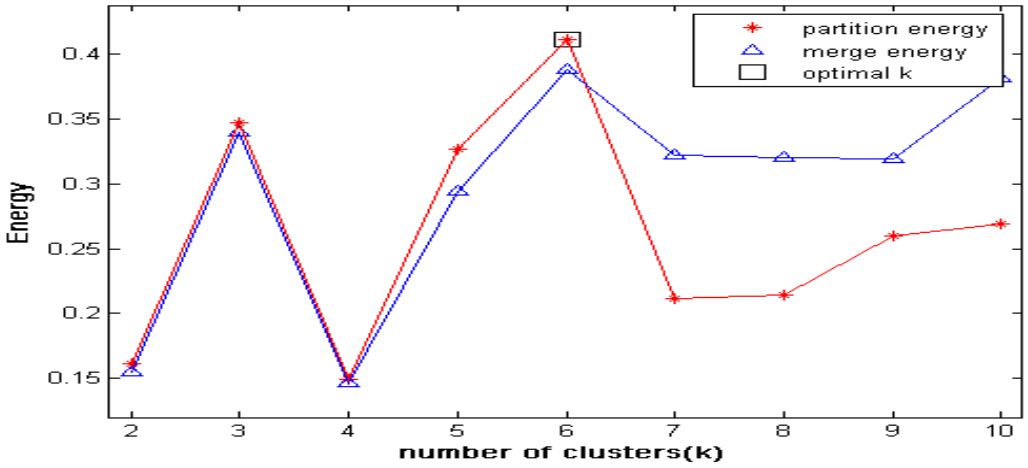


Figure 3. 8--La méthode évolution du système.

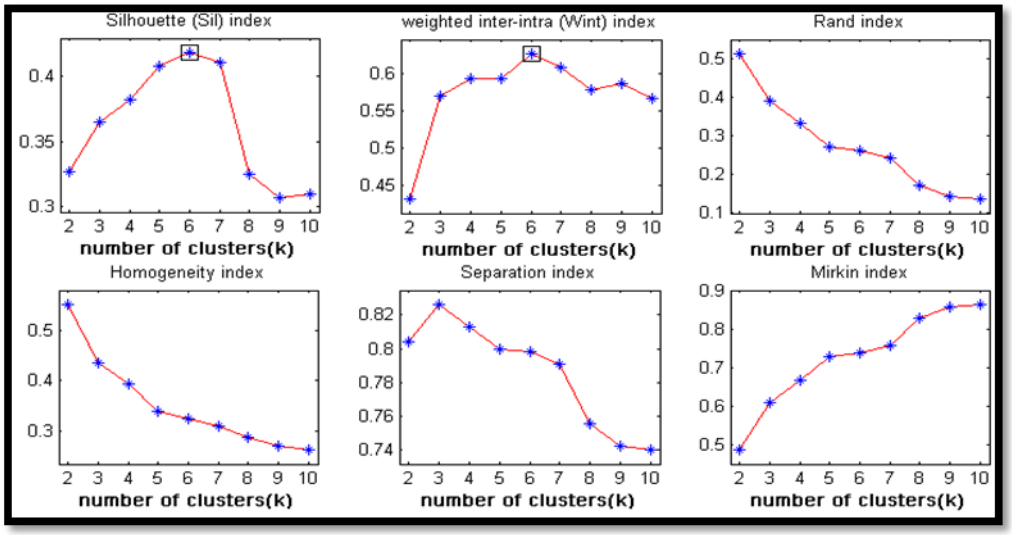


Figure 3.9--Les indices de validité obtenus pour chaque k clusters.

Les résultats obtenus on appliquant l'approche de classification non supervisée à deux niveaux sont représentés par la figure 3.10, et les paramètres météorologiques moyens de chaque cluster sont montrés par la figure 3.11. Le cluster météorologique C3 est caractérisé par une pression régulière pendant toutes les 24h, et d'une température élevée qui dépasse 25 °C durant la journée et un peu plus bas dans la nuit ce cluster est caractérisé aussi par une pression élevée pendant la nuit et basse durant la journée, la vitesse du vent est très faible pendant la nuit et commence à s'augmenter durant la journée, d'après la répartition mensuelle des clusters donnée par la figure 3.12, ce cluster représente les mois chauds de la région d'Annaba. Le sixième cluster est particulièrement concentré dans les mois d'hiver et d'automne, ce cluster est caractérisé principalement par une pression régulière et d'une vitesse du vent élevée pendant la journée, le quatrième cluster est similaire au sixième avec une pression élevée dans la nuit et une vitesse du vent un peu plus forte. Le cinquième cluster météorologique est caractérisé par une pression et un taux d'humidité élevé par rapport aux autres clusters et d'une température et vitesse du vent faibles et presque stables durant tout le jour, le premier cluster est presque similaire au quatrième cluster météorologique avec une pression faible au début de la journée et qui commence à s'accroître par la suite avec une vitesse du vent un peu plus faible, le deuxième cluster semble être un sous cluster de C4 avec une pression un peu plus élevée.

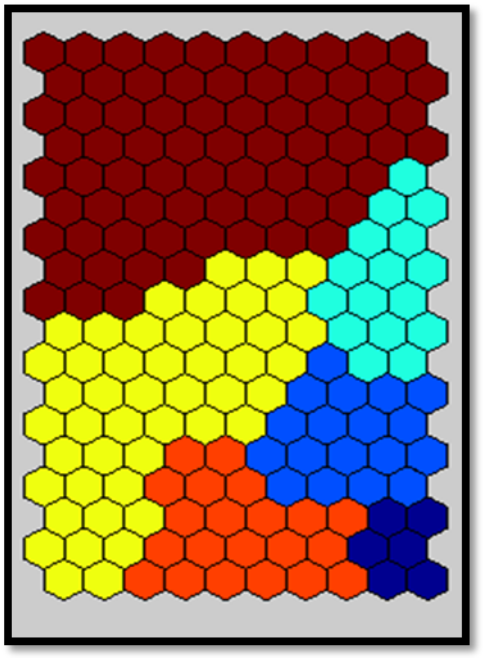


Figure 3.10-Distribution des clusters dans les cartes de kohonen après le regroupement des unités par k-moyennes.

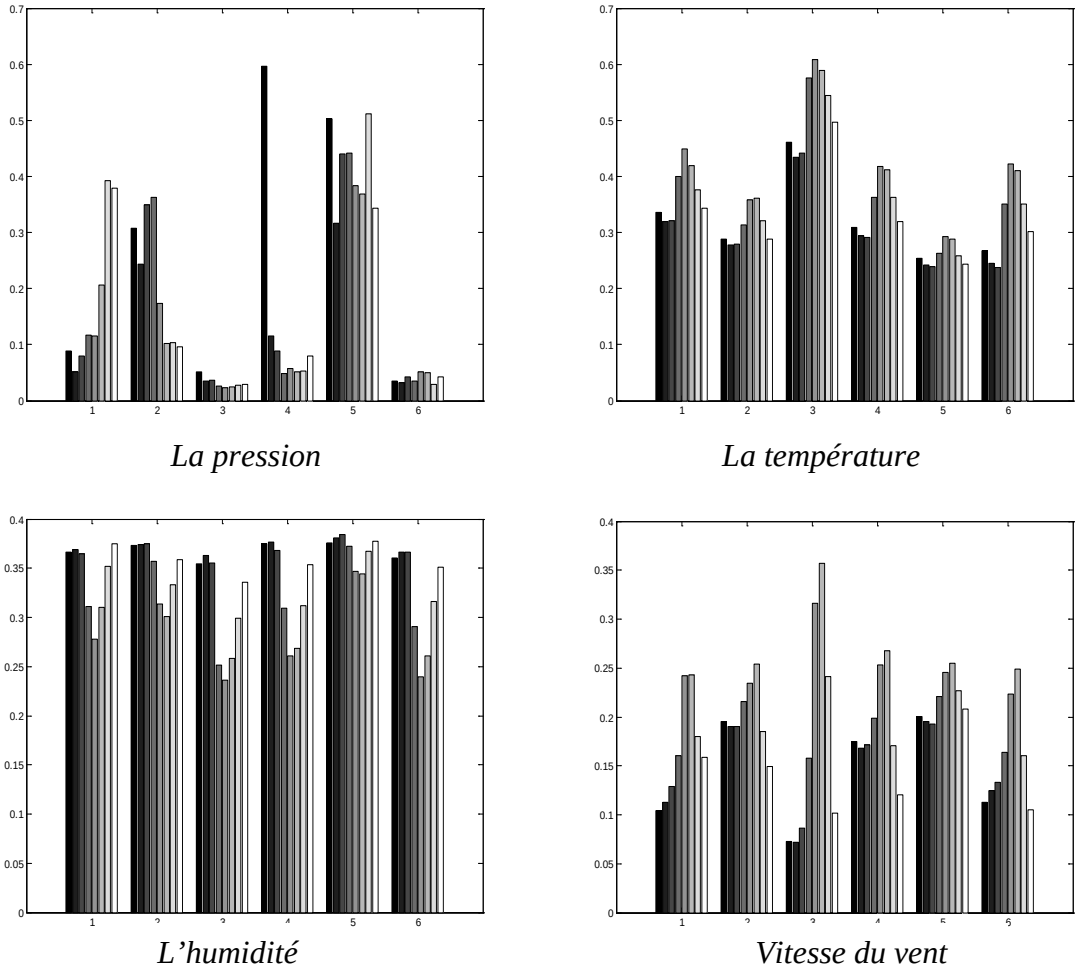


Figure 3.11-Les paramètres météorologiques moyens pour chaque cluster.

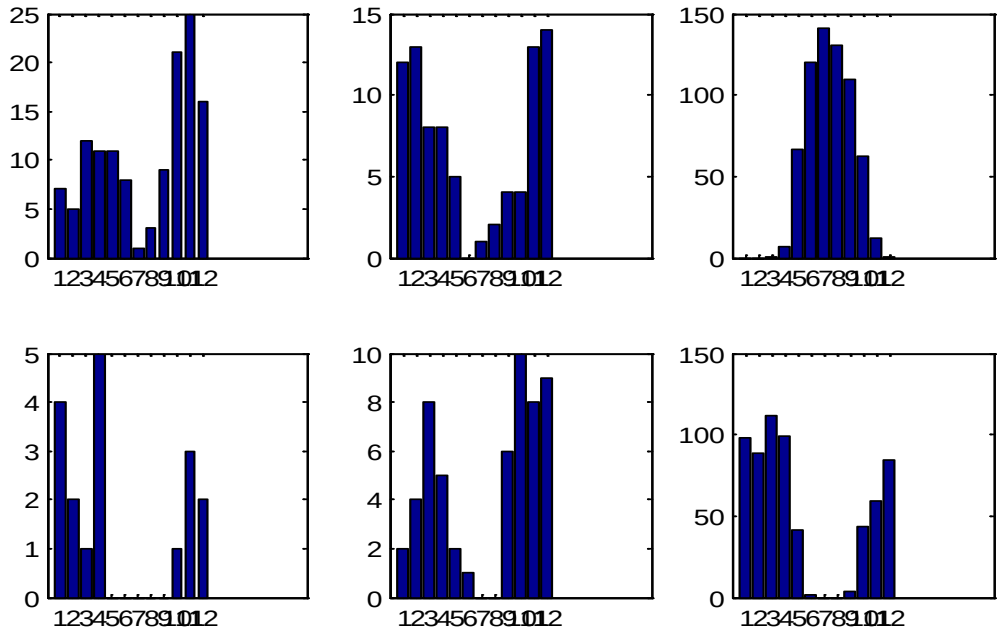


Figure 3.12--Répartition mensuelle des clusters.

3.3.5. Evaluation et test de l'approche proposée

La base de données collectée par la station SAMASAFIA à été utilisée pour tester l'approche de classification non supervisée à deux niveaux décrite précédemment et identifier les types de jours météorologiques pour la région. Les clusters météorologiques obtenus vont servir par la suite à étudier l'influence des paramètres météorologiques (par cluster) sur la pollution atmosphérique dans cette région.

Une topologie d'une SOM définit par une carte bidimensionnelle de taille 12×9, avec un treillis hexagonal et une fonction de voisinage gaussienne a été utilisée pour l'identification des clusters météorologiques. Cette carte a été linéairement initialisée avant de procéder à un apprentissage de type séquentiel. Différents nombres d'époques ont été considérés pour l'apprentissage de la carte de kohonen (100, 300, 350, 500, 1000).Cependant, plus de 300 époques d'apprentissage a mené à des changements très légers dans les vecteurs poids du réseau. L'U-matrice et la carte code-couleur sont montrés par la figure 3.13.

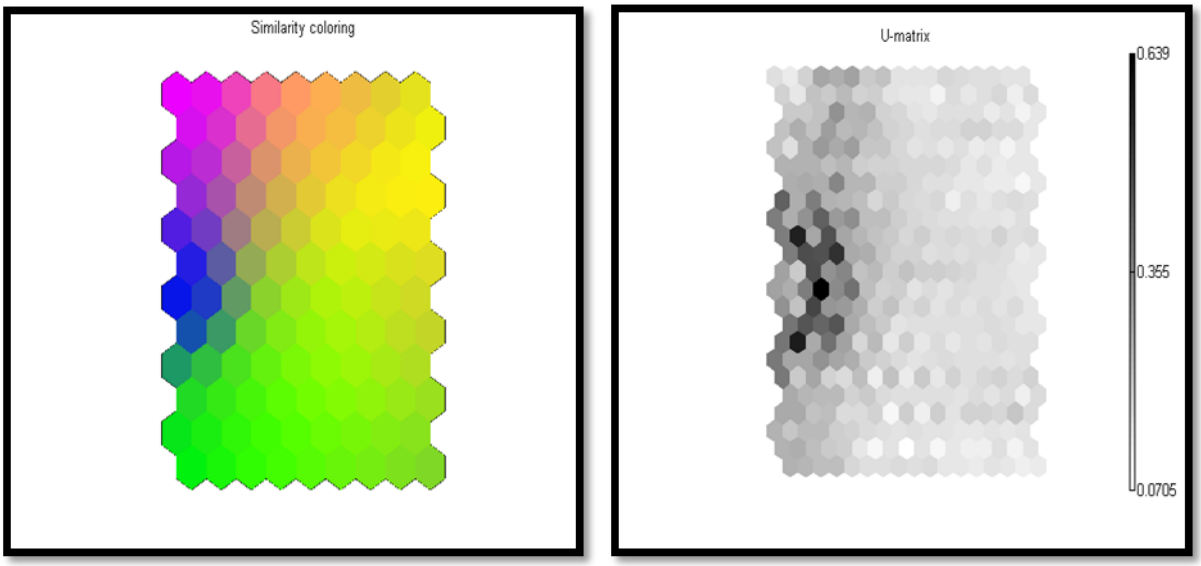


Figure 3.13-(a) La carte code-couleur (simiarity coloring), (b) la carte U-matrice.

Visuellement, et selon les deux cartes représentées par la figure 3.13, cinq cluster peuvent être extraite. Pour bien identifier les frontières de chaque cluster, K-moyennes a été utilisé pour regrouper les unités de la carte de kohonen. Par ce qu'il est très bien adapté pour une utilisation avec la méthode des k-moyennes [30], l'indice de Davies-Bouldin a été utilisé pour déterminer le meilleur regroupement obtenu parmi différentes valeurs de k. L'indice de Davies-Bouldin ainsi que la somme de l'erreur au carrée sont montrés par la figure 3.14 (a). Selon l'indice de Davies-Bouldin, un pic négative est remarqué dans la partition à cinq groupes ce qui indique que le nombre optimale de clusters météorologiques est égale à cinq. La distribution des clusters météorologiques sur la carte de kohonen après le regroupement des unités est représentée par la figure 3.14 (b).

