République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Amar Telidji- Laghouat

FACULTÉ: GENIE CIVIL ET D'ARCHITECTURE

DÉPARTEMENT : GENIE CIVIL

MÉMOIRE DE MASTER

Présenté par : GUENOU Tahar Dhiyaa Eddine

DOMAINE : Sciences et Techniques

FILIERE : Hydraulique

OPTION : Ressources Hydrauliques

Thème

Self-Organizing Map de Kohonen pour l'analyse spatiale de la qualité chimique des eaux souterraines du Sahara algérien.

Jury de soutenance :

Nom et Prénom	Grade	Qualité
Bouache Mohamed	MAA	Président
Tadj Walid	MCB	Examinateur
Chettih Mohamed	Pr.	Rapporteur

Promotion : Juin 2021


بسم الله الرحمن الرحيم

Remerciements

*Tout d'abord je remercie **ALLAH** le tout puissant de m'avoir donné la force, le courage et la patience de mener à bien ce modeste travail.*

*Ma profonde gratitude et mes remerciements les plus chaleureux vont particulièrement à mon encadreur **Mr. Chettih Mohamed** Professeur à l'Université de Laghouat, de m'avoir suivi et dirigé tout au long de la réalisation de ce travail. Aussi, je le remercie pour sa disponibilité permanente, pour son orientation efficace et pour ses idées originales qui ont servi à enrichir ce mémoire.*

*Je remercie aussi Monsieur **Bouache Mohamed,** pour m'avoir fait l'honneur d'accepter la présidence de jury*

*Je suis très honorée par la présence de Monsieur **Tadj Walid** qui a bien voulu examiner ce manuscrit et juger ce travail.*

Je remercie profondément tous les enseignants, les membres techniques et administratifs de département de Génie Civil.

Dédicace

Je dédie ce modeste travail à
mon père et à ma mère.

À mes chers frères et sœurs.

À toute ma famille.

Et

À mes amis.

T.Guenou

ملخص:

المياه الجوفية هي المصدر الوحيد لإمدادات المياه في الصحراء الجزائرية. وقد أدى تكثيف الاستغلال في السنوات الأخيرة إلى عدد من المشاكل، بما في ذلك التطور الهيدروكيمائي للمياه عن طريق الملوحة. في هذا العمل، تم وصف خوارزمية خرائط التنظيم الذاتي وتطبيقها لتصنيف المياه الجوفية من المنطقة القارية المشتركة باستخدام رمز حساب ثابت باستخدام عدة وظائف. وقد أتاحت النتائج التي تم الحصول عليها، التي طبقت على أكثر من 60 عينة من المياه، تصنيف هذه المياه إلى خمس فئات من خلال تسليط الضوء على مصدرها. كما أتاحت الطريقة التي تفسر الكيمياء الخاصة بها. تسليط الضوء على الجمعيات الجيوكيميائية

الكلمات المفتاحية : خريطة ذاتية التنظيم ، التصنيف ، الكيمياء المائية ، المنطقة القارية المشتركة

Groundwater is the only source of water supply in the Algerian Sahara. The intensification of exploitation has given rise in recent years to a number of problems, including the hydrochemical evolution of water by salinization. In this work, the Self-Organizing Maps algorithm was described and applied to classify groundwater from the Continental Interlayer using an established calculation code using several functions. Applied to more than 60 water samples, the results obtained made it possible to classify these waters into five classes by highlighting their origin. The SOM method also made it possible to highlight the geochemical associations that explain their chemism.

Keywords: Self-Organizing Map, Classification, Hydrochemistry, Continental Intercalary.

Les eaux souterraines sont l'unique source d'approvisionnement en eau dans le Sahara algérien. L'intensification de l'exploitation a engendré ces dernières années un certain nombre de problèmes dont l'évolution hydrochimique des eaux par salinisation. Dans ce travail, l'Algorithme Self-Organizing Maps a été décrit et appliqué pour classer les eaux souterraines du Continental Intercalaire à l'aide d'un code de calcul établi faisant appel à plusieurs fonctions. Appliqué à plus de 60 échantillons d'eau, les résultats obtenus, ont permis de classer ces eaux en cinq classes en mettant en évidence leur origine. La méthode SOM a permis également de mettre en évidence les associations géochimiques qui expliquent leur chimisme.

Mots clés : Self-Organizing Map, Classification, Hydrochimie, Continental Intercalaire.

Table des Matières

Chapitre 2: Organigrammes et Programmes de Calcul

Chapitre 3 : Classification des eaux souterraines du Continental Intercalaire par la méthode SOM

Liste des Figures

Liste des tableaux

Introduction Générale

Le travail développé dans ce mémoire s'inscrit dans le cadre d'une étude pour la classification non supervisée de données hydrochimiques par des méthodes géométriques bio-inspirées, plus spécifiquement par les cartes auto-organisatrices de Kohonen.

Grâce aux progrès réalisés en technologies de l'information, il est désormais possible de recueillir, stocker et gérer de grandes masses de données. Par conséquent, il est difficile, quand le volume des informations est important, de décrire d'une façon succincte les différentes caractéristiques associées aux données, d'où la nécessite d'automatiser l'analyse de ces données. De nombreuses méthodes ont été développées qui sont issues du domaine de l'intelligence artificielle.

L'analyse de données par classification est un domaine très utilisé dans l'exploration des données. Le but de la classification automatique est de découper l'ensemble des données étudiées en un ou plusieurs sous-ensembles nommés classes, chaque sous-ensemble devant être le plus homogène possible. Les membres d'une classe ressemblent plus aux autres membres de la même classe qu'aux membres d'une autre classe. On distingue la classification supervisée et non supervisée.

La carte auto-organisée de Kohonen est l'une des techniques les plus performantes des méthodes non supervisées. Créée en 1982 par le Professeur Finlandais Teuvo Kohonen, une carte auto-organisée est un type de réseau neuronal artificiel dont l'apprentissage se déroule de manière non supervisée.

Avec près d'un million de km^2 de superficie, le Sahara septentrional, qui s'étend d'Ouest en Est entre l'Algérie, la Tunisie et la Libye, est l'un des plus grands déserts au monde. À cause de son climat aride, l'alimentation en eau dépend essentiellement des eaux souterraines où la qualité de l'eau est très importante.

Cependant, depuis plusieurs années, l'exploitation par forages des aquifers du Sahara a sévèrement entamé cette réserve d'eau souterraine. Les prélèvements, utilisés autant pour des fins agricoles que pour l'alimentation en eau potable et pour l'industrie, ont passés à 2,5 milliards de m^3/an à travers des forages d'eau dont le nombre a atteint aujourd'hui plus de 5000 forages où les foggaras et les sources qui tarissent, sont remplacées par des forages de plus en plus profonds. Cette intensification de l'exploitation a engendré un certain nombre de problèmes dont principalement l'évolution hydrochimique des eaux par salinisation.

Ce projet propose de caractériser l'évolution de la nappe du Continental Intercalaire du Sahara d'un point de vue hydrochimique afin de comprendre mieux son fonctionnement et son évolution. A cette fin, on propose une nouvelle approche pour la classification des eaux basée sur un type de réseau de neurone particulier, le réseau de Kohonen.

La mémoire est organisée en 3 chapitres :

- Dans le premier chapitre, nous exposerons les différentes méthodes de classification non supervisée qui s'appliquent aux données numériques univaluées. Cependant, la plus grande partie de ce chapitre sera dédiée aux cartes-organisatrices de Kohonen dans leur cadre classique de la classification de données.

- Dans le deuxième chapitre, nous appliquerons la méthode SOM sur des échantillons pour tester l'efficacité du code de calcul sans pour autant détailler l'interprétation des résultats. On tentera d'essayer de maitriser le code et obtenir plus de résultats afin d'explorer la méthode.

- Dans le troisième chapitre, nous appliquerons méthode SOM pour classifier la variabilité spatiale des paramètres physico-chimiques afin de comprendre le fonctionnement de la nappe et mettre en évidence l'origine de l'eau et les associations des éléments chimiques.

Chapitre 1

Cartes auto-organisatrices de Kohonen

I.1. Introduction

Durant ces trois dernières décennies, de nombreuses méthodes d'analyse de données se sont développées en parallèle avec les progrès technologiques et informatiques. L'analyse de données par classification est un domaine très utilisé dans l'exploration des données. Parmi ces méthodes, les méthodes de classification non supervisées basées sur l'intelligence artificielle et qui consiste à répartir les données en groupes homogènes ou classes, dans un but informel ou décisif.