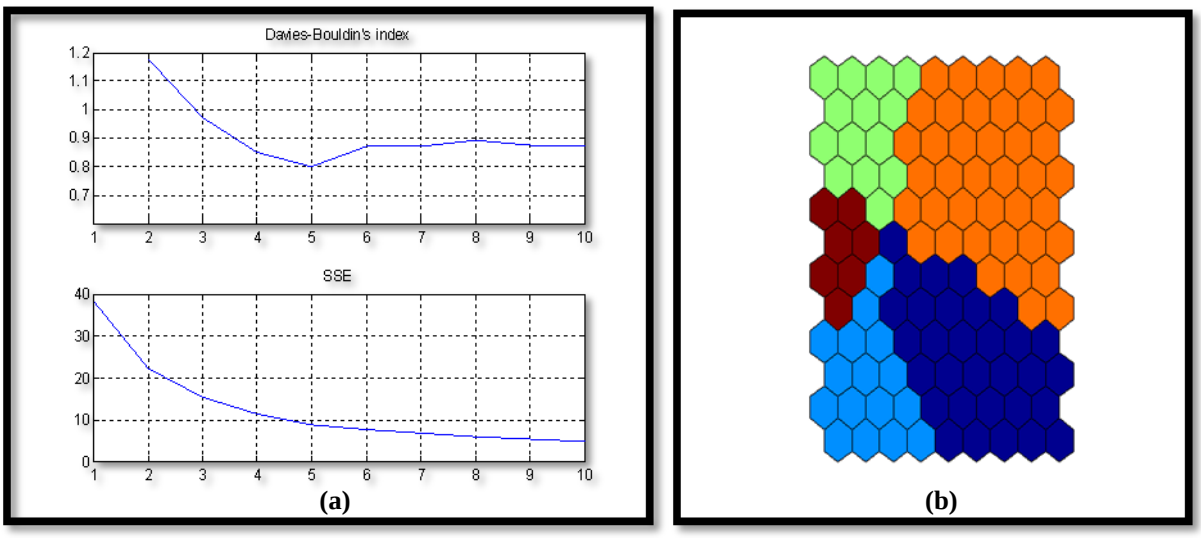


Figure 3. 14- (a) L'indice Davies-Bouldin et la somme de l'erreur au carrée pour k= [1 10]. (b) Distribution des clusters dans la carte.

Les paramètres météorologiques moyens de chaque cluster sont montrés par la figure 3.15, ainsi la distribution mensuelle des clusters par la figure 3.16.

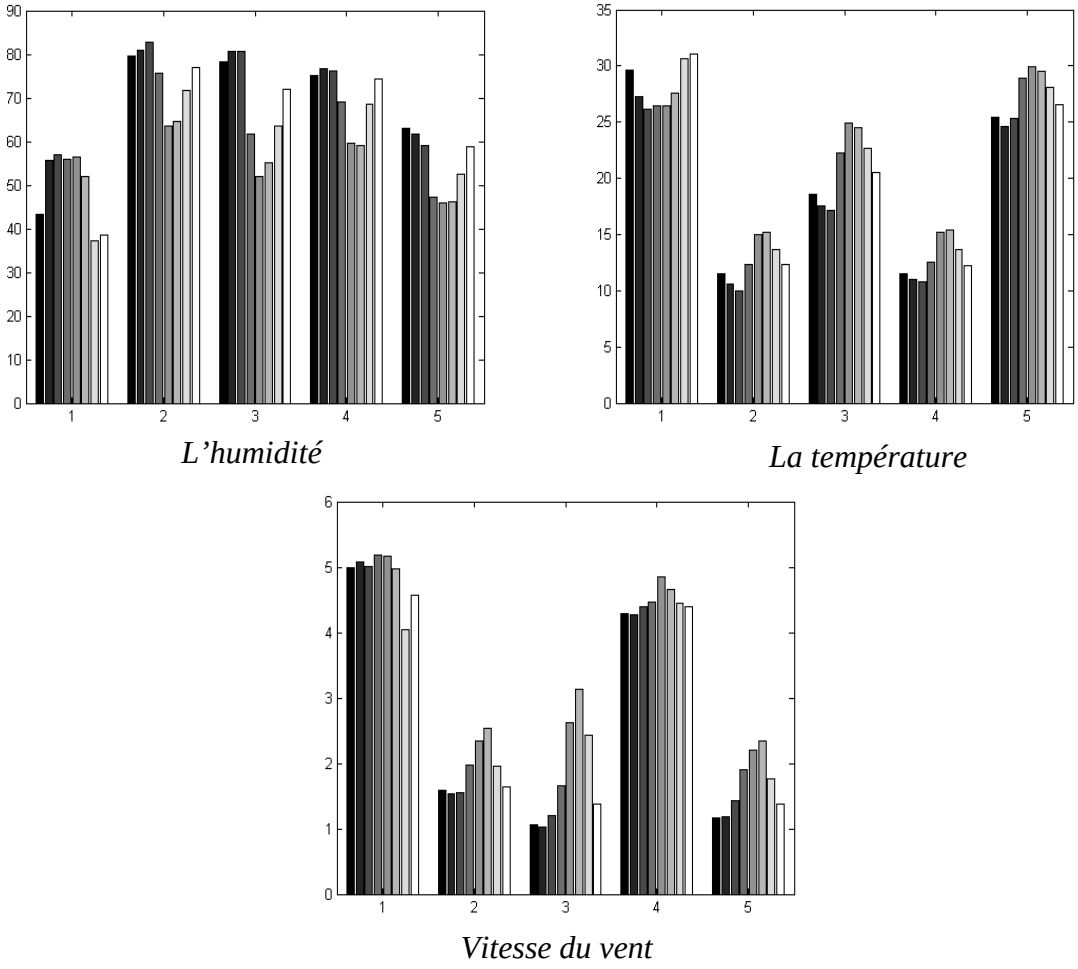


Figure 3. 15-Les paramètres météorologiques moyens pour chaque cluster.

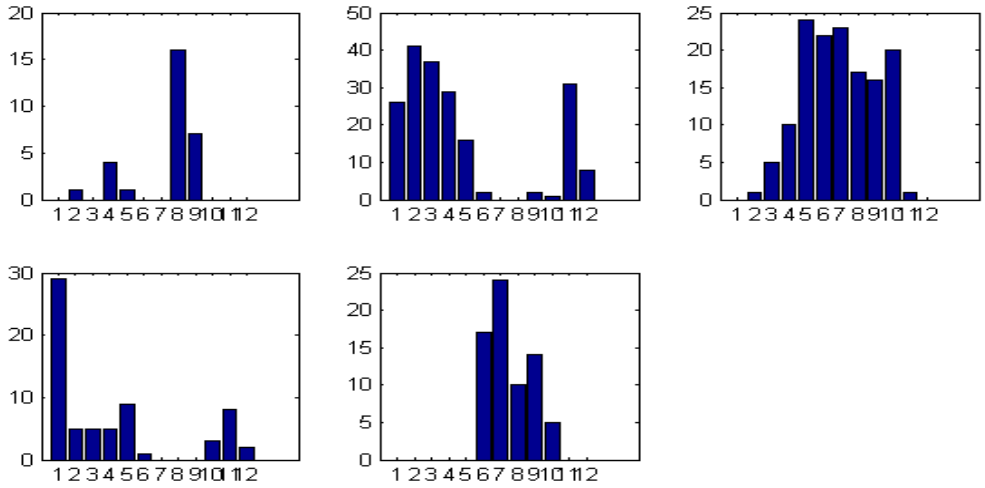


Figure 3.16-Répartition mensuelle des clusters.

Les vecteurs météorologiques du cluster C3 sont caractérisés par une température moyenne entre 20 et 25^0 C et d'une humidité élevée pendant la nuit et moyenne dans la journée, la vitesse du vent est très faible pendant la nuit et commence à s'augmenter durant la journée, d'après la répartition mensuelle des clusters donnée par la figure 3.15, ce cluster représente les mois chauds. Le deuxième cluster est particulièrement distribué dans les mois d'hiver, ce cluster est caractérisé principalement par une basse température et d'une faible vitesse du vent, ce cluster est caractérisé aussi par un taux d'humidité élevé qui dépasse en moyenne 70%. Le premier cluster regroupe les vecteurs météorologiques dont la température est très élevée et qui dépasse en moyenne 25^0c, avec aussi une vitesse du vent élevée par rapport aux autres clusters. Le quatrième cluster est presque similaire au deuxième, avec une vitesse de vent élevée qui dépasse en moyenne 4m/s. Les vecteurs météorologiques du cinquième cluster sont presque similaires à ceux du premier cluster avec une faible vitesse de vent durant les 24 heures. Il est à noter que les clusters météorologiques obtenus sont similaires à ceux obtenus précédemment pour l'ensemble de données capté pendant la période 1995-1999, sauf qu'on n'a pas pris en considération la pression atmosphérique. Ces résultats montrent également l'utilité de l'approche proposée comme un moyen très efficace pour la classification des bases de données volumineuses avec des temps de calculs très réduits par rapport aux autres méthodes de classification.

Chapitre 4

Influence des paramètres météorologiques sur la pollution atmosphérique pour la région d'Annaba

4.1. Introduction

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 [80]. 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. Le niveau de concentration de la pollution atmosphérique est une combinaison des émissions des polluants et des processus chimiques et physiques qui se produisent dans l'atmosphère. Les réactions chimiques des polluants dépendent des conditions atmosphériques ambiantes et sont généralement influencées par le rayonnement d'onde courte, la température de l'air, la vitesse du vent, la direction du vent et l’humidité relative [81]. 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 [1]. Il est extrêmement important de considérer l'effet des conditions météorologiques sur la pollution atmosphérique, puisqu'elles influencent directement les possibilités de dispersion de l'atmosphère. 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 pour disperser les polluants [2].

L'étude des rapports entre les variables dans les bases de données volumineuses telles que la pollution atmosphérique et les paramètres météorologiques peut fournir des informations importantes concernant la nature des dépendances dans les données. Une modélisation (ou une simulation) de l'atmosphère urbaine peut être considérée comme un système qui répond en produisant différents ensembles de sorties, c.-à-d. niveaux de concentration des polluants d'intérêt, sous l’introduction des conditions météorologiques (c.-à-d. les entrées) [82]. Cependant, un des facteurs de complicité dans le sujet de la simulation de la qualité de l’air est lié au comportement non-linéaire des polluants secondaires, c.-à-d. les polluants qui ne sont pas directement émis mais qui sont plutôt formés dans l'atmosphère en raison des réactions chimiques telles que l'ozone (O3) [83]. Les modèles statistiques espaces-temps sont très utiles pour :

- l'éclaircissement des rapports entre les différents polluants atmosphériques,
- mesurer les liens entre les polluants atmosphériques et les paramètres météorologiques,
- valider les modèles de dispersion de pollution atmosphérique,
- définir les rapports espace-temps des épisodes dangereux de la qualité d'air,
- prédire les niveaux de concentration de la pollution atmosphérique urbaine, et
- évaluer l'efficacité de surveillance des réseaux de contrôle.

Pendant les dernières décennies beaucoup d'efforts ont été consacrés à étudier les relations entre la pollution atmosphérique et les paramètres météorologiques [83], [84]. 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 dont la concentration de l'air pollué à un emplacement spécifique est liée au volume de trafic et les variables météorologiques [85]. Les auteurs de [86] ont étudié les relations entre les concentrations des particules ultra-fines ($PM_{2.5}$) 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, en utilisant les modèles de la régression linéaire à effets mixtes. La régression multiple a été utilisée par [87] pour étudier l'effet des paramètres météorologiques captés pendant l'hiver et les périodes d'été sur les concentrations des particules ultra-fines $PM_{2.5}$ et $PM_{2.5-10}$. Ainsi, les relations entre les particules en suspension et les concentrations des gazes sulfuriques avec les facteurs météorologiques, tels que la vitesse du vent, la température, l'humidité relative, la pression et la précipitation, dans les saisons d'hiver ont été statistiquement analysées en utilisant l'analyse multiple par étapes de la régression linéaire. Selon les résultats obtenus de ces études il y a des niveaux de relation modérée et faible pendant quelques mois entre les niveaux de concentrations des polluants et les facteurs météorologiques. La régression linéaire à été également utilisée par [88], pour étudier l'influence des paramètres météorologiques tels que la température, la vitesse et la direction du vent sur les niveaux de concentration de $PM_{2.5}$ et PM_{10}. Les auteurs de ce papier ont découvert que ces paramètres météorologiques sont non linéairement liés aux concentrations des $PM_{2.5}$ et PM_{10}. Pour manipuler ces paramètres, ils ont convertis les variables météorologiques en variables binaires qui sont utilisés par la suite pour construire un modèle linéaire modifié.

L'analyse en composantes principales (ACP) a été utilisée dans de nombreuses études telles que [83], [89] pour identifier les relations cachées entre le polluant examiné et les facteurs qui

favorisent sa formation. Les résultats de l'ACP sont également utilisés en tant que paramètres de guide pour la modélisation des polluants dans le but de simuler leurs comportements et produire ainsi les prévisions fiables qui peuvent être utilisées pour la gestion opérationnelle de la qualité de l'air. D'autre part, les méthodes non linéaires ont été aussi largement utilisées pour la modélisation de la qualité de l'air [83], [5], [6], ou il a été déduit que les RNAs sont plus efficaces que les méthodes linéaires en estimant l'influence des paramètres métrologiques sur les concentrations des polluants tels que l'ozone. Ainsi dans [83] les réseaux de neurones artificiels ont été supportés par une utilisation en parallèle avec l'ACP.

Les modèles des réseaux de neurones artificiels ont été largement utilisés pour modéliser les concentrations de l'air pollué dans le but de les prédire. Étant donné que les RNAs sont capable 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 comparées aux méthodes statistiques telles que la régression linéaire multiple [5], [6] et c'est pour cette raison qu'ils sont considérés dans cette étude. L'objectif de ce travail est d'étudier l'influence des paramètres météorologiques (la vitesse de vent, la température et l'humidité relative) sur la pollution atmosphérique (le monoxyde d'azote (NO), l'oxyde de carbone (CO), l'ozone (O_3), l'oxydes d'azote (NO_x), le dioxyde d'azote (NO_2), le dioxyde de soufre (SO_2) et les particules en suspension[2] (PM_{10})) dans la région d'Annaba pendant la période 2003-2004. Pour étudier efficacement les relations entre les paramètres météorologiques et la pollution atmosphérique nous avons proposé d'utiliser les clusters météorologiques obtenus précédemment. Donc l'impact de la météorologie sur la pollution atmosphérique est étudié au niveau des clusters météorologiques.

4.2. Pollution atmosphérique dans la région d'Annaba

La dégradation de la qualité de l'air a été longtemps perçue en termes de nuisances locales dont les impacts sur la santé humaine sont considérés comme prépondérants. Cette perception se justifie par le fait que les polluants atmosphériques exercent des effets directs et souvent tangibles au niveau local. En Algérie, les infections respiratoires demeurent la première cause de mortalité infantile après la rougeole et la diarrhée [90]. La bronchite chronique, le cancer du poumon et l'asthme sont, entre autres, les maladies engendrées par la pollution. L'année dernière par exemple, le nombre des asthmatiques était de 700 000 et il va en augmentant, car en 2010 il sera de 800 000. En somme, un tiers de la population algérienne a une morbidité

[2] Le terme PM10 vient de l'anglais "Particulate Matter" et signifie donc "matière particulaire". C'est un mélange complexe de substances minérales et organiques inférieurs à 10 microns.

respiratoire. Toutefois, l'Algérie ne dispose pas encore d'études épidémiologiques permettant de faire une corrélation entre les niveaux de pollution atteints et les maladies. Mais les concentrations élevées de certains polluants ont déjà atteint des niveaux dangereux, particulièrement pour des 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 a été citée comme exemple : 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 [90].

La ville d'Annaba est constituée d'une vaste plaine bordée au Sud et à l'Ouest, d'un massif montagneux au Nord, et par la mer à l'Est. Sa topographie en forme de cuvette favorise la stagnation de l'air et la formation d'inversions de températures. Ces situations permettent l'accumulation de polluants et l'élévation des taux de concentration qui en résulte. Les effets des brises de mer, terre, et pente concourent au transport des nuages de polluants. En effet, les nuages de polluants sont entraînés par la brise de terre la nuit vers la mer, et de jour. Ces nuages de polluants retournent sur la ville par effet de brise mer en longeant la montagne de SERAIDI. Les nuages tournent sur la ville sous une forme de cercle. Les polluants se déposent lentement par gravité et l'on assiste à une pollution affectant les trois récepteurs (mer, terre, air) [7].

Les émissions atmosphériques des principaux contaminants se répartissent de façon différente en fonction du secteur d'activité. L'industrie est l'élément moteur de croissance et de dégradation de l'environnement dans la ville de Annaba et sa région, ou la plupart des complexes industrielles sont implantés à proximité de la ville de Annaba tels que le complexe des engrais phosphaté et azotés « ASMIDAL », le complexe sidérurgique d'El-Hadjar « ex : SIDER » et le centrale électrique. 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. Dans la ville d'Annaba, les émissions dues au trafic routier représentent les principaux polluants de l'atmosphère. 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. Parallèlement, le mode de traitement des déchets (ménagers, industriels, hospitaliers, toxiques), qui consiste à mettre ces déchets en décharge dans des sites sauvages, est aussi une source de pollution atmosphérique dans la mesure où les déchets sont incinérés à l'air libre. La décomposition des déchets solides provoque le dégagement d'un gaz

comprenant 50 à 65 % de méthane (important gaz à effet de serre, comme il peut former un mélange explosif avec l'air). Selon les estimations, 7 % du méthane dégagé par les activités humaines proviennent des décharges. Les principaux polluants urbains dans la région d'Annaba sont présentés dans le tableau 4.1 [7].

Tableau 4.1--Les principaux polluants urbains à Annaba.

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 (Nox)	Trafic automobile	Formation d'ozone en basse Atmosphère (NOx + Vapeur) = Contribution aux pluies acides)	Altération des fonctions respiratoires	Le monoxyde émis à l'échappement s'oxyde et se transforme en dioxyde d'azote (NO2) plus toxique.
Dioxyde de souffre SO2	Combustion du Fioul	SO2+ 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 (O3)	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

4.3. Évaluation de Données

La base de données utilisée dans la présente étude couvre la période 2003-2004 a été fournie par le réseau SAMASAFIA (structure responsable de surveillance de la qualité de l'air en Algérie) d'Annaba sur une base de mesure continue de 24 heures. Les polluants atmosphériques surveillés en continu incluent les concentrations du : monoxyde d'azote (NO), monoxyde de carbone (CO), l'ozone (O_3), les particules en suspension (PM_{10}), l'oxydes d'azote (NO_x), le dioxyde d'azote (NO_2) et le dioxyde de soufre (SO_2). La base de données utilisée inclue également trois paramètres météorologiques : la vitesse de vent, la température et l'humidité relative. Ces paramètres météorologiques ont été utilisés pour identifier les

clusters météorologiques de la région d'Annaba on appliquant l'approche proposée dans [91], [92], [93], [94]. Une vue générale des polluants étudiés est montrée par le tableau 4.2.

Tableau 4.2-- Sommaire des données de pollution.

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

Une phase de prétraitement des données brutes est nécessaire pour faire face à deux principaux problèmes : les données manquantes et les données aberrantes. Les données aberrantes sont principalement dues au fonctionnement incorrect des instruments ou de la méthodologie incorrecte pour la collection et l'analyse. Généralement, les valeurs maximum et minimum peuvent être considérées comme données aberrantes et ils doivent être examinés soigneusement, parce qu'elles peuvent causer une déformation dans le modèle obtenu. Les données manquantes sont principalement dues aux échecs des instruments de mesure. Notamment en raison des coupures électriques (problèmes de climatisation) ainsi qu'en raison

des différentes pannes survenues aux différents analyseurs et de quelques blocages d'acquisition [10].

La présence des données manquantes peut infirmer l'analyse statistique, et présenter des composants systématiques des erreurs au sujet de l'évaluation des paramètres du modèle. En outre, si nous estimons les paramètres du modèle en exploitant les données observées sans tenir compte de la présence des données manquantes, les évaluations obtenues pourraient être incertaines parce que beaucoup d'informations au sujet des données manquantes disparaîtraient. Pour cette raison nous avons choisit d'utiliser un modèle neuronal pour chaque polluant étudié, pour minimiser la quantité des données manquantes pris en compte. De plus certaines mesures manquantes ont été remplacées avec des valeurs prévues en se basant sur des valeurs valides avant ou après la mesure manquante. Les niveaux de concentration moyenne pour les polluants considérés pour cette étude sont montrés par la figure 4.1 et la figure 4.2. Ainsi les niveaux de concentration de ces polluants pendant la période 2003-2004 sont représentés par la figure 4.3.

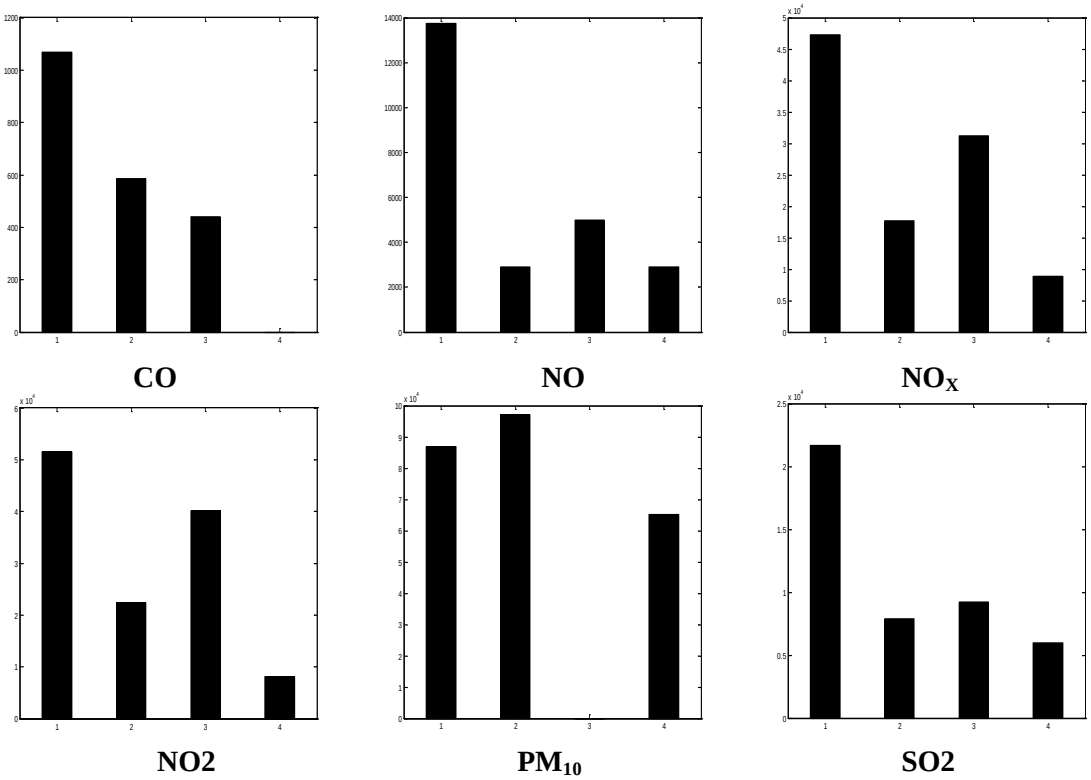


Figure 4. 1--Niveaux de concentration moyenne des polluants pendant 2003.

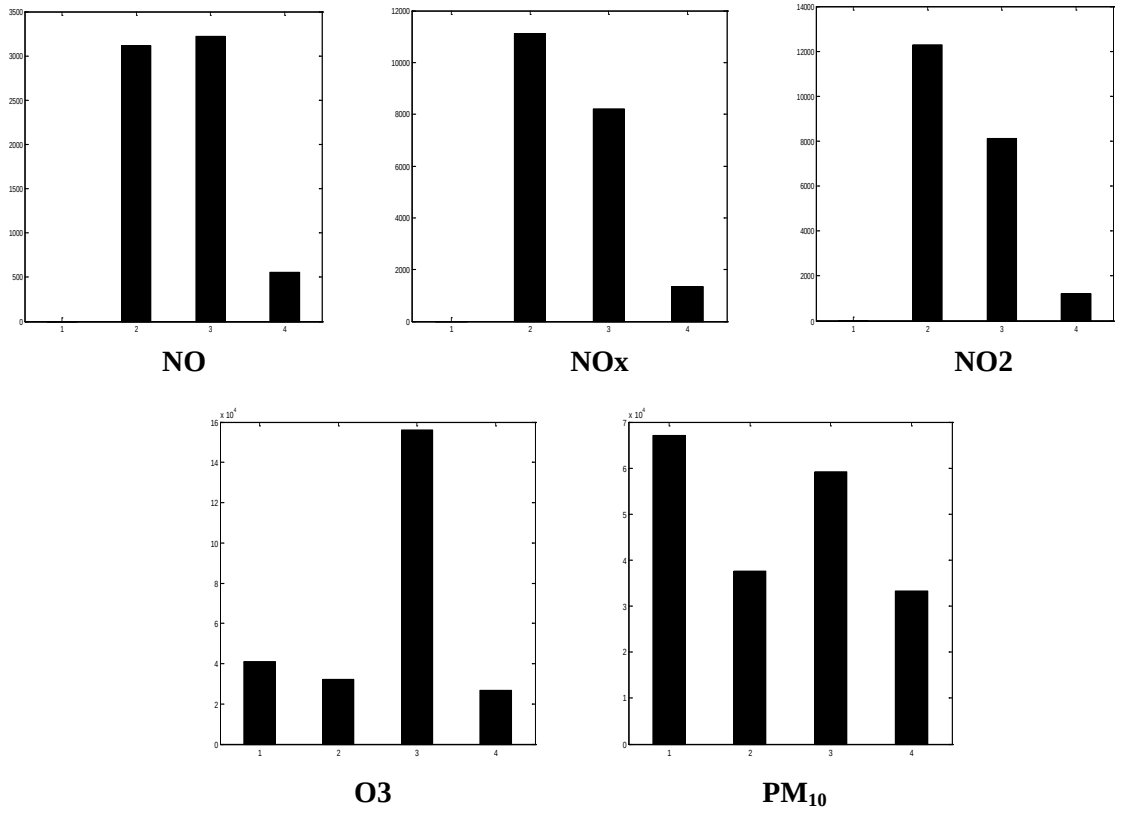


Figure 4.2-- Niveaux de concentration moyenne des polluants pendant 2004.

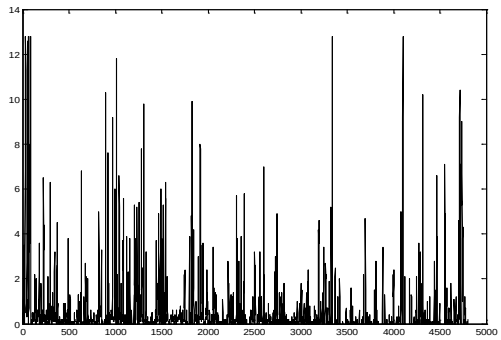
La figure 4.1, et la figure 4.2 montrent également des différences principales dans les niveaux de pollution entre les quatre saisons de chaque année. Les concentrations dans la saison du printemps tendent à être occasionnelles, des épisodes de pollution étaient très élevés dans la saison d'hiver pour la plupart des polluants. Selon [95], 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 stagnants. Ceci est clairement remarquable dans la figure 4.1. Où les niveaux moyens saisonniers du CO sont très élevés en hiver par rapport aux autres saisons. Ainsi l'évolution journalière des concentrations de monoxyde de carbone est généralement caractérisée par deux pointes de concentration observées aux heures d'intensification du trafic du matin et du soir [80]. Les niveaux de concentration moyenne de l'ozone atteignent leurs maximum dans les mois chauds cela est due principalement au fait que l'ozone est un polluant photochimique typique crée par les réactions des hydrocarbures et des oxydes d'azote en présence de la lumière du soleil qui agit en tant que catalyseur [83]. Ces conditions sont très favorables pour la formation de l'ozone dans les mois d'été. Selon [80], 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 en cette période de la journée. 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 [89].

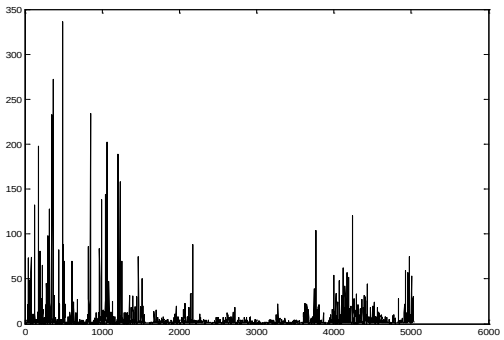
Nombreuses études dans la littérature ont essayé de déterminer statistiquement les effets des paramètres météorologiques sur les concentrations du dioxyde de soufre (SO_2) [96], [97], [98]. Ce polluant pourrait être attribué à la combustion des huiles, qui est en accord avec le caractère industriel de la région (Complexe Sidérurgique d'El-Hadjar) et les émissions domestiques (chauffage) [7]. Les niveaux de concentration du dioxyde de soufre les plus élevée sont observées en hiver. Selon [97] 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 dû à une basse température ayant pour résultat l'émission élevée de SO2 et également les facteurs météorologiques défavorables. 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.

Le monoxyde d'azote (NO) émis à l'échappement s'oxyde et se transforme en dioxyde d'azote (NO2) plus toxique. La figure 4.1 et la figure 4.2 montrent également une augmentation des teneurs en NO2 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. Nombreuses efforts ont été consacrés à l'étude du comportement de ce polluant tel que [99], [100].

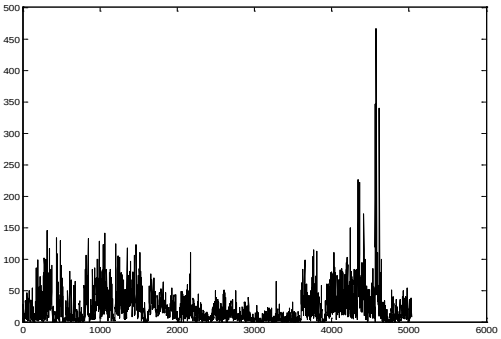
L'oxyde d'azote (NOx) inclut pour ça formation le NO, et le NO2, pour cette raison des épisodes de concentrations élevées ont été marqué en hiver. Concernant les particules en suspension, un cycle hebdomadaire de concentrations est manifesté dans la plupart des sites urbains ou les concentrations de 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 [11] et la figure 4.1, ainsi que la figure 4.2, les concentrations les plus élevées sont mesurées aussi bien en été qu'en hiver.