La carte auto-organisée de Kohonen est l'une des techniques les plus performantes des méthodes non supervisées. Créée en 1982 par le Professeur Finlandais Teuvo Kohonen, une carte auto-organisée est un type de réseau neuronal artificiel dont l'apprentissage se déroule de manière non supervisée. La caractéristique remarquable de cet Algorithme est que les vecteurs d'entrée qui sont proches dans un espace de haute dimension sont également mappés aux nœuds voisins dans un espace 2D. Il s'agit essentiellement d'un procédé de réduction de dimensionnalité, car il relie les entrées de dimension élevée à une représentation discrétisée de faible dimension et conserve la structure sous-jacente de son espace d'entrée.

Les cartes auto-organisées diffèrent des autres réseaux de neurones artificiels car elles appliquent l'apprentissage compétitif par opposition à l'apprentissage de correction d'erreurs comme la rétropropagation avec descente de gradient, et en ce sens, elles utilisent une fonction de voisinage pour préserver les propriétés topologiques de l'espace d'entrée.

Ainsi, dans ce chapitre, nous exposerons les différentes méthodes de classification non supervisée qui s'appliquent aux données numériques univaluées. Cependant, la plus grande partie de ce chapitre sera dédiée aux cartes-organisatrices de Kohonen dans leur cadre classique de la classification de données.

I.2. Méthodes de classification

Les méthodes de classification des données permettent de grouper des objets, des observations ou des individus dans des classes de manière à ce que les objets appartenant à la même classe sont plus similaires entre eux qu'aux objets appartenant aux autres classes.

Dans le domaine de la classification automatique du Machine Learning, il existe deux principaux types de tâches : **supervisées** et **non supervisées**. La classification par la méthode non supervisée est définie comme étant un apprentissage automatique où les données ne sont pas étiquetées (Fig. 1).

Fig. 1 : Apprentissage supervisé et non supervisé

I.2.1. Aperçu sur les méthodes de classification supervisées

L'**apprentissage supervisé** (*Supervised Learning*) est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés. On distingue les problèmes de régression des problèmes de classification. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification (Cornuéjols et al.,2002). Les connaissances a priori sont utilisées pour la création des classes et la saisie des échantillons (Cherel, 2010).

Il existe de nombreuses méthodes de classification supervisée (Jacques, 2018), parmi lesquelles on peut citer : la méthode des k plus proches voisins, l'analyse factorielle discriminante, la régression logistique, les arbres de décision, les forêts aléatoires, les réseaux de neurones, la méthode SVM (Support Vector Machine), la méthode PLS-DA, ainsi que d'autres méthodes.

Le schéma suivant (Fig. 2) résume les principales méthodes supervisées.

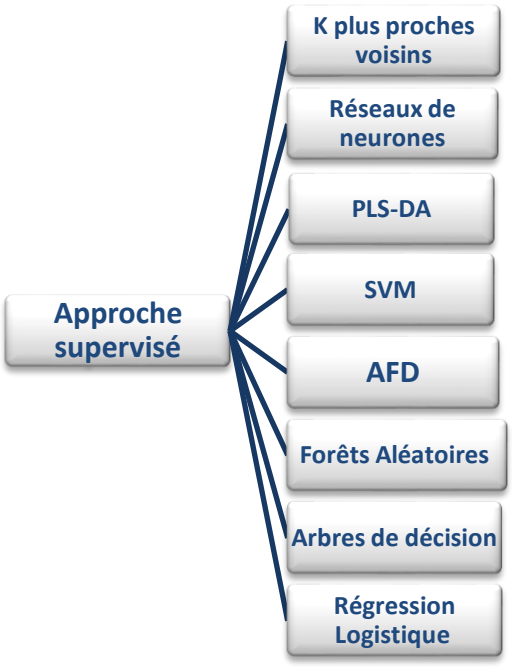


Fig. 2 : Différents types d'approches supervisées

I.2.1.1. Méthode des k plus proches voisins :

En intelligence artificielle, et plus précisément en apprentissage automatique, la **méthode des *k* plus proches voisins** est une méthode d'apprentissage supervisé, en abrégé, la méthode est noté k-NN ou KNN, de l'anglais *k-Nearest Neighbors*. Dans le domaine de la reconnaissance de forme, l'algorithme des ***k* plus proches voisins** (***k*-NN**) est très utilisé et représente une méthode non paramétrique utilisée pour la classification et la régression.

Dans ce cadre, on dispose d'une base de données d'apprentissage constituée de *N* couples « entrée-sortie ». Pour estimer la sortie associée à une nouvelle entrée *x*, la méthode des *k* plus proches voisins consiste à prendre en compte (de façon identique) les *k* échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée *x*, selon une distance à définir. Puisque cet algorithme est basé sur la distance, la normalisation peut améliorer sa précision (Madeh Piryonesi et al., 2020) (Hastie et al., 2001).

I.2.1.2. Analyse Factorielle Discriminante :

L'**Analyse Factorielle Discriminante** (**AFD**) ou simplement **Analyse Discriminante** est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire…) d'un ensemble d'observations à partir d'une série de variables prédictives (descripteurs, variables exogènes…) (Saporta, 2006).

I.2.1.3. Régression Logistique :

La régression logistique est un modèle statistique qui, dans sa forme de base, utilise une fonction logistique pour modéliser une variable dépendante binaire, bien qu'il existe de nombreuses extensions plus complexes. Dans l'analyse de régression, la régression logistique (ou régression logit) est l'estimation des paramètres d'un modèle logistique (Tolles et al.,2016).

I.2.1.4. Arbres de décision :

Un **arbre de décision** est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, la médecine, etc… Il a l'avantage d'être lisible et rapide à exécuter. Il s'agit de plus d'une représentation calculable automatiquement par des algorithmes d'apprentissage supervisé (cnam, 2016).

I.2.1.5. Forêts Aléatoires :

Deux des algorithmes proposés par Leo Breiman : les arbres CART (pour Classification And Regression Trees) introduits dans la première moitié des années 80 et les forêts aléatoires apparues, quant à elles, au début des années 2000. Les **forêts aléatoires** sont des méthodes qui permettent d'obtenir des modèles prédictifs pour la classification et la régression.

L'idée générale derrière la méthode est la suivante : au lieu d'essayer d'obtenir une méthode optimisée en une fois, on génère plusieurs prédicteurs avant de mettre en commun leurs différentes prédictions (Breiman, 2001).

I.2.1.6. Méthode SVM (Support Vector Machine)

Les **machines à vecteurs de support** ou **séparateurs à vaste marge** (en anglais Support Vector Machine, SVM) sont des modèles d'apprentissage supervisés avec des algorithmes d'apprentissage associés qui permettent de résoudre des problèmes tant de classification que de régression ou de détection d'anomalie. La méthode SVM a été développée dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la théorie de Vapnik-Chervonenkis. Les Machines à Vecteurs de Support ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyperparamètres, leurs garanties théoriques, leurs bons résultats en pratique, leur grande flexibilité ainsi que leur simplicité d'utilisation même sans grande connaissance de data mining (Boser et al., 1992 ; DAP, 2020).

I.2.1.7. Méthode PLS-DA:

L'Analyse Discriminante des Moindres Carrés Partiels (**PLS-DA** pour : *Partial Least Squares-Discriminant Analysis*) est une méthode statistique qui porte un certain rapport avec la régression des composants principaux ; au lieu de trouver des hyperplans de variance maximale entre la réponse et les variables indépendantes, elle trouve un modèle de régression linéaire en projetant les variables prévues et les variables observables à un nouvel espace. La régression

PLS est une technique de régression qui vise à prédire les valeurs prises par un groupe de variables Y (variables à prédire, variables cibles, variables expliquées) à partir d'une série de variables X (variables prédictives, les descripteurs, variables explicatives) (Tenenhaus, 1998).

Étant donné que les données X et Y sont projetées dans de nouveaux espaces, la famille de méthodes PLS est connue sous le nom de modèles à facteurs bilinéaires. L'analyse discriminante des moindres carrés partiels (PLS-DA) est une variante utilisée lorsque Y est catégorique (Eriksson, 2001).

I.2.2. Méthodes de classification non supervisées

La classification par les méthodes **non supervisées** est définie comme étant l'apprentissage automatique où les données ne sont pas étiquetées. Il s'agit donc de découvrir les structures sous-jacentes à ces données. En général, les systèmes d'apprentissage non supervisé permettent d'exécuter des tâches plus complexes que les systèmes d'apprentissage supervisé.

L'apprentissage non supervisé (Unsupervised Learning) consiste à ne disposer que de données d'entrée (X) et pas de variables de sortie correspondantes (Ismaili analyticsinsights.io, 2021). A ce titre, il existe de nombreuses méthodes de classification non supervisée dont les plus importantes sont les méthodes géométriques et les méthodes probabilistes, ainsi que d'autres méthodes de moindre importance (Saad Hajjar, 2014).

L'organigramme ci-dessous (Fig. 3) résume les principales méthodes de l'approche non supervisée.