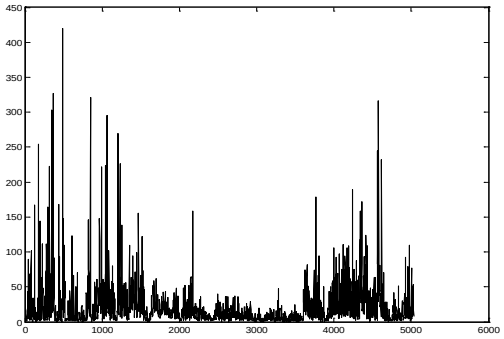
CO



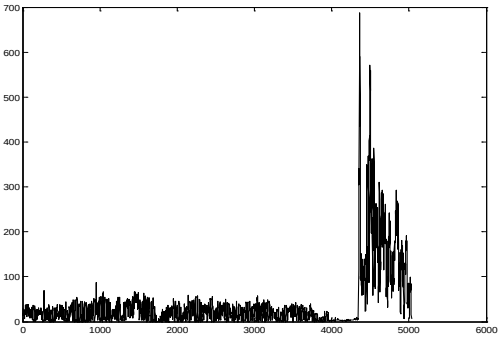
NO



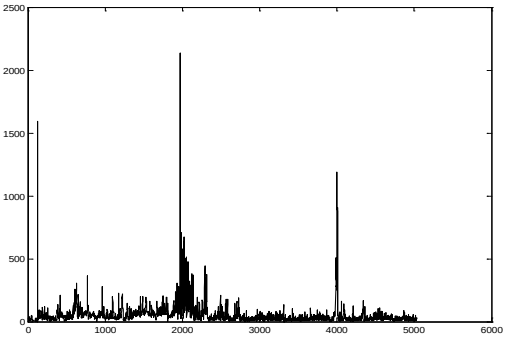
NOx



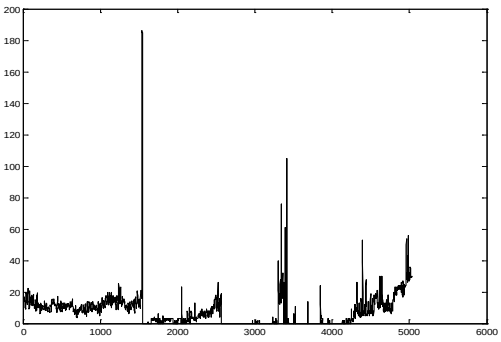
NO2



O3



PS



SO2

Figure 4.3-Niveaux de concentration des polluants à Annaba pendant la période 2003-2004.

4.4. Influence des paramètres météorologiques sur la pollution atmosphérique

Une approche pour prédire et modéliser la qualité atmosphérique de l'air est d'utiliser un modèle atmosphérique détaillé. Un tel modèle vise à résoudre les équations physiques et chimiques fondamentales contrôlant les concentrations des polluants et requiert donc des données détaillées des émissions de polluants et des paramètres météorologiques [101]. Une vue détaillée pour ce type particulier de modèle est fournie par [102]. La deuxième approche consiste à concevoir des modèles statistiques qui essayent de déterminer les rapports fondamentaux entre un ensemble de données d'entrée et les données cibles. Le modèle de régression est un exemple typique de l'approche statistique, ce dernier a été utilisé pour modéliser et prédire la qualité de l'air dans plusieurs études tel que [86]. Une des limitations imposées par les modèles de régression linéaire est qu'elles sont incapables de modéliser les systèmes non linéaires.

Au lieu d'utiliser les règles complexes et les routines mathématiques, les réseaux de neurones artificiels sont capables d'apprendre les principaux modèles d'information dans un domaine multidimensionnel. Les auteurs de [83], [5] ont conclu que les RNAs donnent généralement de meilleurs résultats une fois comparés aux méthodes statistiques linéaires, particulièrement si le problème analysé inclut un comportement non linéaire. Les RNAs peuvent également être utilisés en combinaison avec d'autres techniques traditionnelles telles que l'ACP [83]. L'inconvénient de l'approche neuronale est qu'aucune compréhension profonde des phénomènes physiques n'est gagnée en utilisant le réseau de neurones, puisqu'elle ressemble au comportement d'une boite noir. D'autre part, les problèmes de modélisation et de prédiction météorologique offrent un domaine très large pour examiner et développer les algorithmes non linéaires. Dans ce sens les RNAs ont été récemment établis comme un outil fiable pour l'analyse des séries temporelles et quelques résultats prometteurs ont été rapportés [103]. Le rapport entre les niveaux de concentration des polluants dans la région d'Annaba et les paramètres météorologiques est complexe et extrêmement non linéaires, par conséquent les réseaux neuronaux ont été utilisés pour modéliser l'influence des conditions météorologiques telles que la vitesse de vent, la température et l'humidité relative sur la qualité de l'air.

4.5. Application de l'ACP pour l'analyse environnementale

La modélisation de la qualité de l'air mis en œuvre généralement des bases de données multidimensionnelles, ou l'une des premières questions à répondre est liée à l'extraction des caractéristiques dans l'ensemble de données étudié. L'analyse en composante principale (ACP) est une méthode appropriée d'analyse de données multidimensionnelle qui permet d'identifier les relations cachées entre le polluant examiné et les facteurs qui favorisent sa formation. L'ACP a été largement utilisée pour identifier les sources possibles de la pollution ambiante. Parmi des techniques multivariables, l'ACP est souvent utilisée comme un outil exploratoire pour identifier les sources principales des émissions des polluants de l'air [1]. Les résultats de l'ACP peuvent être également utilisés en tant que paramètres de guide pour la modélisation des polluants dans le but de simuler leurs comportements et produire ainsi les prévisions fiables qui peuvent être utilisées pour la gestion opérationnelle de la qualité de l'air. En appliquant la méthode ACP sur l'ensemble de données, on obtiendra comme résultat un nouvel ensemble de variables non corrélées dites composantes principales (CPs). Les (CP_S), sont ordonnés selon leur pourcentage de variance original qu'ils expliquent. Chaque CP est perpendiculaire aux autres CPs, par conséquent il n'y a pas d'information redondante dans le nouvel ensemble de données obtenus. Par sélection des CPs les plus significatifs (sur la base d'un critère approprié), nous réussissons à réduire la dimension de l'ensemble de données original et également sélectionner la plupart des caractéristiques et paramètres d'analyse [83]. La méthode ACP a été utilisée pour projeter les données météorologiques et l'ensemble de polluants captés par la station de Sidi Amar (Annaba) pendant l'année 2004 uniquement. La matrice normalisée de covariance (c.-à-d. coefficient de corrélation) a été calculée et les résultats de l'ACP sont présentés respectivement dans les tableaux 4.3 et le tableau 4.4. Selon la matrice de corrélation montrée par le tableau 4.3, l'augmentation des valeurs de concentration des NO_2, NOx et NO est liée à la diminution des valeurs de température et de vitesse du vent, cependant l'humidité présente une faible corrélation avec ces trois polluants. L'oxyde d'azote (NOx) est le polluant le plus affecté par le trafic routier, avec un rapport approximativement linéaire [85]. Par conséquent la concentration de NOx est presque proportionnelle au nombre de véhicules. C'est raisonnable, puisque NOx inclut pour ça formation le NO, venant directement de l'échappement et le NO_2 qui est créé quand NO s'oxyde avec l'oxygène. Pour cette raison les niveaux de concentration des NOx sont fortement corrélés avec les concentrations de NO et NO_2.

Tableau 4.3-Coefficient de corrélation des données horaires des polluants à Annaba pendant l'année 2004.

	HU	***NO***	***NOx***	***NO_2***	***O_3***	***PM_{10}***	***TE***	***V-V***
HU	1	0,004	0,013	0,005	-0,070	-0,210	-0,610	-0,379
NO	0,004	1	0,615	0,306	-0,032	0,046	-0,049	-0,136
NOx	0,013	0,615	1	0,933	-0,113	0,057	-0,159	-0,143
NO_2	0,005	0,306	0,933	1	-0,115	0,048	-0,152	-0,107
O_3	-0,070	-0,032	-0,113	-0,115	1	0,196	0,408	0,039
PS	-0,210	0,046	0,057	0,048	0,196	1	0,383	-0,050
TE	-0,610	-0,049	-0,159	-0,152	0,408	0,383	1	0,147
V-V	-0,379	-0,136	-0,143	-0,107	0,039	-0,050	0,147	1

Les concentrations de NO_2 sont aussi liées positivement avec les niveaux de NO. L'augmentation des niveaux de concentration de l'ozone est liée à la diminution des valeurs de NOx et NO_2. D'autre part, l'ozone semble augmenter avec la température et diminuer avec l'humidité accrue. Comme la température peut être associée au rayonnement solaire, et l'humidité avec la présence du nuage, les deux résultats vérifient les mécanismes de base du système atmosphérique simulé : le rayonnement solaire amplifie la génération de l'ozone due aux réactions photochimiques. Les particules en suspension se corrèlent négativement avec l'humidité relative. Le coefficient de corrélation entre les PM_{10} et l'humidité relative est de -0,21. L'humidité et la pluie enlèvent les particules atmosphériques et diminuent la quantité de la poussière suspendue du sol en faisant le sol humide. L'humidité relative dans la région d'étude dépasse en moyenne 60%. Par conséquent, plus l'humidité relative est élevée, plus le taux d'arrangement de la poussière de l'atmosphère est élevé aussi. La température de l'air contribue également à expliquer les variations de concentrations des particules en suspension. Selon le tableau 4.2, les particules en suspension sont corrélées positivement avec la température. L'influence de la température de l'air sur les PM_{10} reflètent une tendance que les états atmosphériques les plus favorables de dispersion sont observées sous l'air chaud que l'air froid. La corrélation linéaire entre les particules en suspension et la vitesse de vent rapporte également une faible relation. La valeur de ce coefficient de corrélation est basse, et n'est pas donc statistiquement significative. Selon les résultats obtenus nous pouvons distinguer qu'il n y a pas de forte relation entre la vitesse du vent et la majorité des polluants dans la région

d'Annaba. La raison pour ce faible rapport est due principalement à la topographie de la région qui prend la forme d'une cuvette (un plat entouré d'un massif montagneux) qui favorise la stagnation par conséquent, le vent ne peut pas transporter la pollution loin de la région. D'après [104] Les effets de la vitesse du vent fonctionnent bien et plus efficacement sur les émissions de niveau élevé d'altitude que sur les émissions de niveau inférieur.

Les résultats d'application de l'ACP sur la base météorologique de la région d'Annaba sont représentés par le tableau 4.5, seulement les composantes principales associées à des valeurs propres supérieurs à l'unité (>1) ont été extraites. Quatre CPs dont les valeurs propres sont supérieures à 1 comptent pour 99% de la variation des données. La dimension mathématique des données peut donc être réduite de 24 à 4 en utilisant l'ACP.

Tableau 4. 4--Résultats de l'ACP pour les données horaires des polluants à Annaba pendant l'année 2004.

	C1	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***	***C8***
HU	-0,014	-0,430	-0,863	-0,13	-0,224	-0,006	-0,045	-0,001
NO	-0,001	0,009	-0,018	0,095	-0,019	0,738	-0,041	0,665
NOx	-0,011	0,044	-0,105	0,596	-0,031	0,484	-0,021	-0,628
NO_2	-0,014	0,054	-0,126	0,773	-0,057	-0,467	0,002	0,401
O_3	0,998	-0,051	-0,0003	0,023	0,022	0	0,001	0
PS	0,046	0,884	-0,439	-0,13	0,070	-0,005	-0,006	0,001
TE	0,031	0,158	0,181	-0,04	-0,967	-0,004	-0,042	-0,017
V-V	0,000	0,005	0,037	-0,005	0,052	-0,041	-0,996	-0,012
%Var	89.79	04.8	02.8	02	0,20	0,148	0,02	0.02
Cumul %Var	89.79	94.6	97.54	99.61	99.81	99.96	99.98	100

La première composante principale (CP1) exprime 89,79 % de la variance, cette composante est principalement caractérisée par un niveau de concentration élevé de l'ozone. La situation change pour CP2, celle-ci est influencée par des valeurs basses d'humidité relative, cette composante est aussi caractérisée par des valeurs élevées de température, et de hauts niveaux de concentration des particules en suspension (PM_{10}) (le paramètre le plus influençant pour CP2). CP3 est corrélé négativement avec d'humidité relative et les niveaux de concentration des particules en suspension, NO_2 et NOx. La quatrième composante principale est

positivement corrélée avec NO_2 (le paramètre le plus influençant pour CP4) et NOx, et négativement corrélée avec l'humidité relative.

4.6. Modélisation de l'influence des paramètres météorologiques sur la pollution atmosphérique à l'aide des réseaux de neurones artificiels

4.6.1. Introduction

Les réseaux de neurones artificiels (RNAs) sont considérés comme une branche de l'intelligence artificielle développés dans les années 50 visant à imiter l'architecture biologique du cerveau humain [105]. L'utilisation des réseaux neurones artificiels (RNAs) a été introduite par McCulloch et Pitts [106] dans leur proposition d'un modèle mathématique d'un neurone artificiel, tandis qu'en 1957, F.Rosenblatt a développé le premier modèle du perceptron, dont il a construit le premier neuro-ordinateur basé sur ce modèle qui a été appliqué au domaine de la reconnaissance des formes. En 1969 Minsky et Papert ont pu publier leur ouvrage, où ils ont montré les limitations théoriques du perceptron. Ces limitation concerne notamment l'impossibilité de traiter par ce modèle les problèmes non linéaires. Par ce fait la plupart des efforts consacrés aux réseaux de neurones artificiels ont été réorientés et les chercheurs ont quitté ce domaine de recherche. Seulement quelques chercheurs ont continué leurs efforts, plus notamment Teuvo Kohonen, Stephen Grossberg, James Anderson, et Kunihiko Fukushima [107]. L'intérêt pour les réseaux de neurones a réapparu seulement après que quelques résultats théoriques importants ont été atteints dans les années '80 (notamment la découverte de la retro-propagation de l'erreur), ainsi que les nouveaux développements de matériel qui ont augmenté les capacités de traitement. Depuis cette période plusieurs efforts ont été mis en œuvre dans ce domaine de recherche. La littérature relative aux RNAs est très étendue, fournissant une vue profonde et complète sur les RNAs [108], [109].

Un réseau de neurones artificiels est un système répartis en parallèles composés de plusieurs éléments de traitement non linéaires reliés ensemble, appelé neurones. Ces neurones sont présentés comme des modèles de neurones biologiques et en tant que composants conceptuels pour les circuits qui pourraient accomplir des tâches informatiques. Un réseau de neurones artificiels est capable de mémoriser des connaissances de façon expérimentale et de les rendre disponibles pour utilisation. Il ressemble au cerveau humain en deux points [108]:

1. la connaissance est acquise à travers d'un processus d'apprentissage;

2. les poids des connections entre les neurones sont utilisés pour mémoriser la connaissance.

Les réseaux de neurones servent aujourd'hui à toutes sortes d'applications dans divers domaines. Un aperçu des domaines ou les RNA ont étés d'un grand apport sont [110]:

1. Traitement de signal**:** égalisation non linéaire, l'élimination de bruit, reconnaissance de signaux radar ou sonar.
2. Automatique: identification/modélisation des systèmes non linéaire….etc.
3. Compression de données.
4. Systèmes de classifications ou de diagnostiques.
5. etc.

4.6.2. Types des RNAs

Les quatre types de RNA les plus utilisés sont énumérés ci-dessous [110], [108].

1. *Perceptron a une ou plusieurs couches cachées (PMC), ou le single or Multi-Layered Perceptrons (MLP)*: C'est le type de réseaux le plus utilisé, bon pour la classification, et très bon pour l'approximation de fonctions.

2. *Réseaux à base de fonction radial ou Radial-basis function (RBF)*: Les réseaux de neurones à fonctions de base radiales sont des réseaux de type *feedforward* avec une seule couche cachée. La particularité de ces réseaux réside dans le fait qu'ils sont capables de fournir une représentation locale de l'espace grâce à des fonctions de base radiales φ (.)

3. *Réseaux de Hopfield*: Le modèle de Hopfield est un réseau a couche unique, avec un retour des sorties sur les entrées. Ce modèle est basé sur le concept de mémoire Associative.

4. *Les cartes de Kohonen ou Kohonen Self-organising Maps (SOM)*: développé par Kohonen, ce modèle permet de coder des motifs présentés en entrée, tout en conservant la topologie de l'espace d'entrée.

4.6.3. Modèle de neurone et réseau

Un neurone est essentiellement constitué d'un intégrateur qui effectue la somme pondérée de ses entrées. Le résultat u de cette somme est ensuite transformé par une fonction de transfert φ qui produit la sortie y du neurone. Le résultat est comparé à un seuil et le neurone devient excité si ce seuil est dépassé. Le réseau de neurone est tolérant aux faute, du fait du grand nombre de neurones et de leurs interconnections. Ainsi, la défectuosité d'un élément mémoire (neurone) n'entraînera aucune perte réelle d'information, mais seulement une faible dégradation en qualité de toute l'information contenue dans le système [108]. Le modèle

mathématique d'un neurone artificiel est illustré à la figure 2.1. les d entrées du neurone correspondent au vecteur x = [x1x2 …xd]T, alors que le vecteur w = [w1,1w1,2 …w1,d]T représente le vecteur poids du premier neurone. La sortie u de l'intégrateur est donnée par l'équation suivante [109] :

$$u = \sum_{j=1}^{D} w_{1,j} x_j - b \qquad \textbf{(4.1)}$$

$$= w_{1,1}x_1 + w_{1,2}x_2 + ... + w_{1,D}x_D - b$$

Cette sortie correspond a une somme pondérée des poids et des entrées moins ce qu'on nomme le biais b du neurone. Le résultat u s'appelle le niveau d'activation du neurone, alors que le biais b s'appelle aussi le seuil d'activation du neurone. Lorsque le niveau d'activation atteint ou dépasse le seuil b, alors l'argument de f devient positif (ou nul). Sinon, il est négatif. On ajoute la fonction d'activation φ, la sortie du neurone sera donnée par l'équation:

$$y = \varphi(u) = \varphi(w^T x - b) \qquad \textbf{(4.2)}$$

Différentes fonctions de transfert pouvant être utilisées comme fonction d'activation du neurone. Les trois fonctions les plus utilisées sont les fonctions «seuil» (en anglais «hard limit»), «linéaire» et «sigmoïde». Il existe principalement deux types de neurones artificiels, Le neurone de Mc Culloch-Pitt qui se caractérise principalement par une fonction d'activation de type limitation 0 ou 1, et le perceptron qui se caractérise par une fonction d'activation différentiable et continue [109], la figure 4.4 représente également le modèle formel d'un neurone biologique.

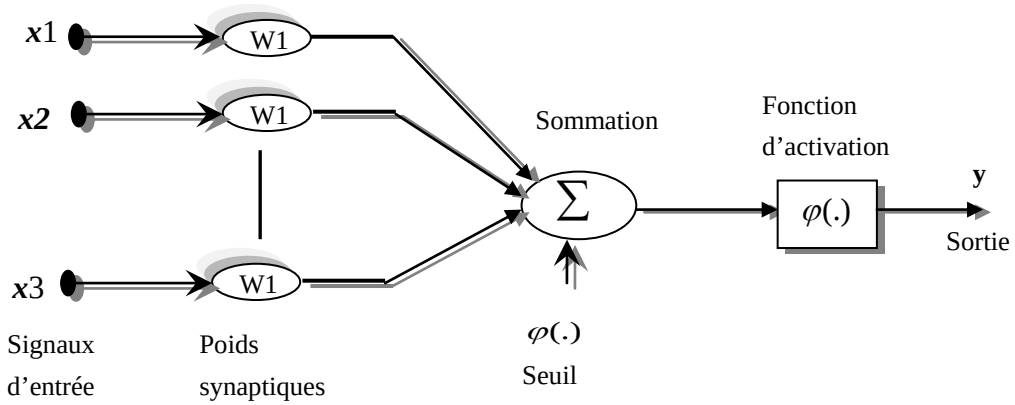


Figure 4.4–-Le modèle d'un neurone formel.

4.6.4. Topologies des réseaux de neurones

La topologie d'un réseau de neurone est définie par son architecture (ou structure) et la nature de ses connexions. L'architecture du réseau se compose de beaucoup d'éléments de traitement simples qui sont organisés en ordre de couches. Le réseau se compose habituellement d'une

couche d'entrée, des couches cachées et d'une couche de sortie. Sous sa forme simple, chaque neurone est relié à d'autres neurones de la couche précédente par des poids synaptiques adaptables. Il est à noter que les réseaux multicouches sont beaucoup plus puissants que les réseaux simples à une seule couche [109]. Quant à ces modes de connections, la distinction principale que nous pouvons faire est entre [111] :

- Les réseaux à propagation en avant (réseau feed-forward) : les neurones sont arrangés par couche. L'information circule d'une couche à la suivante. Il n'y a pas de connexion entre neurones d'une même couche ou couches précédentes (e.g figure 4.5 (a)).
- Réseau à connexions récurrentes : les connexions récurrentes ramènent l'information en arrière par rapport au sens de propagation défini dans un réseau multicouche (e.g figure 4.5 (b)).

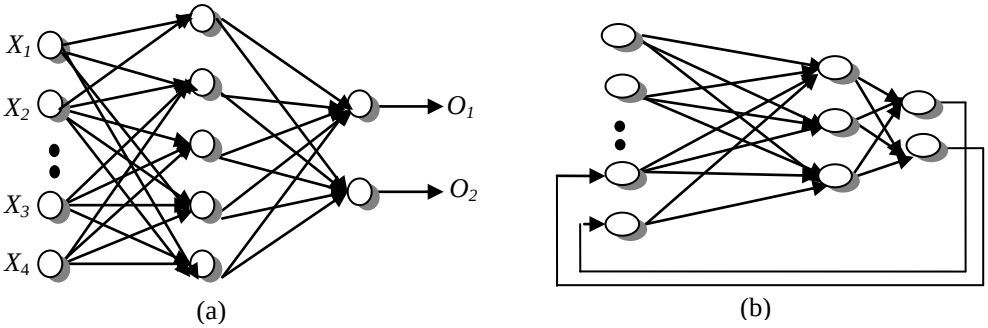


Figure 4. 5--Architecture :(a)d'un perceptron simple (b) et réseau à connexions récurrentes.

4.6.5. Apprentissage

Parmi les propriétés désirables pour un réseau de neurones artificiel, la plus fondamentale est surement la capacité d'apprendre de son environnement, d'améliorer ses performances à travers un processus d'apprentissage, idéalement, un réseau acquiert davantage de connaissances à chaque itération de l'algorithme d'apprentissage. Le rôle de l'apprentissage donc est de modifier les poids des connexions entre les neurones d'entrée et ceux de sortie, de manière à obtenir une réponse que l'on souhaite reproduire par le réseau de neurones [112]. Il existe divers algorithmes d'apprentissage pour ajuster les poids d'un réseau neuronal, et éventuellement sa topologie. Généralement l'architecture du réseau est définie par l'utilisateur, puis éventuellement optimisée par l'ordinateur. Tandis que l'ajustement des poids est totalement pris en charge par le système.

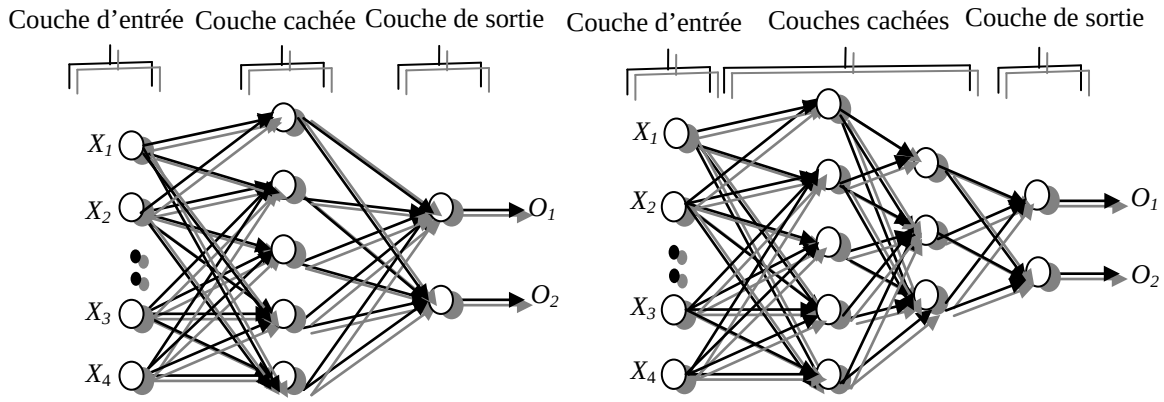
L'apprentissage peut s'effectuer de diverses manières. Il peut être supervisé ou non supervisé. Dans l'apprentissage supervisé, un superviseur (ou expert humain) fournit une valeur ou un vecteur de sortie (appelé cible ou sortie désirée) que le réseau de neurones doit associer au

vecteur d'entrée **x**. Tandis que dans l'apprentissage non supervisé les données ne contiennent pas d'informations sur une sortie désirée. Il n'y a pas de superviseur. Il s'agit de déterminer les paramètres du réseau de neurones suivant un critère à définir.

4.6.6. Le perceptron multicouche

Les Perceptrons Multicouches[3] (PMC) sont des réseaux de neurones pour lesquels les neurones sont organisés en couches successives, les connections sont toujours dirigées des couches inférieures vers les couches supérieures et les neurones d'une même couche ne sont pas interconnectés. Un neurone ne peut donc transmettre son état qu'au neurone situé dans une couche postérieure à la sienne. La topologie d'un perceptron multicouches est définie par son architecture (ou structure) et la nature de ses connexions. L'architecture d'un PMC peut alors être décrite par le nombre de couches et le nombre de neurones dans chaque couche (e.g figure 4.6). L'architecture d'un PMC est constituée de [113]:

- Un ensemble de neurones d'entrée dont le rôle est de recevoir les signaux externes et de les diffuser aux unités de la couche suivante. Les neurones d'entrée sont organisées en une couche appelée couche d'entrée.
- Une couche de sortie qui produit la réponse du réseau au signal d'entrée.
- Une ou plusieurs couches cachées se trouvant entre la couche d'entrée et la couche de sortie.



(a) Une couche cachée (b) Deux couches cachées

Figure 4.6--Perceptron multicouches feedforward.

Le Perceptron Multicouches est probablement le réseau neuronal le plus utilisé, en effet il a été prouvé par [114] qu'une fonction continue quelconque peut être arbitrairement bien approximée par un réseau feed-forward à couche cachée unique, ou chaque neurone dans la

[3] En anglais, ce type de réseau est appelé : Multi Layer Perceptron (MLP)

couche cachée a comme fonction d'activation une fonction non linéarité sigmoïdal (le perceptron). Cependant, certains problèmes (utilisant des neurones de types discontinus) ne peuvent être résolus avec une seule couche cachée, mais peuvent l'être par des réseaux à deux couches cachées. Alors, la détermination de l'architecture optimale se fait en utilisant les théories statistiques concernant le choix d'un modèle.

4.6.6.1. Apprentissage par retro-propagation

L'algorithme de rétro-propagation du gradient (RP) est certainement à la base des premiers succès des réseaux de neurones artificiels. Sa mise en application a permis au domaine du connexionnisme de sortir de la période de silence. Il figure aujourd'hui parmi les algorithmes d'apprentissage les plus utilisés [113]. L'apprentissage par rétro-propagation de l'erreur, se passe en deux phases alternées. Pendant la première phase l'entrée X est présentée et propagée en avant par le réseau pour calculer les valeurs de sorties y^p pour chaque neurone de sortie. La valeur de sortie du réseau est comparée à sa valeur désirée d_o, ayant pour résultat un signal d'erreur pour chaque neurone de sortie [110]. Dans la seconde phase, on applique l'algorithme de mise à jour, dans la dernière couche de poids, puis on propage les erreurs de sortie à travers cette couche de poids sur l'avant-dernière couche, ce qui permet alors de réitérer l'algorithme en mettant à jour l'avant-dernière couche de poids, et ainsi de suite. L'algorithme de rétro-propagation, nécessite toutefois que les fonctions d'activations des neurones soient continues et dérivables. Les fonctions qui sont le plus couramment utilisées sont probablement les fonctions de type sigmoïdal.

L'algorithme de retro-propagation procède à l'adaptation des poids neurone par neurone en commençant par la couche de sortie. Soit l'erreur observée *ej(n)* pour le neurone de sortie *j* et la donnée d'apprentissage n [111]:

$$e_j(n) = d_j(n) - y_j(n) \qquad \textbf{(4.3)}$$

Ou $d_j(n)$ correspond à la sortie désirée du neurone *j*, et $y_j(n)$ à sa sortie observée. Pour chaque neurone de sortie, il n'ya pas de problème d'implémentation de la règle, car l'erreur ($e = d - y$) est directement disponible. Par contre pour les neurones de la couche cachée (si l'on considère un réseau a une couche cachée), on est dans l'obligation de rétro propager l'erreur de sortie pour obtenir une erreur effective a la sortie de chaque neurone de la couche cachée. Le processus de calcul du gradient de l'erreur et d'adaptation des poids est répété jusqu'au moment où un minimum de l'erreur, ou un point qui en est suffisamment proche, est trouvé. La fin de la phase d'apprentissage peut être décidée lorsque l'erreur moyenne observée à la

sortie du réseau devient inférieure à un seuil prédéfini, ou encore lorsque l'amplitude moyenne du gradient de cette erreur devient très faible, puisque, par définition, le gradient est nul au minimum de la fonction de coût [108]. Pour en savoir plus sur le fonctionnement de l'algorithme de retro-propagation consulter [110].

4.6.6.2. L'apprentissage : un problème d'optimisation

L'apprentissage d'un réseau de neurones artificiels constitue un problème d'identification de modèle qui englobe deux problèmes d'optimisation sous-jacents [115]:

- le problème d'optimisation de l'architecture du modèle de réseau,
- le problème d'optimisation des paramètres du modèle de réseau.

Apres la détermination d'une topologie appropriée, le PMC doit subir une phase d'apprentissage dans le but de définir ces paramètres. Plusieurs méthodes peuvent être utilisées pour l'amélioration de l'algorithme d'apprentissage, tel que la méthode du moment, pas d'apprentissage adaptatif et les techniques d'ordre supérieur [110]. Durant la phase d'apprentissage, l'erreur commise diminue, jusqu'à tendre asymptotiquement vers zéro si l'architecture du réseau à été bien choisie. Mais il convient de remarquer que plus l'erreur d'apprentissage est faible, plus le réseau apprend à reconnaitre les formes qui lui sont présentées, mais il risque d'être capable de la généralisation. Dans ce qui va suivre nous discutons les problèmes engendrés par le sur apprentissage ainsi que l'utilisation d'un ensemble de donnée restreint lors de la modélisation.

4.6.6.3. Sur apprentissage

La capacité de généralisation est l'une des raisons qui motive l'étude et le développement des réseaux de neurones artificiels. Elle peut être définie par la capacité d'élargir les connaissances acquises après apprentissage à des données nouvellement rencontrées par le réseau de neurones. Cette capacité de généralisation est très liée à la notion de sur-apprentissage. Ces deux caractéristiques sont complètement antagonistes. On parle de sur-apprentissage quand le réseau a parfaitement appris les exemples proposés. Il sera donc incapable de généraliser [112]. L'algorithme d'apprentissage peut donc, être exécuté jusqu'a ce qu'un minimum global (ou local) soit atteint. Toutefois, si l'on procède pour chaque itération à une validation par une base d'essai non utilisée par l'apprentissage, il peut arriver un moment où l'amélioration de l'erreur sur la base d'apprentissage conduit à une augmentation de l'erreur sur la base de validation. On dit alors qu'il y a un sur apprentissage, "over training". Pour résoudre ce problème, une méthode d'apprentissage supervisé basée sur

l'erreur de validation est implémentée. 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 sure que ceci est un minimum global (ou du 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.