I.2.2.1. Méthodes Géométriques

Ces méthodes tentent de décomposer le problème à résoudre en petits problèmes classiques de géométrie et de combiner les solutions de ces problèmes pour avoir la solution globale. La résolution se fait par des méthodes déductives ou des algorithmes de parcours de graphes (Ait Aoudia, 1994).

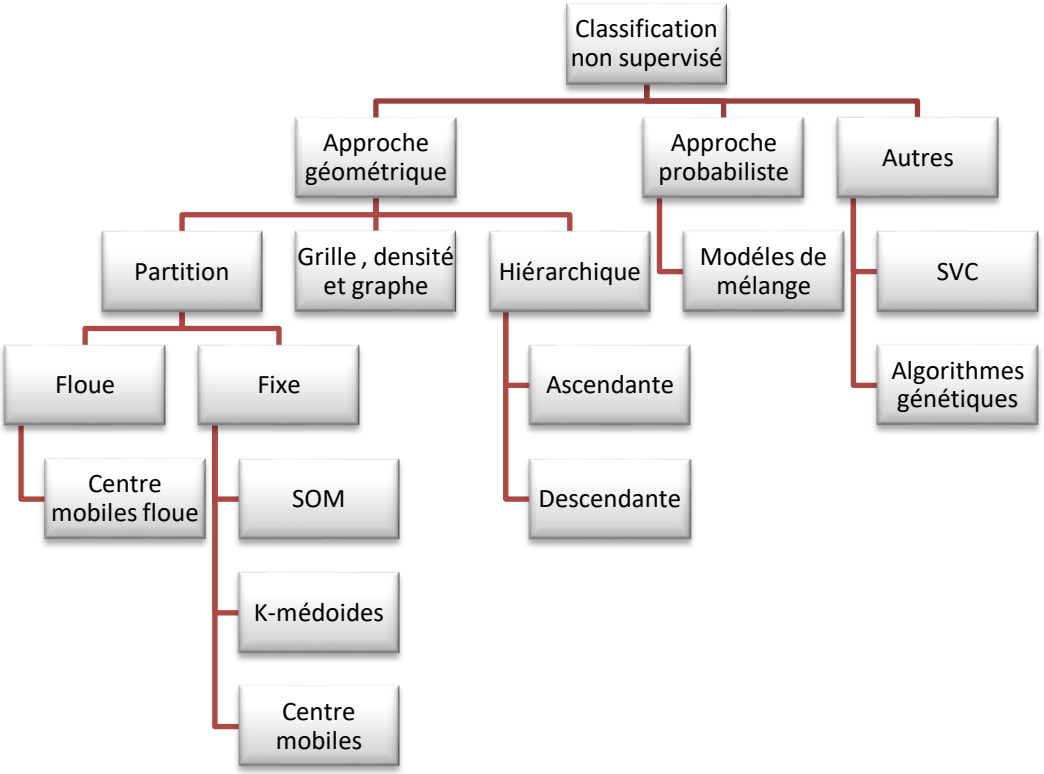


Fig. 3 : Différentes méthodes de l'approche non supervisée

La classification de données se fonde sur une mesure de proximité entre les individus. Parmi ces méthodes, on trouve :

- Classification hiérarchique ascendante et descendante ;
- Classification par partitionnement fixe et flou ;
- Méthodes basées sur la densité, sur les graphes et sur les grilles.

a) - Classification hiérarchique :

Dans le domaine de l'analyse et de la classification automatique de donnée la notion de regroupement hiérarchique recouvre différentes méthodes de partitionnement de données, et se catégorise en deux grandes familles : les méthodes '*ascendantes*' et les méthodes '*descendantes*'.

- **Classification hiérarchique ascendante** : La classification ascendante hiérarchique est une méthode de classification itérative dont le principe est simple. Elle s'agit de regrouper successivement les objets en forme d'un arbre binaire de classification appelé : dendrogramme. (larmarange.github.io, 2019).

- **Classification hiérarchique descendante** : Elle contraire à la classification hiérarchique ascendante, utilisée lorsque l'ordre est descendant.

b) - Classification à base de Grille, Densité et Graphe :

Les classifications à base de grilles ont un principe similaire aux méthodes à base de densité. Elles consistent à diviser chaque dimension de l'espace des observations en intervalles de longueur égale formant ainsi une grille composée d'unités qui ne se chevauchent pas. (Wang et al., 1997).

Les classifications à base de densité sont considérées comme des zones de haute densité, séparées par des zones de basse densité formées par des objets appelés bruits ou objets frontaliers (Saad Hajjar, 2014).

Les méthodes à base de graphes se fondent sur la théorie des graphes pour classifier les données (Augustson et Minker, 1970).

c) - Classification par partitionnement :

Le partitionnement de données est une méthode de classification non supervisée où la structure classificatoire recherchée est la partition (Berkhin, 2002). En définissant une fonction d'homogénéité ou un critère de qualité sur une partition le problème de classification devient un problème parfaitement défini en optimisation discrète (Partition (lri.fr)).

On distingue deux catégories de partition : fixe et floue. Dans une partition floue, les classes se chevauchent et une observation peut appartenir à plus d'une classe, tout en quantifiant cette appartenance par un certain degré qui varie entre 0 et 1. Par contre, avec une partition fixe, chaque observation appartient à une et une seule classe.

Dans la partition fixe on distingue la méthode k-médoïdes, la méthode SOM (Self-Organizing Maps) et la méthode centre mobile.

- **Méthode K-médoïdes :**

Un médoïde est un représentant d'une classe choisit comme étant l'objet le plus central de la classe, il décompose le flux de données sous la forme d'un ensemble de lots de données. Chaque lot est partitionné et résumé sous la forme de médoïdes pondérés (Labroche, 2012). Cet algorithme est plus robuste vis-à-vis des données aberrantes que celui des k-means.

- **Méthode SOM :**

Les cartes auto-adaptatives (SOM) est une classe de réseau de neurones artificiels fondée sur des méthodes d'apprentissage non-supervisées. Ces réseaux sont utilisés pour cartographier un espace réel, c'est-à-dire pour étudier la répartition de données dans un espace à grande dimension (Ritter et al., 1992). Cette méthode fait l'objet de ce travail, elle sera plus décrite et détaillée plus loin.

- **Méthode des centres mobiles :**

La méthode des centres mobiles (Forgy, 1965), appelé aussi K-means, conduit à une partition fixe de l'ensemble de données en optimisant une fonction coût, et chaque individu sera affecté à la classe dont le centre de gravité en est le plus proche. Elle permet de partitionner en différentes classes des individus pour lesquels on dispose de mesures. La méthode des centres

mobiles s'applique lorsque l'on sait à l'avance combien de classes on veut obtenir (Méthode des K-means (univ-lyon2.fr)).

I.2.2.2. Méthodes Probabilistes :

La méthode probabiliste est une méthode non constructive, initialement utilisée en analyse combinatoire et popularisée par Paul Erdős pour démontrer l'existence d'un type donné d'objet mathématique. Cette méthode a été appliquée à d'autres domaines des mathématiques tels que la théorie des nombres, l'algèbre linéaire et l'analyse réelle. Son principe est de montrer que si l'on prend au hasard des objets d'une catégorie, la probabilité que le résultat soit d'un certain type est plus que zéro. Bien que la démonstration utilise la théorie des probabilités, la conclusion finale est déterminée de façon certaine. Parmi les Méthodes à approche probabiliste un outil populaire aujourd'hui dans la classification réside dans les modèles de mélanges.

- **Modèle de mélange :**

Les modèles de mélanges, apparus dans les travaux de Pearson en 1984, sont utilisés avec succès dans bon nombre de disciplines comme l'astronomie, la biologie, la génétique, l'économie, les sciences de l'ingénieur, le marketing, la reconnaissance d'images, etc, ….

Le modèle de mélange consiste à supposer que les données proviennent d'une source contenant plusieurs sous-populations homogènes appelées composants. La population totale est un mélange de ces sous-populations. Le modèle résultant est un modèle de mélange fini (Roche et Djrobie, 2016).

I.2.2.3. Autres méthodes :

D'autres méthodes de classification non supervisée existent aussi comme les méthodes à base d'algorithmes génétiques (Raghavan et Birchard, 1979), et la méthode à base de machines à vecteurs supports non supervisée connue sous le nom de SVC (Support Vector Clustering) qui consiste à projeter les données dans un espace à dimension supérieure, appelé espace de redescription, en utilisant une fonction noyau (Ben-Hur et al., 2001).

a) - Algorithmes génétiques

Les algorithmes génétiques appartiennent à la famille des algorithmes évolutionnistes. Leur but est d'obtenir une solution approchée à un problème d'optimisation, lorsqu'il n'existe pas de méthode exacte (ou que la solution est inconnue) pour le résoudre en un temps raisonnable. Les algorithmes génétiques utilisent la notion de sélection naturelle et l'appliquent à une population de solutions potentielles au problème donné. La solution est approchée par « bonds » successifs, comme dans une procédure de séparation et évaluation, à ceci près que ce sont des formules qui sont recherchées et non plus directement des valeurs.

b) - Méthode SVC (Support Vector Clustering)

L'algorithme Support Vector Clustering, créé par Hava Siegelmann et Vladimir Vapnik, applique les statistiques des vecteurs de support, développées dans l'algorithme de SVM pour

catégoriser les données non étiquetées, il est l'un des algorithmes de clustering les plus largement utilisés dans les applications industrielles.