4.6.6.4. Validation croisée

La validation croisée consiste à extraire de la base d'apprentissage, une portion des observations qui serviront non pas à l'apprentissage, mais à l'évaluation après apprentissage de l'erreur commise par le réseau. Elle consiste donc à diviser l'ensemble de données en *n* sous ensembles [110], [108]. Il en résulte *n* modèles à identifier en utilisant *n* sous ensembles de données. Prendre pour chaque modèle un sous ensemble pour la validation et utiliser la concaténation des sous ensembles restant pour l'apprentissage. A la fin de la validation croisée on obtient *n* modèles. On peut alors choisir le meilleur ou le plus appropriée à notre région opérative, ou alors combiner tous les modèles obtenus. Une combinaison linéaire peut être utilisée à cet effet. Cette méthode s'avère plus efficace quand la taille de l'espace de données est restreinte et/ou limité.

4.6.6.5. Evaluation de la qualité d'un modèle

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) [116], [117].

L'utilisation des indicateurs de performance pour évaluer et comparer les modèles neuronaux a été initialement discutée par willmot et al dans [118]. Ainsi, leurs recommandations ont été suivies par plusieurs chercheurs [116], [117], [119]. 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 le moyen des erreurs absolues pris sur l'ensemble des données prévus. cet indicateur est calculé selon l'équation :

$$\frac{|p_1 - a_1| + ... |p_n - a_n|}{n} \quad \textbf{(4.4)}$$

Où *n* est le nombre de points. L'avantage de MAE est qu'il est moins sensible que l'erreur carrée aux valeurs aberrantes et il rapporte également les erreurs dans la grandeur et la balance originales.

- *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' RMSE est calculé selon l'équation :

$$\sqrt{\frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{n}} \quad \textbf{(4.5)}$$

RMSE donne une indication quantitative sur l'erreur de simulation obtenue pendant la phase de modélisation. La RMSE 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 :

$$1 - \frac{\sum_i |p_i - a_i|^2}{\sum_i (|p_i - \bar{a}| + |a_i - \bar{a}|)^2} \quad \textbf{(4.6)}$$

- 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:

$$r = \frac{S_{PA}}{\sqrt{S_p S_A}} \quad \textbf{(4.7)}$$

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

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

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

- 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.

$$\frac{|p_1 - a_1| + ... + |p_n - a_n|}{|a_1 - \overline{a}| + ... + |a_n - \overline{a}|} \qquad \textbf{(4.8)}$$

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 être haut, 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.

4.6.7. Application des RNAs

Il a été montré dans la section 4.3, que les niveaux de concentration des polluants étudiés varient d'une saison à l'autre influencés par plusieurs paramètres tels que les conditions météorologiques, les activités industrielles et le trafic routier. Par conséquent, les niveaux de corrélations entre les concentrations des polluants et les paramètres qui les influencent sont mis en cause. Par conséquent l'approche que nous avons proposée est basée sur l'utilisation des clusters météorologiques pour étudier l'influence des paramètres météorologique sur la pollution atmosphérique dans la région d'Annaba. Les réseaux de neurones peuvent fournir une solution intéressante pour les problématiques de modélisation et de prédiction de la qualité de l'air. En effet, et comme nous avons montré précédemment, leur utilisation ne nécessite pas l'existence d'une modélisation formelle de la qualité de l'air. Par ailleurs, leurs capacités de mémorisation, d'apprentissage et d'adaptation représentent des fonctions très utiles pour tout système de prédiction ou de modélisation de la qualité de l'air. Afin de modéliser l'impact des paramètres météorologiques sur la pollution atmosphérique à Annaba nous avons choisis d'utiliser une architecture classique basée sur le perceptron multicouche (PMC) avec une seule couche cachée. Le réseau est composé d'une couche d'entrée, d'une couche cachée et de la couche de sortie. D'après [120] ce type de réseaux de neurones artificiels est le plus adéquat pour modéliser l'impact des paramètres météorologiques sur la pollution atmosphérique. En effet les modèles de PMC sont largement utilisés dans la modélisation et la prédiction des concentrations des polluants d'air puisqu'ils peuvent capturer

efficacement les rapports fortement non-linéaires entre les variables, et ses performances sont meilleures une fois comparée aux autres modèles linéaires et neuronaux [5], [6]. En outre, le PMC peut être formé pour approximer n'importe quelles fonctions fortement non-linéaires sans en connaitre au préalable la distribution de données. Dans ce qui suit, nous discutons l'architecture et l'apprentissage du PMC élaboré pour modéliser l'influence des paramètres météorologique sur la pollution atmosphérique dans la région d'Annaba.

4.6.7.1. Architecture du modèle neuronal

Dans l'objectif d'évaluer les performances de l'approche proposée, un modèle de PMC pour chaque polluant dans chaque cluster météorologique de la région d'Annaba est créé. Nous avons choisi de construire un modèle pour chaque polluant pour réduire au maximum la quantité des données manquantes, par conséquent prendre le maximum de données pour modéliser chaque polluant. La topologie d'un PMC est définie par son architecture (ou structure) et la nature de ses connexions. L'architecture proposée pour chaque modèle neuronal est le résultat de toute une phase de comparaison et de test des déférentes topologies possibles. Cette topologie est basée sur une architecture classique : le perceptron multicouche (PMC) avec une seule couche cachée. Cette architecture est décrite par 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 dont le nombre de neurones est un choix arbitraire. La figure 4.7 représente l'architecture du PMC utilisé pour modéliser le comportement de l'ozone au niveau du premier cluster météorologique.

Quelques tests avec des couches cachées modifiées ont été effectués afin d'observer une éventuelle incidence sur les résultats obtenus. Ainsi pour tirer profit au maximum des RNAs, il est primordial de définir intelligemment le nombre de neurones et leurs fonctions d'activation dans chacune des couches crées.

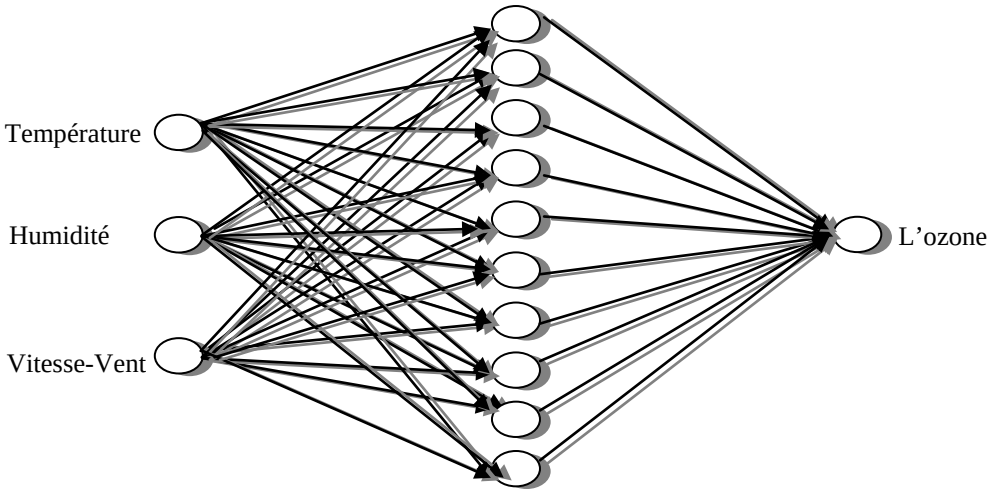


Figure 4.7--structure du PMC de l'ozone du premier cluster.

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 se compose 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. 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) sont calculés. Par la suite, on retient 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 premier cluster météorologique. Selon ce tableau, le PMC le plus adéquat est obtenu pour un nombre de neurone dans la couche cachée égale à dix.

Tableau 4.5-Les performances des différents PMC, selon le nombre de neurone.

Neurone	*Nombre d'époque*	*Erreur de validation*	*IA*
09	350	4,52 e-5	0,993
10	350	2,97 e-5	0,995
12	350	4,42 e-4	0,917
13	350	4,16 e-4	0,900
14	350	5,17 e-5	0,975
15	350	2,99 e-4	0,935
16	350	0,00064	0,897
18	350	0,00185	0,921

4.6.7.2. Apprentissage des modèles neuronaux

La base de données utilisée pour créer chaque modèle neuronal contient des paramètres météorologiques (la température ambiante de l'air, l'humidité relative et la vitesse du vent) et le polluant étudié. L'apprentissage dans ce cas est supervisé, car on dispose d'un ensemble d'exemples (paramètres météorologiques) auxquels sont associés des nivaux de concentration des polluants. La période considéré (2 ans) semble être restrictif, 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 sure qu'il s'agit d'un minimum global (ou du 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 à notre région opérative. La figure 4.8, représente l'évolution de l'erreur d'apprentissage et de validation pour le PM_{10} dans le cluster3, ainsi l'évolution des erreurs d'apprentissage et de validation pour l'ensemble des polluants dans tous les clusters

météorologiques est représentée par la figure 4.9. Les résultats obtenus sont représentés au tableau 4.7, Les indicateurs statistiques décrits dans la section 4.5 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.

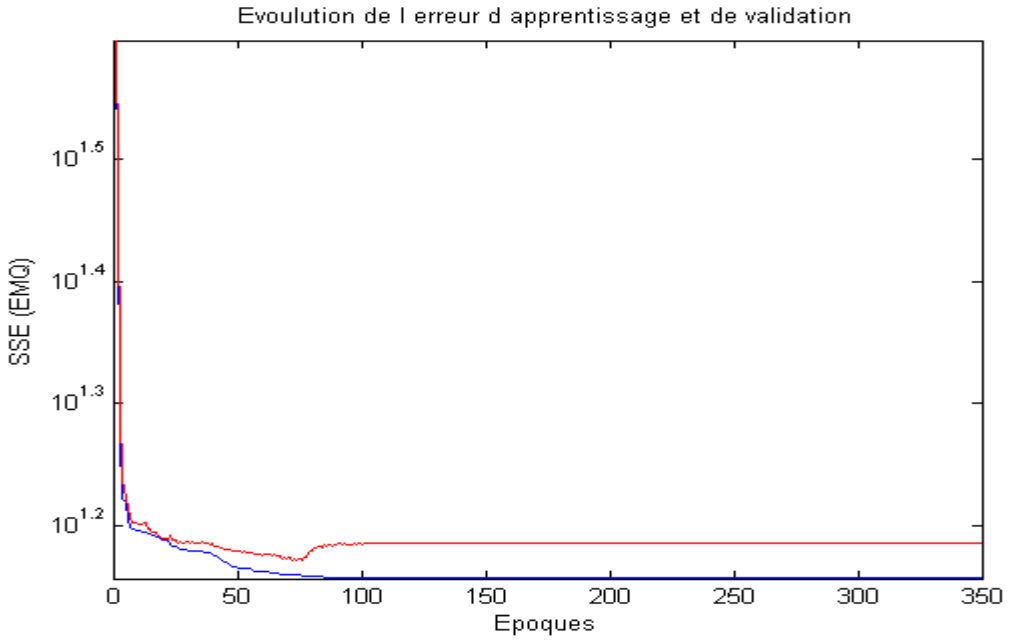


Figure 4. 8--Evolution de l'erreur d'apprentissage et de validation pour PM_{10}dans le cluster3.

Les performances de chaque modèle neuronal ont été évaluées à l'aide de l'indice de l'accord (IA). L'intervalle des valeurs de l'indice (IA) pour l'ensemble des polluants dans le premier cluster est de 63.4% pour NO à 99.5% pour l'ozone, avec un moyen de 77.41% pour l'ensemble des polluants. Les mesures de performances pour le deuxième cluster sont presque similaires à ceux du cinquième cluster météorologique, dont les valeurs moyennes de l'indice (IA) sont respectivement 74,2% et 74,9%. L'intervalle de valeurs de l'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. 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 74%. Ceci montre également que les niveaux de concentration des polluants sont non linéairement liés aux 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 elle est plus proche de zéro. L'intervalle des valeurs de l'RMSE pour le premier cluster météorologique est de 4,5% à 11.5% pour l'ensemble des polluants. En effet les valeurs de l'erreur RMSE

sont proches pour l'ensemble des polluants dans tous les clusters. Ces résultats sont très satisfaisant et montre l'efficacité de l'approche considérée pour la modélisation des émissions des polluants dans un environnement urbain.

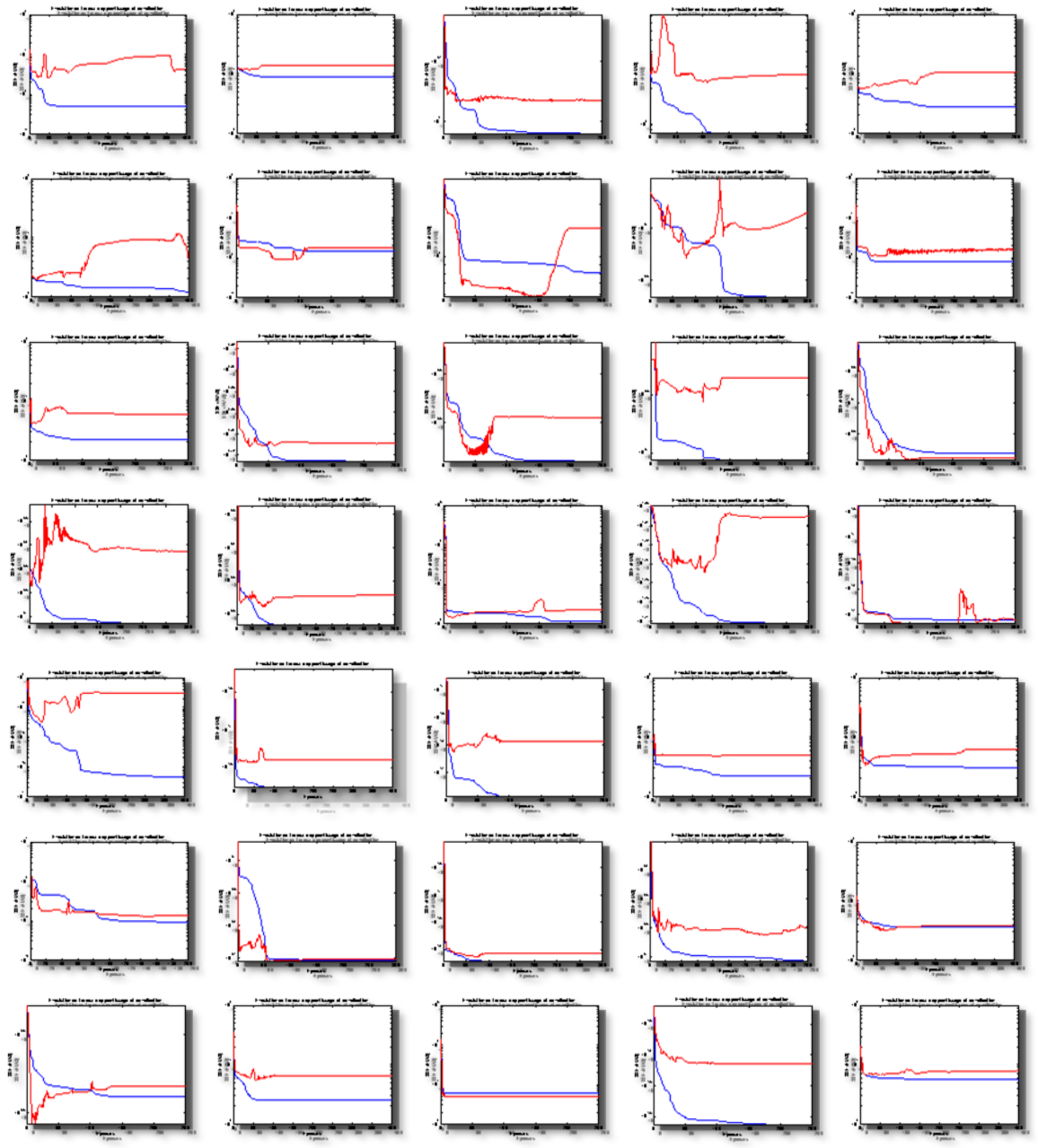


Figure 4.9-Evolution de l'erreur d'apprentissage et de validation pour l'ensemble des polluants dans les clusters météorologiques.