Dans cette méthode, les points de données sont cartographiés de l'espace de données à un espace de fonctionnalité dimensionnel élevé appelé espace de redescription (Saad Hajjar, 2014) à l'aide d'une fonction noyau (Ben-Hur et al., 2001).

I.2.3. Cartes auto-organisatrices de Kohonen

Les cartes auto-organisatrices de Kohonen (SOM pour Self Organizing Maps) désignent un ensemble de méthodes d'apprentissage non supervisé (Kohonen, 2001). Il s'agit d'un cas particulier des réseaux de neurones artificiels où l'apprentissage alterne des phases de compétition et de coopération entre les neurones.

La méthode permet une réduction des dimensions des données : les données de départ sont projetées sur un treillis (ou grille) de neurones le plus souvent à deux dimensions. De ce point de vue, les SOM s'apparentent aux méthodes factorielles (ACP, AFC, AFD, . . .) où les points sont projetés sur des axes minimisant la variance du nuage de points. Néanmoins, les dimensions du treillis ne sont pas liées à celle des données. Plus généralement, il s'agit de méthodes dont l'inspiration et le raisonnement sont très différents.

La projection sur une grille permet de prendre en compte le voisinage des points afin de préserver la topologie des données. Autrement dit, si deux objets sont similaires dans l'espace original, alors leur position sur la grille devrait être elle aussi similaire (Soubiran, 2016).

I.2.3.1. Réseaux de Neurones Artificiels

Les Réseaux de Neurones sont un moyen populaire de mettre en œuvre l'intelligence artificielle. L'idée est qu'ils se comportent comme les neurones d'un cerveau. Le terme de « réseau de neurones » suggère un lien fort avec la biologie. Ce lien existe : les méthodes mathématiques qui seront décrites par la suite, ont été appliquées avec succès à la modélisation des systèmes nerveux vivants (Fig. 4). Néanmoins, le terme est plus métaphorique que scientifique : si le lien avec la biologie a constitué une motivation majeure des pionniers du domaine, les réels développements des réseaux de neurones sont de nature purement mathématique et statistique ; leurs applications se situent dans des domaines qui n'ont généralement aucun rapport avec la neurobiologie (Dreyfus, 2002).

En terme mathématique un neurone est une fonction algébrique non linéaire et paramétré, on distingue deux types de réseaux : bouclés et non bouclés.

– Historique

Walter Pitts, un logicien, et Warren McCulloch, un neuroscientifique, ont créé le premier modèle mathématique d'un réseau neuronal en 1943. Publiés dans leur ouvrage fondateur : *A logical calculus of the ideas immanent in nervous activity*, ils ont proposé une combinaison de mathématiques et d'algorithmes visant à imiter les processus de la pensée humaine.

En 1949, la règle de Hebb ou théorie des assemblées de neurones a été établie par Donald Hebb. Elle est à la fois utilisée comme hypothèse en neurosciences et comme concept dans les réseaux neuronaux en mathématiques.

Frank Rosenblatt, un psychologue, a soumis un document intitulé : *Le perceptron, un automate percevant et reconnaissant*, au Cornell Aeronautical Laboratory en 1957. Il a déclaré qu'il construirait un système électronique ou électromécanique qui apprendrait à reconnaître les similitudes ou les identités entre les modèles d'informations optiques, électriques ou tonales, d'une manière qui peut être étroitement analogue aux processus perceptifs d'un cerveau biologique (Rosenblatt, 1957).

Le modèle connexionniste du célèbre perceptron de Rosenblatt (1962), fut sans doute le premier modèle de catégorisation perceptive à base de réseau neuromimétique doté d'une capacité d'apprentissage (Victorri, 2006). Mais très vite, la voie de recherche ainsi ouverte a été abandonnée au profit du calcul symbolique prôné par les promoteurs de l'intelligence artificielle classique, en particulier à la suite de critiques sévères de Minsky et Papert (1969), qui ont mis en évidence les limites, jugées à l'époque indépassables, des performances du perceptron. Cela va avoir une grande incidence sur la recherche dans ce domaine. Cependant, en 1972 Kohonen présente ses travaux sur les mémoires associatives et propose des applications à la reconnaissance de formes. En 1982 Hopfield présente son étude d'un réseau complétement rebouclé, dont il analyse la dynamique. Mais, il a fallu attendre une vingtaine d'années après les critiques de Minsky et Papert sur les propriétés du perceptron avant que le connexionnisme ne revienne sur les devants de la scène, avec la publication du livre du groupe de recherche PDP (Parallel Distributed Processing), édité par McCleland et Rumelhart (1986), qui a donné une formidable impulsion aux recherches dans ce domaine. La découverte quasi simultanée par plusieurs équipes de chercheurs (Le Cun 1986, Rumelhart et al. 1986, Parker 1985) d'un nouvel algorithme d'apprentissage, la méthode de rétropropagation du gradient de l'erreur, a montré que l'on pouvait dépasser largement les limites qui avaient handicapé le perceptron.

Aujourd'hui, les réseaux neuronaux sont utilisés dans de nombreux domaines (entre autres, vie artificielle et intelligence artificielle) à cause de leur propriété en particulier, leur capacité d'apprentissage.

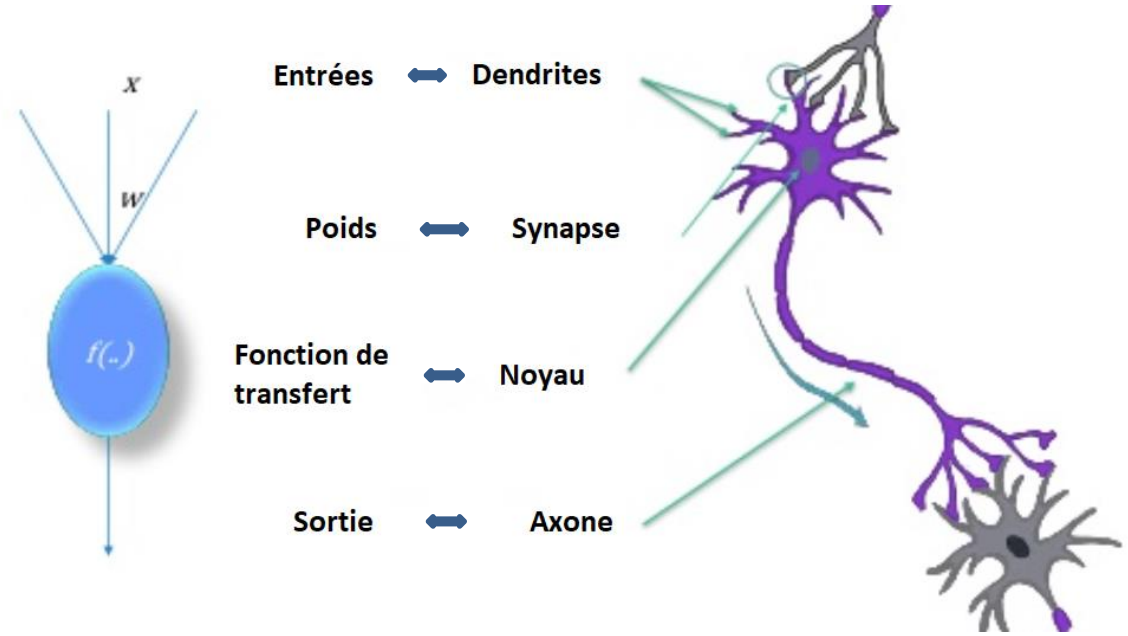


Fig. 4 : Neurone Biologique et Neurone Formel

I.2.3.2. Self-Organizing Map (SOM)

Les cartes auto-organisatrices ont été introduites pour la première fois par Kohonen qui cherchait à représenter des données multidimensionnelles et de grande taille. Pour y parvenir, Kohonen cherche à partitionner, par apprentissage, les données en groupements similaires dont la structure de voisinage peut être matérialisée et visualisable par un espace discret de faible dimension (1, 2 ou 3D) appelé carte topologique ou auto-organisatrice.

Elles sont utilisées pour cartographier un espace réel, c'est-à-dire pour étudier la répartition de données dans un espace à grande dimension. En pratique, cette cartographie peut servir à réaliser des tâches de discrétisation, quantification vectorielle ou classification.

Techniquement, la carte réalise une quantification vectorielle de l'espace de données. Cela signifie discrétiser l'espace ; c'est-à-dire le diviser en zones, et affecter à chaque zone un point significatif dit vecteur référent.

a) - Composition et principe :

Une carte auto-organisatrice est composée d'une grille de neurones de faible dimension. Quand la grille est unidimensionnelle (Fig. 5), chaque neurone a deux voisins (Saad Hajjar, 2014).

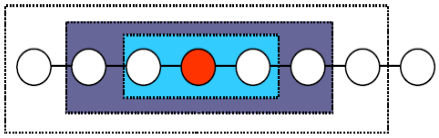


Fig. 5 : - Carte auto-organisatrice Unidimensionnelle de 8 neurones.

Quand la grille est bidimensionnelle (Fig. 6), l'arrangement des neurones se fait d'une façon rectangulaire où chaque neurone possède 4 voisins (topologie rectangulaire) ou d'une façon hexagonale où chaque neurone possède 6 voisins (topologie hexagonale).

Fig. 6 : - Cartes auto-organisatrices bidimensionnelles de 25 neurones.
(a) – Voisinage rectangulaire
(b) – Voisinage hexagonal

Les neurones sont reconnus par leur numéro et leur emplacement sur la grille. Les données sont projetées de leur espace initial, ou espace d'entrée, vers la carte ou espace de sortie. A chaque neurone de la carte est associé un vecteur référent, appelé aussi vecteur prototype ou prototype, appartenant à l'espace d'entrée.

En désignant par K le nombre total des neurones de la carte, le vecteur référent du neurone k est reconnu par w_k avec $k \in \{1, \ldots\ldots, K\}$ et $w_k \in \mathbb{R}^p$. L'objectif de l'apprentissage de la carte consiste à mettre à jour les vecteurs référents de façon à approximer au mieux la distribution des vecteurs d'entrée tout en reproduisant l'auto organisation des neurones de la carte. L'apprentissage de la carte se fait en mode séquentiel appelé aussi incrémental, ou en mode différé (batch).

b) – Apprentissage Séquentiel :

Après une initialisation aléatoire des valeurs de chaque neurone, on soumet une à une les données à la carte auto adaptative. Selon les valeurs des neurones, il y en a un, appelé neurone gagnant ou vainqueur (BMU pour Best Matching Unit), qui répond le mieux au stimulus ; c'est celui dont la valeur est la plus proche de la donnée présentée. Ce neurone est alors gratifié d'un changement de valeur pour qu'il réponde encore mieux à un autre *stimulus* de même nature que le précédent. Par là-même, on gratifie un peu aussi les neurones voisins du gagnant avec un facteur multiplicatif du gain inférieur à un. Ainsi, c'est toute la région de la carte autour du neurone gagnant qui se spécialise. En fin d'algorithme, lorsque les neurones ne bougent plus, ou seulement très peu, à chaque itération, la carte auto-organisatrice recouvre toute la topologie des données. Chaque itération t de l'apprentissage séquentiel comprend deux étapes. La première étape consiste à choisir au hasard une observation $x(t)$ de l'ensemble, et à la présenter au réseau dans le but de déterminer son neurone vainqueur. Le neurone vainqueur BMU, d'une observation est le neurone dont le vecteur référent en est le plus proche au sens d'une distance donnée (ex : distance euclidienne). Si c est le neurone vainqueur du vecteur $x(t)$, c est déterminé comme suit :

$$d(w_c(t), x(t)) = \min_{k \in \{1,\dots,K\}} d(w_c(t), x(t)) \qquad (1)$$

Dans la deuxième étape, le neurone vainqueur est activé. Son vecteur référent est mis à jour pour se rapprocher du vecteur d'entrée présenté au réseau.

La sélection de BMU pour chaque élément d'entrée, consiste à trouver le nœud de carte le plus proche, par mesure de distance euclidienne (Kohonen, 1998). Le processus de mise à jour tient compte de l'ensemble des données d'entrée à la fois, sur une série d'itérations (Kohonen, 1993 ; Clark, 2018).

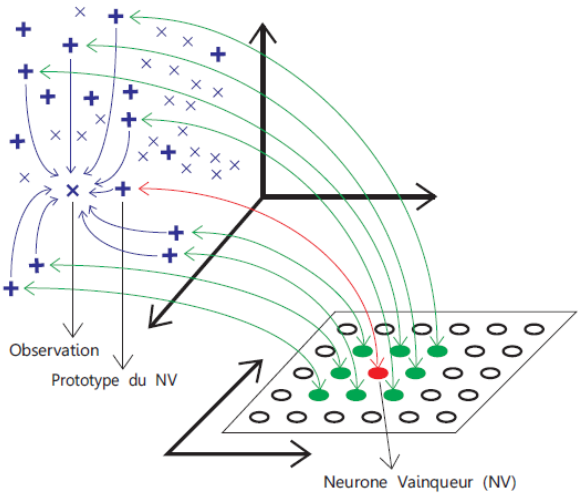


Fig. 7 : - Mise à jour des vecteurs prototypes
(D'après Saad Hajjar, 2014)

Cette mise à jour ne concerne pas seulement le neurone vainqueur comme dans les méthodes de l'apprentissage par compétition (Winner take all), mais aussi les neurones qui lui sont voisins et qui voient alors leurs vecteurs référents s'ajuster vers ce vecteur d'entrée (Fig. 7).

L'amplitude de cet ajustement est déterminée par la valeur d'un pas d'apprentissage $\alpha(t)$ et la valeur d'une fonction de voisinage $h(t)$. Le paramètre $\alpha(t)$ règle la vitesse de l'apprentissage. Il est initialisé avec une grande valeur au début puis décroît avec les itérations en vue de ralentir au fur et à mesure le processus d'apprentissage. La fonction $h(t)$ définit l'appartenance au voisinage.

La fonction $h(t)$ dépend à la fois de l'emplacement des neurones sur la carte et d'un certain rayon de voisinage. Dans les premières itérations, le rayon de voisinage est assez large pour mettre à jour un grand nombre de neurones voisins du neurone vainqueur, mais ce rayon se rétrécit progressivement pour ne contenir que le neurone vainqueur avec ses voisins immédiats, ou bien même le neurone vainqueur seulement. La règle de mise à jour des vecteurs référents est la suivante (Kohonen, 1984 ; Ritter et Schulten, 1986 ; Saad Hajjar, 2014) :

$$w_k(t+1) = w_k(t) + \alpha(t) h_{ck}(t)[x(t) - w_k(t)] \qquad k \in \{1, \dots, K\} \qquad (2)$$

où c est le neurone vainqueur du vecteur d'entrée $x(t)$ présenté au réseau à l'itération t et $h_{ck}(t)$ est la fonction de voisinage qui définit la proximité entre les neurones c et k.

Une fonction de voisinage entre le neurone vainqueur c et un neurone k de la carte vaut 1 si le neurone k se trouve à l'intérieur du carré centré sur le neurone c et 0 dans les autres cas.

Le rayon de ce carré est appelé rayon de voisinage. Il est large au début, puis se rétrécit avec les itérations pour contenir seulement le neurone c avec ses voisins immédiats à la fin de l'apprentissage ou même seulement le neurone c (Saad Hajjar, 2014). Sur la figure 8, si le rayon de voisinage est 1, la fonction de voisinage vaut 1 pour les neurones contenus à l'intérieur du carré interne (vert) et 0 pour les autres neurones de la carte.

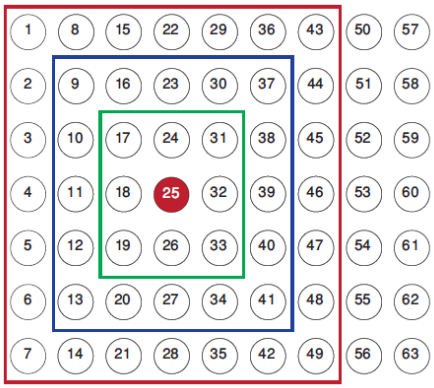


Fig. 8 : - Carte auto-organisatrice rectangulaire de 63 neurones.
(D'après Saad Hajjar, 2014)

Une fonction de voisinage plus flexible et plus commune est la fonction gaussienne illustrée dans la figure 9 et définie comme suit :

$$h_{ck}\big(\sigma(t)\big) = exp\left(-\frac{d_2^2(r_c, r_k}{2\sigma^2(t)}\right)$$

$$= exp\left(-\frac{\|r_c - r_k\|^2}{2\sigma^2(t)}\right) \qquad (3)$$

où r_c et r_k sont respectivement l'emplacement du neurone c et du neurone k sur la carte, et $\sigma(t)$ est le rayon du voisinage à l'itération t du processus d'apprentissage. Avec une telle fonction de voisinage, l'amplitude de l'ajustement est graduée selon l'éloignement du neurone vainqueur qui réserve à lui-même l'amplitude maximale.