Le pouvoir de généralisation des modèles neuronaux a été également testé par le cœfficient de corrélation. En se basant sur les résultats obtenus, les cœfficients 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 de mesures est de 0,578 pour le troisième cluster à 0,668 pour le premier cluster

météorologique. Selon le tableau 4.9 qui montre également les résultats moyens des indicateurs statistiques pour chaque polluant, l'ozone (O_3) est le polluant le plus corrélé avec les paramètres météorologiques.

Tableau 4.6-Résultats des indicateurs statistiques pour les modèles neuronaux.

	CO	***NO***	***NOx***	***NO_2***	***O_3***	***PM_{10}***	***SO_2***
Cluster1							
IA	0,802	0,634	0,642	0,654	0,995	0,975	0,717
RMSE	0,066	0,076	0,115	0,110	0,054	0,045	0,093
RAE	0,494	0,919	0,876	0,907	0,202	0,328	0,745
MAE	0,021	0,042	0,083	0,082	0,027	0,029	0,065
Corr.	0,700	0,509	0,505	0,473	0,943	0,952	0,599
Cluster2							
IA	0,746	0,680	0,615	0,739	0,849	0,771	0,797
RMSE	0,128	0,036	0,054	0,054	0,112	0,031	0,069
RAE	0,691	0,877	0,803	0,776	0,588	0,904	0,693
MAE	0,060	0,013	0,030	0,037	0,085	0,018	0,027
Corr.	0,619	0,584	0,504	0,607	0,758	0,669	0,698
Cluster3							
IA	0,691	0,842	0,764	0,771	0,764	0,501	0,622
RMSE	0,098	0,021	0,040	0,081	0,098	0,085	0,238
RAE	0,800	0,923	0,783	0,742	0,672	0,929	0,890
MAE	0,051	0,014	0,024	0,052	0,057	0,057	0,156
Corr.	0,544	0,743	0,655	0,663	0,616	0,394	0,431
Cluster4							
IA	0,785	0,683	0,647	0,603	0,846	0,691	0,793
RMSE	0,060	0,063	0,068	0,128	0,128	0,076	0,104
RAE	0,748	0,887	0,874	0,854	0,635	0,844	0,621
MAE	0,031	0,022	0,035	0,083	0,096	0,049	0,075
Corr.	0,678	0,514	0,540	0,453	0,744	0,573	0,655
Cluster5							
IA	0,605	0,800	0,786	0,793	0,854	0,814	0,596
RMSE	0,085	0,029	0,046	0,041	0,169	0,094	0,134
RAE	0,781	0,981	0,796	0,772	0,587	0,713	0,821
MAE	0,034	0,016	0,026	0,023	0,124	0,058	0,102

Corr.	0,467	0,693	0,662	0,678	0,758	0,705	0,471

Tableau 4.7-Les indicateurs statistiques moyens pour tous les polluants dans chaque cluster.

	Cluster1	***Cluster2***	***Cluster3***	***Cluster4***	***Cluster5***
IA	0,774	0,742	0,707	0,721	0,749
RMSE	0,079	0,069	0,094	0,089	0,085
RAE	0,638	0,761	0,819	0,780	0,778
MAE	0,049	0,038	0,058	0,055	0,054
Corr.	0,668	0,634	0,578	0,593	0,633

Tableau 4.8-Les résultats moyens des indicateurs statistiques pour chaque polluant.

	CO	***NO***	***NOx***	***NO_2***	***O_3***	***PM_{10}***	***SO_2***
IA	0,725	0,727	0,690	0,712	0,861	0,750	0,705
RMSE	0,087	0,045	0,064	0,082	0,112	0,066	0,127
RAE	0,702	0,917	0,826	0,810	0,536	0,743	0,754
MAE	0,039	0,021	0,039	0,055	0,077	0,042	0,085
Corr.	0,601	0,608	0,573	0,574	0,763	0,658	0,570

Les résultats obtenus (en terme du cœfficient de corrélation) prouvent qu'approximativement 62,12% de la variation des variables dépendantes (concentration des polluants) peuvent être expliquée par les variables indépendantes (paramètres météorologiques). D'une manière générale, les résultats des indicateurs : RAE, MAE montrent également que l'approche proposée est très utile pour la modélisation du comportement des niveaux de concentration des polluants en se basant sur les clusters météorologiques.

4.7. Conclusion

La base de données collecté par la station «SAMASAFIA» a été utilisé pour étudier l'influence des paramètres météorologique sur la pollution atmosphérique. Les relations entre les niveaux des concentrations de la qualité de l'air ambiantes et les paramètres météorologiques ont été évaluées en utilisant La méthode ACP, cette méthode s'est avéré un outil efficace pour l'investigation de la pollution atmosphérique et les données météorologiques. Les résultats de l'analyse en composante principale ont indiqué des rapports fondamentaux parmi les variables de la base de données ou plusieurs significations et conclusions ont été tirées. Les réseaux de neurones artificiels ont été utilisés pour modéliser le

comportement non linéaire des paramètres météorologiques et l'ensemble des polluants considéré dans cette étude. Bien que la période considéré (2 ans) semble être restrictif et le manque d'horaire continu de données, il s'avère qu'un réseau de neurone artificiel avec une seule couche cachée basée sur l'algorithme de rétro propagation décrit précédemment est très efficace pour modéliser le comportement non linéaire des émissions de polluants dans chaque cluster météorologique.

Conclusion générale.

Conclusion générale

Dans ce mémoire, nous avons présenté une approche non supervisée pour l'identification des types de jours météorologiques pour la région d'Annaba. En raison de la quantité énorme de données fournie par la station météorologique de l'aéroport d'Annaba, des outils efficaces d'analyse sont indispensables pour extraire les caractéristiques utiles, fournissant des informations plus simples et plus maniables. Bien que la méthode SOM a prouvé son efficacité pour le regroupement des données multidimensionnelles, il est difficile d'identifier les groupes et leurs frontières clairement si le nombre de nœuds de la carte est grand. Une approche à deux niveaux de classification a été utilisée dans le cadre de cette étude : dans le premier niveau la méthode SOM a été utilisé pour réduire la dimensionnalité des données et extraire les prototypes météorologiques qui sont ensuite regroupés a l'aide d'un deuxième niveau de classification. Plusieurs méthodes de classification non supervisée ont été utilisées pour regrouper les vecteurs référents de la carte de kohonen tel que : l'algorithme PAM (Partitioning Around Medoids), K-moyennes, et la classification hiérarchique (méthode de Ward). Dans le but de découvrir la méthode la plus appropriée pour le regroupement des unités de la SOM, nous avons procédé à une comparaison basée sur les critères de validation des regroupements. Selon les résultats obtenus des indices de validation de regroupement, k-moyennes s'avère l'algorithme le plus approprié pour regrouper les nœuds de la carte de kohonen pour ce type de données. Il a été remarqué qu'une classification des données météorologiques uniquement par k-moyennes engendre un temps de calcul beaucoup plus grand que celui généré par l'approche à deux niveaux de regroupement. Même avec un nombre d'échantillons relativement petit la plupart des algorithmes de classification automatique (spécialement les algorithmes hiérarchiques) deviennent intraitable et très lourd, pour cette raison, il est nécessaire de regrouper un ensemble de prototypes plutôt que de regrouper directement les données.

Le nombre optimal de clusters météorologiques a été sélectionné en utilisant deux catégories de critères de validation de la classification automatique (critères interne et externe). Ainsi, six clusters météorologiques différents (C1-C6) ont été détectés avec des frontières claires, donc les paramètres météorologiques de chaque cluster peuvent être facilement interprétés. L'approche proposée a été également testée en utilisant la base de données collectée par la station SAMASAFIA de Sidi-Amar.

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 ou 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 l'approche proposée précédemment peut être également utilisée pour l'identification des types de jours de pollution pour l'étude et l'analyse des paramètres atmosphériques dans la région. Ainsi, il serait intéressent d'utiliser les différentes topologies des cartes de kohonen (rectangulaire, hexagonal, toroidal, 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 outils et méthodes de la classification floue pour développer de nouvelles approches de classification basées sur les cartes de kohonen et les méthodes floues.

Références.

[1] A.K. Gupta, Kakoli Karar, S. Ayoob, Kuruvilla John., "Spatio-temporal characteristics of gaseous and particulate pollutants in an urban region of Kolkata, India", *Atmospheric Research*, vol. 87, pp. 103-115, 2008.

[2] Ziomas, I.C., Melas, D., Zerefos, C.S., Bais, A.F., "Forecasting peak pollutant levels from meteorological variables", *Atmospheric Environment*, vol. 29, pp. 3703–3711, 1995.

[3] Kalkstein, L.S., "A new approach to evaluate the impact of climate on human mortality," *Environmental Health Perspectives*, vol. 96, pp. 145–150, 1991.

[4] Scott c. Sheridan, "The Redevelopment of a Weather-Type Classification Scheme for North America", *Int. J. Climatol*, vol. 22, pp. 51–68, 2002.

[5] M.W. Gardner, S.R. Dorling, "Neural network modelling and prediction of hourly NOx and NO_2 concentrations in urban air in London", *Atmospheric Environment*, vol. 31, pp. 709–719, 1999.

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

[7] Mebirouk Hayet, Mebirouk-Bendir Fatiha, "principaux acteurs de la pollution dans l'agglomération de annaba. Effets et développements", Colloque International sur l'Eau et l'Environnement, alger, 2007.

[8] Fadel Derradji , Nacer Kherici , Saadane Djorfi , Michèle Romeo , Raoul Caruba, "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", *La Houille Blanche*, vol. 1, pp. 73-80, 2005.

[9] *http://www. météofrance.com.*

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

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

[12] Laurent Candillier, Contextualisation, "visualisation et évaluation en apprentissage non supervisé", thèse de doctorat, Dept. Informatique, Univ. Charles de gaules-Lille 3, 2006.

[13] Eder, B.K., Davis, J.M., Bloomfield, P, "An automated classification scheme designed to better elucidate the dependence of on meteorology," *Journal of Applied Meteorology*, pp. 1182–1199, 1994.

[14] Irini Reljin, Branimir Reljin, Gordana Jovanovi, "clustering and mapping spatial-temporal datasets using som neural networks," *Journal Of Automatic Control*, university of Belgrade, vol. 13(1), pp. 55-60, 2003.

[15] kwan,C,R Xu and L. Haynes , "a new data clustering and its applications," *in Proc of SPIE-the international society for optical engineering*, vol. 4384, pp. 1-5, 2001.

[16] Laitinen,N.,J.Ranatanen,S.Laine,O.Antikainen,E. Rasanen, S. Airaksinen and J.Yliruusi, "visualization of particle size and shape distributions using self-organizing maps," chemometrics and intelligent laboratory systems, vol. 62, pp. 47-60, 2002.

[17] B. C. Hewitson, R. G. Crane, "Self-organizing maps: applications to synoptic climatology," *Climate Research*, vol. 22, pp. 13–26, 2002.

[18] Tereza Cavazos, "Using self-organizing maps to investigate extreme climate events: an application to wintertime precipitation in the balkans," *Journal of climate*, vol. 13, pp. 1718–1732, 2000.

[19] Ignacio J. Turias, Francisco J. Gonzalez, M Luz Martin, Pedro L. Galindo, "A competitive neural network approach for meteorological situation clustering," *Atmospheric Environment*, vol. 40, pp. 532–541, 2006.

[20] [1] P.Arabie and L.J. Hubert. "An overview of combinatorial data analysis. Clustering and classification", pages 5-63, 1996.

[21] V. Ganti, J. Gehrke, and R. Ramakrishnan., "CACTUS-clustering categorial data using summaries", in KDD'99:proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 73-83, New York, USA, 1999.

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

[23] Suwardi Annas, Takenori Kanai and Shuhei Koyama, "Principal Component Analysis and Self Organizing Map for Visualizing and Classifying Fire Risks in Forest regions, agricultural information research" *Atmospheric Environment*, vol. 16(2), pp. 44-51, 2007.

[24] Aguilera, P.A., Frenich, A.G., Torres, J.A., Castro, H., Vidal, J.L.M., Canton, M, Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. *Water Research*, vol. 35, pp. 4053–4062, 2001.

[25] J. Tison , Y.-S. Park , M. Coste , J.G. Wasson , L. Ector , F. Rimet, F. Delmas, "Typology of diatom communities and the influence of hydro-ecoregions: A study on the French hydro system scale", *Water Research*, vol. 39, pp. 3177–3188, 2005.
[26] Mohamed Tarek Khadir, Damien Fay,and Ahmed Boughrira, "Day Type Identification for Algerian Electricity Load using Kohonen Maps", *transaction on engineering, computing and technology*, vol. 15. pp. 1305-5313, 2006.
[27] jain,A.K, Murty,M.,and Flynn, P. "Data clustering: a review", *ACM computing surveys*, vol. 31(3), pp. 264-322, 1999.
[28] pavel berkhin, "survey of clustering data mining techniques". Technical report, Accrue Software, San Jose, CA, 2002.
[29] Grabmeier et Rudolph, J. Grabmeier et A. Rudolph. "Techniques of cluster algorithms in data mining". *Data Mining and Know ledge Discovery*, vol. 6(4), pp. 303-360, 2002.
[30] Sébastien Guerif, "Réduction de dimension en apprentissage numérique non supervisé", thèse de doctorat, Univ. Paris 13, 2006.
[31] Mounzer Boubou, "contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions", thèse de doctorat, Univ Claude Bernard-lyon 1, 2006.
[32] Alexandre Blansché, "classification non supervée avec pondération d'atributs par des méthodes évolutionnaires", thése de doctorat, Discip. Informatique, Univ. Louis Pasteur, 2006.
[33] François-Xavier Jollois, "contribution de la classification automatique à la fouille de données", thèse de doctorat, Dept. Informatique, Univ.de Metz, 2003.
[34] Zhao, Y. and Karypis, G. "criterion functions for document clustering: Experiments and analysis". Technical Report TR:01-40, Dept. of computer science, university of Minnesota, Minneapolis,MN, 2002.
[35] Maria halkidi, Yannis batistakis, Michalis vazirgiannis, On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, vol. 17:2/3, pp. 107–145, 2001.
[36] Rezaee, R., Lelieveldt, B.P.F., and Reiber, J.H.C, "A New Cluster Validity Index for the Fuzzy c-Mean". *Pattern Recognition Letters*, vol. 19, pp. 237–246, 1998.
[37] M. Halkidi, M. Vazirgiannis, Y. Batistakis. "Quality scheme assessment in the clustering process", In Proceedings of PKDD, Lyon, France,2000.
[38] Kohonen , T, *The self-organising map*, Proceedings IEEE, , 78 (9), 1990.
[39] G. Dreyfus, J.m. Martinez, M. Samuelides, M.b. Gordon, F. Badran, S. Thiria, L. Hérault, *Reseaux De Neurones : Méthodologie Et Application*, ISBN : 2-212-11464-8, Eyrolles 2004.
[40] Aicha El Golli , Brieuc Conan-Guez , and Fabrice Rossi, "A Self Organizing Map for dissimilarity data", In proc of IFCS'2004.
[41] Praveen Boinee, "insights into machine learning: data clustering and classification algorithms for astrophysical experiments", Ph. D. thesis, Dept. of Mathematics and Computer Science, Univ. of udine – italy, 2006.
[42] mikko t. kolehmainen, Data exploration with self-organizing maps in environmental informatics and bioinformatics, Ph. D. thesis, Dept of Computer Science and Engineering, Helsinki University of Technology, 2004.
[43] Juha Vesanto, SOM-Based Data Visualization Methods, *Intelligent Data Analysis*, Vol. 3(2), pp. 111-126, 1999.
[44] Johan Himberg, A SOM Based Cluster Visualization and Its Application for False Coloring, *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, vol. 3, pp. 587-592, 2000.
[45] Patrick ROUSSET, "Applications des algorithmes d'auto-organisation à la classification et à la prévision", thése de doctorat, Univ. Paris I, 1999.
[46] Vincent Lemaire, Cartes auto-organisatrices pour l'analyse de données, proc. COnférence en Recherche d'Information et Application (CORIA).2006.
[47] Site de l'université d'Helsinki *http://www.cis.hut.fr/research/som-research.*
[48] Patrice Wira, Réseaux neuromimétique, "modularité et statistiques: estimation du mouvement pour l'asservissement visuel de robots", thése de doctorat, discip. E.E.A. Univ. haute-alsace, 2002.
[49] Juha Vesanto, "Neural Network tool for data mining: SOM toolbox", *Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000)*, Oulu, Finland, pp. 184-196, 2000.
[50] Olli Simula, Juha Vesanto, Esa Alhoniemi and Jaakko Hollmén, "Analysis and Modeling of Complex Systems Using the Self-Organizing Map", Chapter in Neuro-Fuzzy Techniques for Intelligent Information Systems, 1999.
[51] SAMUEL KASKI, "Data Exploration Using Self-Organizing Maps", Doctor of Technology thesis, Univ. Helsinki, 1997.
[52] A. Ultsch and H. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. *In Proc. INNC'90, Int. Neural Network Conf.*, pages 305-308, Dordrecht, Netherlands, 1990.

[53] V. Tryba, S. Metzen, and K. Goser. “Designing basic integrated circuits by self-organizing feature maps”, *In Neuro-Nimes '89. International Workshop. Neural Net-works and their Applications*, pages 225- 235, 1989.
[54] Juha Vesanto and Esa Alhoniemi , “*Clustering of the Self-Organizing Map”*, Volume 11(3) of IEEE Transactions on Neural Networks, special issue on data mining, 2000.
[55] Bradley, P. and U. Fayyad, “Refining initial points for K-means clustering”, *International Conference on Machine Learning* (ICML-98), 1998.
[56] Kanungo, T., D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, A. Wu. "An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(7), pp. 881-892, 2002.
[57] S.P. Bradley, U. M. Fayad, and C.Reina. Scaling clustering algorithms to large databases. In Knowledge Discovery and Data Mining,pages 9-12, 1998.
[58] L. Kaufman and P.J. Rousseeuw. “Finding groups in data”. *John Wiley and Sons*, Inc., 1990.
[59] Ng, R. T. and Han, J. “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, 1994.
[60] D. Chessel, J. Thioulouse & A.B. Dufour, “introduction à la classification hiérarchique”, Rapport technique, 2004.
[61] Lotfi KHODJA, Contribution à la Classification Floue non Supervisée, thèse de doctorat, Discip. Electronique - Electrotechnique – Automatique, Univ de Savoie, 1997.
[62] Wong M.A. “A hybrid clustering method for identifying high density clusters”, *J. of Amer. Statist. Assoc*., vol. 77, pp. 841-847, 1982.
[63] Dave,R.N., “Validating Fuzzy Partitions Obtained Throughc-Shells Clustering”, *Pattern Recognition Letters*, vol. 17, pp. 613–623, 1996.
[64] Dimitriadou, E, Dolnicar, S & Weingassel, A, “An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets”, *Psychometrika*,, vol. 67(1), pp. 137-160, 2002.
[65] Famili, A., Liu, G., Liu, Z., “Evaluation and Optimization of Clustering in Gene Expression Data Analysis”, *Journal of Bioinformatics*, 2003.
[66] Gengxin Chen, Saied A.Jaradat, Nila Banerjee, “evaluation and comparison of clustering algorithms in anglyzing es cell gene expression data”, *Statistica Sinica* vol. 12, pp. 241-262, 2002.
[67] Theodoridis, S.and Koutroubas, K.Pattern Recognition. Academic Press, 1999.
[68] Halkidi,M., Vazirgiannis, M.,andBatistakis, I., “Quality Scheme Assessment in the Clustering Process”, *in Proceedings of PKDD*,Lyon, France, 2000.
[69] David L. Davies and Donald W. Bouldin, “A Cluster Separation Measure”. *IEEE Transactions On Pattern Analysis And Machine Intelegence*, *PAMI*, vol. 1(2), pp. 224-227, 1997.
[70] Kaijun Wang, Junying Zhang, Hongyi Zhang and Tao Guo, Estimating the Number of Clusters via System Evolution for Cluster Analysis of Gene Expression Data, Technical report. Xidian University, P. R. China, 2007.
[71] A. Strehl. “Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining”. Ph.D thesis, The University of Texas at Austin, May 2002.
[72] Lam, K.C., Cheng, S., A synoptic climatological approach to forecast concentrations of sulfur dioxide and nitrogen oxides in Hong Kong. *Environmental Pollution*, vol. 101, pp. 183–191.1998.
[73] Alex J. Cannon, Paul h. Whitfield Synoptic, Edward r. Lord, “Synoptic Map-Pattern Classification Using Recursive Partitioning and Principal Component Analysis”, *Monthly Weather Review*, vol. 130, pp. 1187-1206. 2002.
[74] Bjorn a. Malmgren, Amos Winter, “Climate Zonation in Puerto Rico Based on Principal Components Analysis and an Artificial Neural Network”, *Journal of Climate*, vol. 12, pp. 977-985, 1999.
[75] S.M. Shiva Nagendra, Mukesh Khare, “Principal Component Analysis of Urban Traffic Characteristics and Meteorological Data”, *Transportation Research Part D* 8. pp. 285-297. 2003.
[76] Ph. Preux, “Fouille de données : Notes de cours”, Rapport technique. Univ. de Lille 3, 2007.
[77] P. Demartines and J. Hérault. “Curvilinear component analysis: A self-organizing neural network for non-linear mapping of data sets”. *IEEE Transactions on Neural Networks*, vol. 8(1), pp. 148–154, 1997.
[78] Hautaniemi, S,O.Yli-Harja,J. Astola,P. Kauraniemi, A. Kallioniemi, M. Wolf, J. Ruiz, S. Mousses and O. kallioniemi , “Analysis and visualisation of gene expression microarray data in human cancer using self-organizing maps,” *Machine Learning*, vol. 52, pp. 45-66, 2003.
[79] Kiang,M.Y.,M.Y.Hu and D.M.Fisher, “An extended self organizing map network for market segmentation a telecommunication example,” *Decision Support Systems*, *DECSUP*-11061- N° of pages 12, 2004.
[80] Réseau de surveillance de la qualité de l’air, Bilan annuel sur la qualité de l’air pour l’année 2007, « SAMASAFIA », 62p.
[81] Elminir, H.K., “Dependence of urban air pollutants on meteorology”, *Science of the Total Environment*, vol. 350, pp. 225–237. 2005.

[82] M. Sharma, S. Aggrawal, P. Bose, Meteorology-based forecasting of air quality index using neural network, *Int. Conference ICONIP'02-SEAL'02-FSKD'02 on Neural Network*, Singapore, pp. 374–378, 2002.

[83] Kostas D. Karatzas, Stamoulis Kaltsatos. "Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece". *Simulation Modelling Practice and Theory*, vol. 15, pp. 1310-1319, 2007.

[84] M Kermani , K Naddafi , M Shariat , AS Mesbah, "Chemical Composition of TSP and PM_{10} and their Relations with Meteorological Parameters in the Ambient Air of Shariati Hospital District", *Iranian J Publ Health*, Vol. 32, No. 4, pp.68-72, 2003.

[85] Magne Aldrin, Ingrid Hobæk Haff. "Generalised additive modelling of air pollution, traffic volume and meteorology". *Atmospheric Environment*, vol. 39, pp. 2145-2155, 2005.

[86] Levy,J.I.,Bennett,D.H.,Melly,S.J.,Spengler,J.D., "Influence of traffic patterns on particulate matter and polycyclic aromatic hydrocarbon concentrations in Roxbury, Massachusetts", *Journal of Exposure Analysis and Environmental Epidemiology* vol. 13, pp. 364–371, 2003.

[87] P. D. Hien, V. T. Bac, H. C. Tham, D. D. Nhan, L. D. Vinh. "Influence of meteorological conditions on $PM_{2.5}$ and $PM_{2.5–10}$ concentrations during the monsoon season in Hanoi, Vietnam". *Atmospheric Environment*, vol. 36, pp. 3473-3484, 2002.

[88] Chaloulakou,A., Kassomenos,P., Spyrelli,N., Demokritou, P.,Koutrakis,P., "Measurements of PM_{10} and $PM_{2:5}$ particle concentrations in Athens, Greece". *Atmospheric Environment*, vol. 37, pp. 649–660, 2003.

[89] M. Statheropoulos, N. Vassiliadis, A. Pappa. "Principal component and canonical correlation analysis for examining air pollution and meteorological data", *Atmospheric Environment*, vol. 32, pp. 1087-1095, 1998.

[90] *www.samasafia.dz\journaux.*

[91] Khedairia Soufiane, Khadir Mohamed Tarek, "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", in Proc of the 10th Maghrebian Conference on Information Technologies (MCSEAI'08), pp. 18-23, Oran, Alegria, April 2008.

[92] Khedairia Soufiane, Khadir Mohamed Tarek, "Self-Organizing Map and K-Means for Meteorological Day Type Identification for the Region of Annaba -Algeria-", In Proceedings of the IEEE Conference on computer information systems and industrial management applications (CISIM), Ostrava, The Czech Republic, 26 -28 June, 2008.

[93] Khedairia Soufiane, Khadir Mohamed Tarek, Cartes de Kohonen et K-Moyennes pour L'identification des Types de Jours Météorologiques de la Région d'Annaba (Algérie), Conférence Internationale Francophone d'Automatique, Buccharest, Roumanie, Sept. 2008.

[94] Khedairia Soufiane, Khadir Mohamed Tarek, 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.

[95] Andrew C. Comrie, Jeremy E. Diem, "Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona". *Atmospheric Environment*, vol. 33, pp. 5023-5036, 1999.

[96] H. A. Bridgman, T. D. Davies, T. Jickells, I. Hunova, K. Tovey, K. Bridges, V. Surapipith, "Air pollution in the Krusne Hory region, Czech Republic during the 1990s", *Atmospheric Environment*, vol. 36, pp. 3375-3389, 2002.

[97] F. Sezer Turalioğlu, Alper Nuhoğlu, Hanefi Bayraktar, Impacts of some meteorological parameters on SO2 and TSP concentrations in Erzurum, Turkey, *Chemosphere*, vol. 59, pp. 1633-1642, 2005.

[98] A.K. Gupta, Kakoli Karar, S. Ayoob, Kuruvilla John, "Spatio-temporal characteristics of gaseous and particulate pollutants in an urban region of Kolkata, India", *Atmospheric Research*, vol. 87, pp. 103-115, 2008.

[99] P. R. Hargreaves, A. Leidi, H. J. Grubb, M. T. Howe, M. A. Mugglestone, "Local and seasonal variations in atmospheric nitrogen dioxide levels at Rothamsted, UK, and relationships with meteorological conditions", *Atmospheric Environment*, vol. 34, pp. 843-853, 2000.

[100] 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.

[101] Collet, r. Oduyemi, k.: Air Quality Modelling: a Technical Review of Mathematical Approaches. *Meteorological Applications*, vol. 4, 1997.

[102] Hamdy K. Elminir. Hala Abdel-Galil, "Estimation of Air Pollutant Concentrations from Meteorological Parameters Using Artificial Neural Network", *Electrical Engineering*, vol. 57, No. 2, pp. 105–110, 2006.

[103] COMRIE, A.: Comparing Neural Networks and Regression Models for Ozone Forecasting, *Journal of Air Waste Manage*, vol. 47, 1997.

[104] Chow K. K. and Lim J. T., "Monitoring of suspended particulate In Petaling Jaya, In Yip Y. H. and Low K. S. (eds.): Urbanization and Ecodevelopment", University of Malaya, pp178-185, 1984.

[105] Ben Krose, Patrick van der Smagt, An introduction to neural networks, Nov, 1996.

[106] W.S. McCulloch, W.H. Pitts, “A logical calculus of the ideas immanent in nervous activity”, *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.
[107] Claude TOUZET, “les réseaux de neurones artificiels : introduction au connexionnisme”, 1992.
[108] Bernard Gosselin, “Application De Réseaux De Neurones Artificiels A La Reconnaissance Automatique De Caractères Manuscrits”, thèse de doctorat, Faculté Polytechnique de Mons, 1996.
[109] Marc Parizeaumm, “Réseaux de Neurones, Rapport technique”, Univ Laval, 2004.
[110] Ben Krose Patrick van der Smagt, “An Introduction to Neural Networks”, 1996.
[111] Mohamed Tarek Khadir, “Réseaux De Neurones Artificiels”, Uni. Badji Mokhtar Annaba, 2005.
[112] Mohamed Ryad Zemouri, “Contribution à la surveillance des systèmes de production à l’aide des réseaux de neurones dynamiques : Application à la e-maintenance”, thèse de doctorat, l’Université de Franche-Comté, 2003.
[113] Fouad Babran et sylvie THIRIA, Les perceptrons multicouches:de la régression non linéaire aux problèmes inverse, Rapport scientifique CEDRIC, 2001.
[114] Cybenko,G.,1989.Approximations by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, 303-314.
[115] Stéphane Breton, Une approche neuronale du contrôle robotique utilisant la vision binoculaire par reconstruction tridimensionnelle, thèse de doctorat, Université de Haute-Alsace, 1999.
[116] Sami Lallahem, Structure Et Modélisation Hydrodynamique des Eaux Souterraines : Application à l'aquifère Crayeux de la Bordure Nord du Bassin De Paris, thèse de doctorat, université des sciences et technologies de Lille, 2002.
[117] Mikko T.Kolehmainen, “Data Exploration with Self-Organizing Maps in Environmental Informatics and Bioinformatics”, doctoral thesis, Helsinki university of technology, 2004.
[118] Willmott, C.,Ackleson, S.,Davis, R., Fuddema, J., Klink, K., Legates, D.,O'Donnell, J., and Rowe, C., “Statistics for the evaluation and comparison of models”, *Journal of Geophysical Research*, vol. 90, pp. 8995-9005, 1985.
[119] Koffi Yao Blaise, Lasm Théophile, Ayral Piere Alain, Anne Johannet, 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*, Vol.17 No.3(2007), pp.313-328, 2007.
[120] J. Kukkonen, L. Partanen, A. Karppinen, J. Ruuskanen, H. Junninen, M. Kolehmainen, H. Niska, S. Dorling, T. Chatterton, R. Foxall, G. Cawley, “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*, vol. 37, pp. 4539–4550, 2003.

.