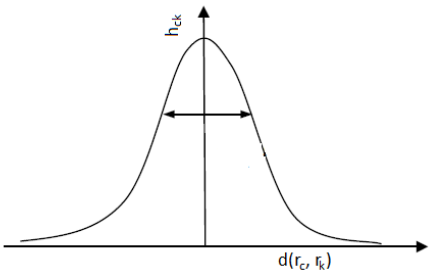


Fig. 9 : - Fonction de voisinage Gaussienne.

c) – Apprentissage en mode différé

En mode différé, à chaque itération t, toutes les observations sont présentées au réseau et la mise à jour des vecteurs prototypes se fait en prenant en compte toutes les observations de l'ensemble de données.

Chaque vecteur prototype est une moyenne pondérée des vecteurs d'observations $(x_i, i \in \{1, \dots, n\})$ quand le carré de la distance euclidienne est utilisé pour le calcul du neurone vainqueur, les poids correspondants étant les valeurs de la fonction de voisinage $h(t)$ (Kohonen, 2001). La règle de mise à jour des vecteurs prototypes est donnée par :

$$w_k(t+1) = \frac{\sum_{i=1}^{n} h_{kc_i}(t) x_i}{\sum_{i=1}^{n} h_{kc_i}(t)} \quad (4)$$

où h_{kc_i} est la valeur de la fonction de voisinage entre le neurone vainqueur c_i du vecteur x_i et le neurone k.

La mise à jour des vecteurs prototypes peut être formulée autrement en utilisant le fait que les observations qui ont le même neurone vainqueur ont la même valeur pour la fonction de voisinage et appartiennent à la région de Voronoï dont le centre est leur neurone vainqueur :

$$w_k(t+1) = \frac{\sum_{l=1}^{K} h_{kl}(t) n_l \bar{x}_l}{\sum_{l=1}^{K} n_l h_{kl}(t)} \qquad k \in \{1, \dots, K\} \quad (5)$$

Où n_l est le nombre d'observations appartenant à la région de Voronoï représentée par le neurone l et $\bar{x}_l$ est la moyenne des observations de cette même région (Kohonen, 2001 ; Saad Hajjar, 2014).

Vers la fin de l'apprentissage, quand le rayon de voisinage devient trop petit pour activer seulement le neurone vainqueur, chaque vecteur prototype constitue le centre de gravité des observations qu'il représente et on retombe alors sur l'algorithme des centres-mobiles, ce qui garantit une meilleure approximation de la fonction de densité des observations (Kohonen, 2001).

d) – Evaluation de performance

Une carte auto-organisatrice est évaluée pour ses capacités de quantification et ses capacités de préservation de la topologie. Pour mesurer le degré de déploiement de la carte sur les données ou le degré de quantification, on calcule la moyenne des erreurs de quantification (Mean Quantization Error) qui est définie par :

$$mqe = \frac{\sum_{i=1}^{n} d_2^2(x_i, w_{c_i})}{n} \quad (6)$$

Où d_2^2 est le carré de la distance euclidienne et c_i le neurone vainqueur de l'observation x_i.

Plusieurs critères existent pour quantifier la préservation de la topologie, nous pouvons orienter le lecteur aux travaux des auteurs suivants : Bauer et Pawelzik, 1992 ; Zrehen et Blayo, 1992 et Kiviluoto, 1996.

e) – Choix des paramètres du réseau

Comme dans le cas des réseaux de neurones classiques, il est souvent reproché aux cartes auto-organisatrices le nombre de paramètres à régler avant l'apprentissage, surtout qu'un mauvais choix de l'un de ces paramètres peut conduire à des résultats incohérents (Saad Hajjar, 2014). Les vecteurs prototypes initiaux ont une grande influence sur les résultats finaux au cas où ils sont initialisés aléatoirement. Une solution à ce problème consiste à exécuter plusieurs lancées de l'algorithme d'apprentissage, avec à chaque fois des vecteurs initiaux différents, et à adopter les résultats obtenus par la lancée qui permet de minimiser le plus l'erreur de quantification définie dans l'équation (6).

Sinon, le choix des prototypes initiaux peut se faire de manière déterministe en réalisant une analyse en composantes principales (ACP) des données. Les vecteurs prototypes seront alors positionnés sur les deux axes formés par les deux premiers vecteurs propres de la matrice de covariance des observations. Dans ce cas, l'initialisation des prototypes est linéaire. Une telle initialisation est à préconiser du fait qu'elle diminue le risque des torsions de la carte (Saad Hajjar, 2014).

f) – Critères de convergence

L'algorithme d'apprentissage des cartes auto-organisatrices dans sa version séquentielle ou en mode différé optimise le critère suivant (Ritter et al., 1992 ; Ritter et Schulten, 1988 ; Saad Hajjar, 2014) :

$$G_{SOM} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} h_{kc_i} \; d_2^2(x_i \, , w_k) \tag{7}$$

Mais ce critère pose un problème pour les observations qui se trouvent sur le bord du pavage de Voronoï et ayant plus qu'un neurone vainqueur, ce qui entraîne des points de discontinuités pour la fonction de voisinage h_{kc_i}. Cependant, au cas où les observations sont les réalisations d'une distribution discrète, la probabilité qu'une observation se trouve entre deux vecteurs prototypes sur le bord du pavage de Voronoï est nulle. Pour cette raison, il est possible d'adopter ce critère de convergence pour les ensembles de données finis et discrets (Heskes, 1999 ; Cheng, 1997 ; Saad Hajjar, 2014).
L'organigramme suivant résume les principales étapes de la méthode SOM :

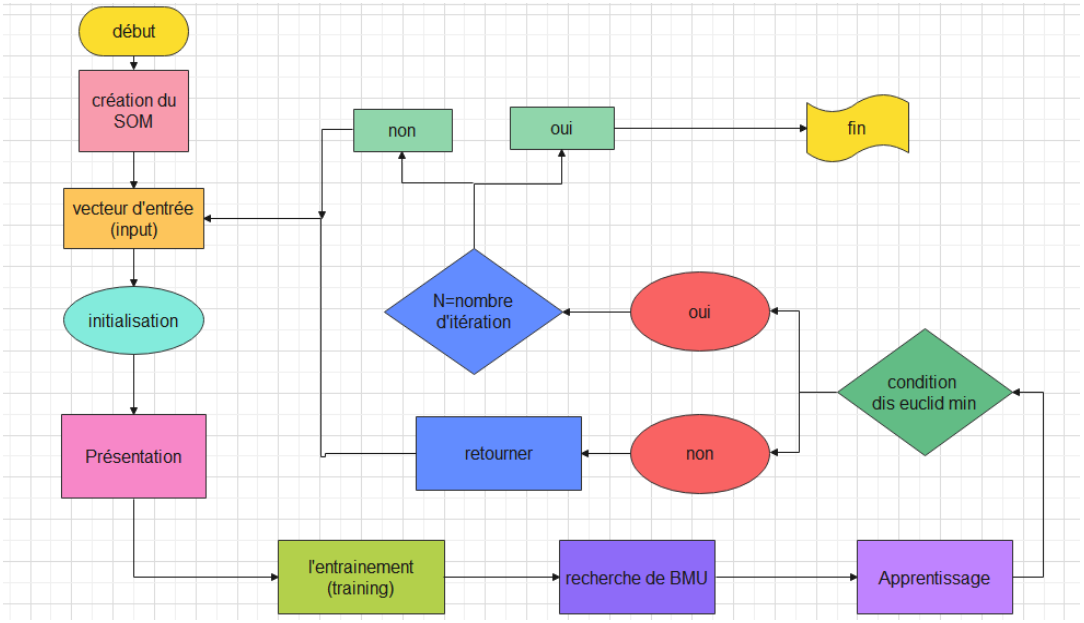


Fig. 10 : - Organigramme simplifié de la méthode SOM

I.2.3.3. Cartes auto-organisatrices pour la classification

Les cartes auto-organisatrices sont souvent utilisées pour la classification non supervisée de données avec préservation de la topologie. A cette fin, chaque neurone représente une classe, et chaque observation est affectée au neurone dont le vecteur référent est le plus proche. De plus, deux neurones voisins sur la carte représenteront des observations qui sont proches dans l'espace d'entrée (Saad Hajjar, 2014).

Pour avoir une meilleure distribution des vecteurs prototypes sur les données, le rayon de voisinage doit atteindre des valeurs assez petites pour n'activer que le neurone vainqueur. Le vecteur prototype correspondant sera alors le centre de gravité de la région de Voronoï qu'il représente. Donc, la classification moyennant les petites cartes se réduit à la méthode des centres-mobiles avec préservation de la topologie (Baçao et al., 2005). La classification moyennant des cartes de petites dimensions présente plusieurs inconvénients notamment : la sensibilité de l'algorithme aux prototypes initiaux, le choix de la topologie de la carte qui est dictée par le nombre des classes et la difficulté de détecter des classes de formes quelconques.

Pour pallier ces inconvénients, des cartes de grandes dimensions sont utilisées, et la reconnaissance de classes de différentes formes peut se faire en inspectant la carte visuellement moyennant les U-matrix (Kohonen, 2001). Une U-matrix ou matrice de distance unifiée (pour Unified distance matrix) est une représentation d'une carte auto-adaptative où les distances euclidiennes entre les poids associés aux neurones voisins sont représentées par une image en tons de gris ou couleur (Fig. 11). Les U-Matrix sont utilisées pour visualiser des données exprimées dans un espace de grandes dimensions sur image 2D (Ultsch et al., 1990).

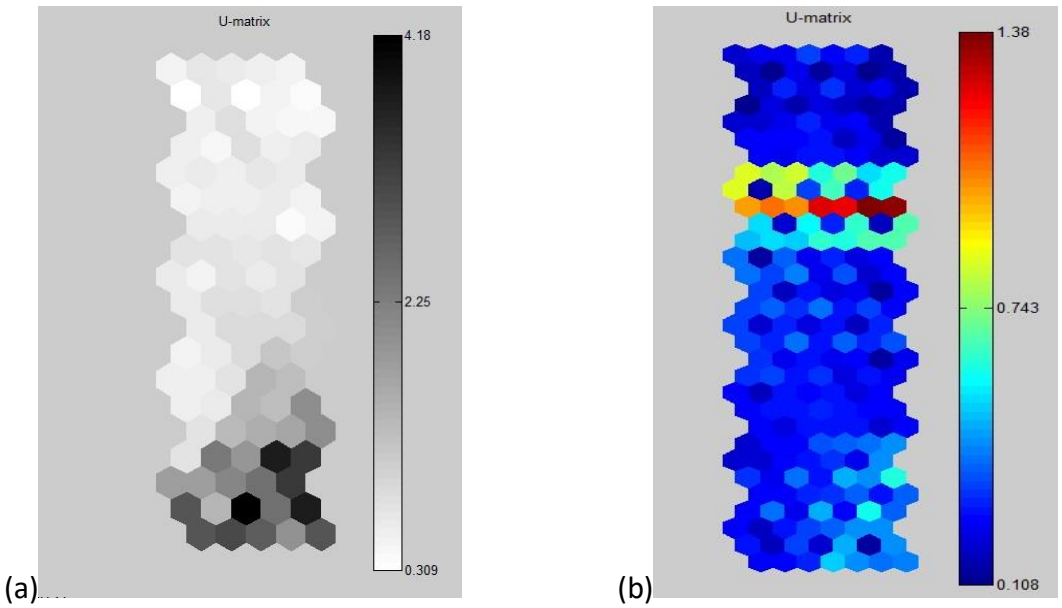


Fig. 11 : Représentation de la matrice de distance unifiée (U-matrix) de la carte auto-organisée.
(a) - en Gris ; (b) - en Couleur

Cependant, une telle détection des classes n'est pas toujours triviale, voire même impossible. Pour ces raisons, nous avons recours à des méthodes automatisées en classifiant les prototypes de la carte moyennant un algorithme de classification standard comme la classification hiérarchique ou la méthode des centres-mobiles (Vesanto et Alhoniemi, 2000).

Il s'agit donc d'une classification en deux étapes : une projection des données sur une carte formée d'un grand nombre de neurones, dans une première étape, suivie d'une classification des prototypes dans une deuxième étape où chaque observation appartiendra à la classe du vecteur prototype de son neurone vainqueur (Fig. 12). Cette technique présente plusieurs avantages (Dong et Xie, 2005 ; Vesanto et Alhoniemi, 2000) :

- Choix libre de la topologie de la carte ;
- Sensibilité réduite quant au choix des prototypes initiaux ;
- Sensibilité réduite aux points aberrants ;
- Réduction de la complexité algorithmique.

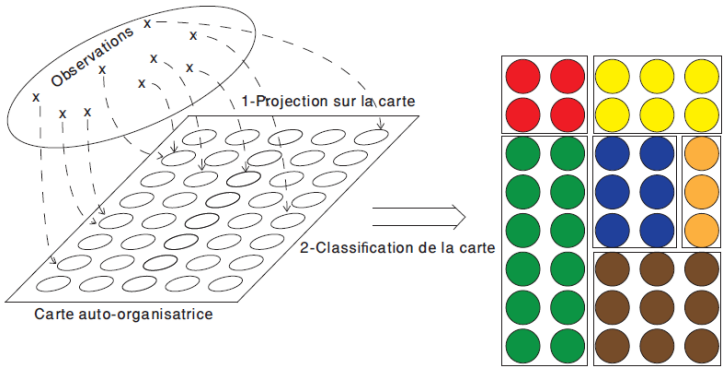


Fig. 12 : - Classification en deux étapes.
(D'après Saad Hajjar, 2014)

I.3. Conclusion :

Dans ce chapitre, nous avons décrit sommairement les différentes méthodes de classification supervisée et non supervisé en intéressant de la méthode SOM qui fait partie de la famille des méthodes non supervisés.

Dans ce chapitre, nous sommes intéressés plus particulièrement aux cartes auto-organisatrices de Kohonen comme méthode de classification automatique où nous avons présenté sommairement le concept de la méthode. Nous avons également décrit les principales étapes de calcul en 2 modes différé et séquentiel en présentant les relations mathématique du chaque mode ensuite on a résumé la méthode SOM par un organigramme simplifié.

Les méthodes exposées dans ce chapitre s'appliquent à des données quantitatives où chaque observation est représentée par un point dans l'espace. Cependant, les données issues des expériences de la vie réelle sont souvent représentées sous plusieurs formes par exemple, une observation peut être de type qualitatif, modale, intervalle, ou multi-valuée, donc il est nécessaire de généraliser les méthodes classiques pour prendre en compte les différents types de données.

Le chapitre 2 présente une détaille de programme de la méthode SOM, avec une première application sur un exemple donné.

Chapitre 2

Organigrammes et Programmes de Calcul

II.1. Introduction :

Dans ce chapitre, on cherche à appliquer la méthode SOM sur des échantillons pour tester l'efficacité du code de calcul sans pour autant détailler l'interprétation des résultats. On tentera d'essayer de maitriser le code et obtenir plus de résultats afin d'explorer la méthode. Notre objectif est de présenter sommairement les différents algorithmes et leurs organigrammes, ainsi que les fonctions les plus pertinentes utilisées par le programme principal. Par la suite, on présentera les résultats obtenus avec différentes options permettant de meilleures visualisations des résultats.

II.2. Description du modèle :

Généralement, on peut résumer la méthode SOM par l'organigramme simplifié de figure 13. L'organigramme débute par le nombre d'itération, le choix du vecteur d'entrée comme variables aléatoires, la recherche optimale de la distance minimale et par l'actualisation des poids. Cependant, plusieurs étapes sous-jacentes existent pour construire les cartes auto-organisatrices de Kohonen basées sur la méthodologie décrite précédemment et qui nécessitent certaines fonctions appropriées à l'aide de nombreux paramètres qui seront illustrés plus loin.

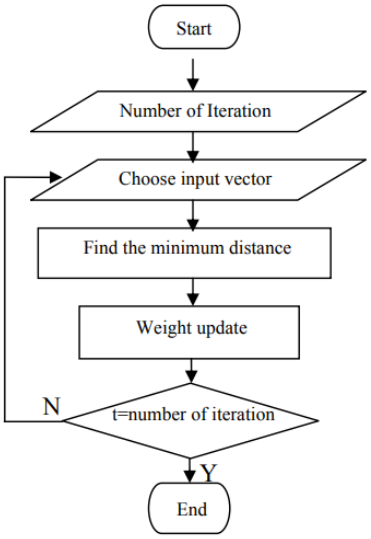


Fig.13 : - Organigramme simplifié de la méthode SOM.
(d'après Anifah et al., 2013)

La carte auto-adaptive est construite par deux espaces indépendants (Lemaire, 2006), le premier espace est de grande dimension correspondant à la couche d'entrée du réseau de neurones, et le deuxième est d'une dimension réduite correspondant à la couche de sortie (Fig.14). Pour conserver la typologie des données, il est essentiel de trouver la projection entre les deux espaces.

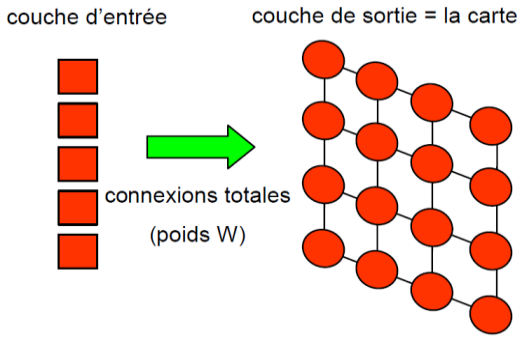


Fig.14 : - Description de la carte (d'après Lemaire, 2006)

Chaque nœud dans la carte possède des coordonnées ***fixes*** sur la carte, et des coordonnées ***adaptables*** *W* (poids) dans l'espace d'entrée original.

La structure topologique de la carte introduit la notion de voisinage et de distance entre les neurones de la carte. Dans ce travail, on a opté pour un voisinage hexagonal.

On peut résumer les principales étapes de l'algorithme SOM par les points suivants :

1) Création des données
2) Visualisation
3) Analyse des résultats

II.2.1. Création des données :

La boîte à outils SOM de Matlab a une structure spéciale permettant de regrouper les informations concernant l'ensemble de données :

- On commence par importer les données à l'aide de la fonction som_data_struct Si les données existent déjà dans la base de données, on utilise la fonction som_read_data pour lire le fichier de données ;
- On suite, on procède à la Normalisation des données à l'aide de la fonction som_normalize et la fonction som_denormalize pour la Dénormalisation des données.

On introduit les commandes sous Matlab comme suit :

```
%**************************************************************************
%               First, the data is read from ascii file
clf reset;
f0 = gcf;
echo on
clc
try,
sD = som_read_data('file.dat');
catch
echo off
  warning('File "file.dat" not found. Using simulated data instead.')
 D = randn(100000,11);
 D2 = randn(100000,11); D2(:,2) = sort(D2(:,2));
 sD = som_data_struct([D; D2],'name','file (simulated)',...
                              'comp_names',{'Ca','Mg','Na','K','Cl','SO4','HCO3','NO3'...
                        'CE','Ph','T°'});
sD = som_label(sD,'add',[i:j]','L(i)');
echo on
end
sD = som_normalize(sD);
sM = som_make(sD);
sM = som_autolabel(sM,sD);
%**************************************************************************
```

Dans ce travail, l'algorithme SOM est basé sur les distances euclidiennes, l'échelle des variables sont très importantes pour déterminer la carte auto-organisatrice. Si la gamme de valeurs d'une variable est beaucoup plus grande que celle des autres variables, cette variable dominera probablement complètement l'organisation de la carte. Pour cette raison, les composants de l'ensemble de données sont généralement normalisés.

- Ensuite, on fait appel à la fonction som_make qui crée, initialise et entraîne la carte auto-organisatrice. Les fonctions som_randinit et som_lininit seront utilisées pour initialiser les vecteurs prototypes de la carte. La taille de la carte est en fait un

argument facultatif. La fonction som_randinit initialise la carte avec des valeurs aléatoires. La fonction som_lininit initialise la carte linéairement.

- Pour le training ou l'entrainement ou encore l'apprentissage, on utilise la fonction som_seqtrain. Cette fonction utilise un algorithme séquentiel pour former la carte auto-organisatrice.

II.2.2. Visualisation :

La visualisation de base de la méthode SOM se fait avec la fonction som_show, mais plusieurs fonctions sont utilisées avec la fonction principale som_show. L'organigramme suivant (Fig.15) résume les principales étapes de l'opération visualisation de la méthode SOM. Les principales commandes de la visualisation sous Matlab sont exécutées comme suit :

```
%*************************************************************************
clc
%                VISUAL INSPECTION OF THE MAP
echo off
f1=figure;
[Pd,V,me,l] = pcaproj(sD,2); Pm = pcaproj(sM,V,me);
Code = som_colorcode(Pm);
hits = som_hits(sM,sD);
U = som_umat(sM);
Dm = U(1:2:size(U,1),1:2:size(U,2));
Dm = 1-Dm(:)/max(Dm(:)); Dm(find(hits==0)) = 0;
subplot(1,3,1)
som_cplane(sM,Code,Dm);
hold on
som_grid(sM,'Label',cellstr(int2str(hits)),...
            'Line','none','Marker','none','Labelcolor','k');
hold off
title('Color code')
subplot(1,3,2)
som_grid(sM,'Coord',Pm,'MarkerColor',Code,'Linecolor','k');
hold on, plot(Pd(:,1),Pd(:,2),'k+'), hold off, axis tight, axis equal
title('PC projection')
subplot(1,3,3)
som_cplane(sM,'none')
hold on
som_grid(sM,'Label',sM.labels,'Labelsize',8,...
            'Line','none','Marker','none','Labelcolor','r');
hold off
title('Labels')
echo on
%*************************************************************************
```

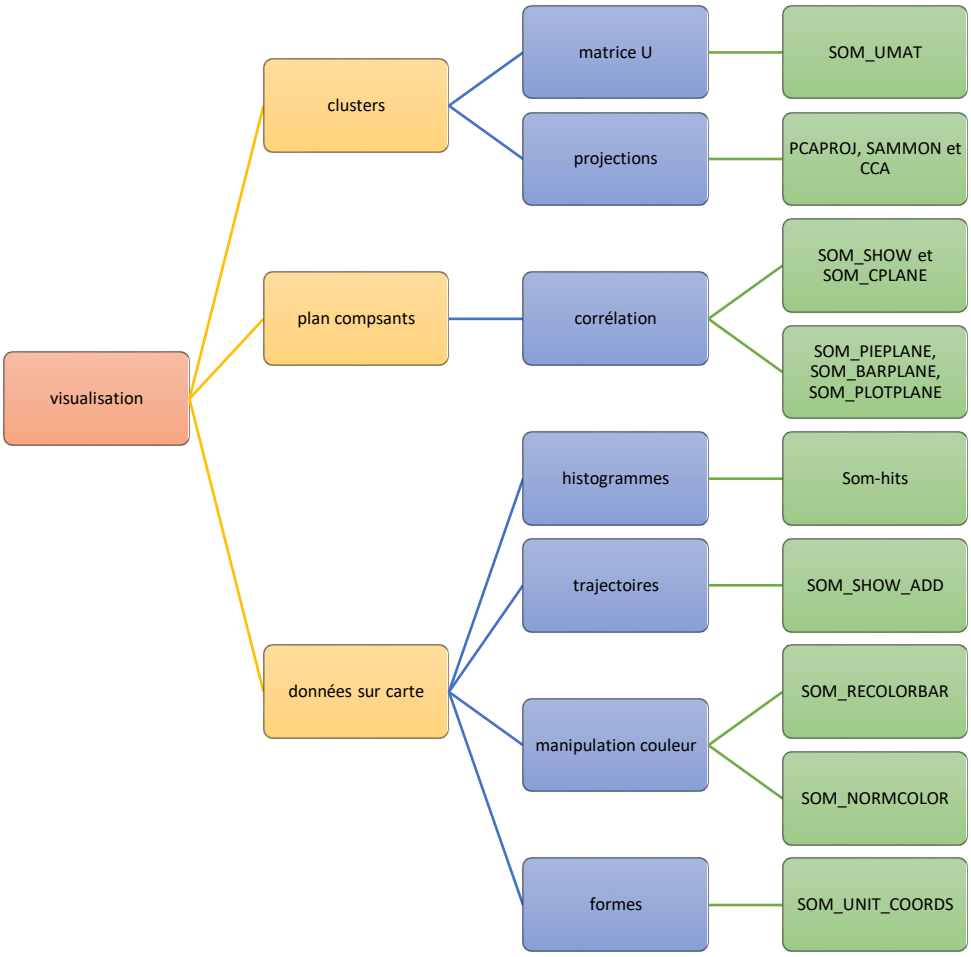


Fig.15 : - Organigramme montrant les étapes et les différentes fonctions permettant la visualisation de la carte auto-organisatrice.

I.2.3. Analyse des résultats :

L'analyse des résultats s'effectue en deux étapes :

- Recherche du BMU (neurone gagnant) qui s'effectue à l'aide de la fonction som_bmus ;
- Mesurer la qualité du modèle qui s'effectue à l'aide de la fonction som_quality pour quantifier l'erreur.

```
- %**************************************************************************
- % Denormalize and add species information
- names = sD.comp_names; names{end+1} = 'L(i)';
- D = som_denormalize(sD.data,sD); dlen = size(D,1);
- s = zeros(dlen,1)+NaN; s(strcmp(sD.labels))=1;
- s(strcmp(sD.labels,'P(i)'))=2; s(strcmp(sD.labels,'F(j)'))=3;
- D = [D, s];
- M = som_denormalize(sM.codebook,sM); munits = size(M,1);
- s = zeros(munits,1)+NaN; s(strcmp(sM.labels,'L(i)'))=1;
- s(strcmp(sM.labels))=2; s(strcmp(sM.labels))=3;
- M = [M, s];
- f2=figure;
- bmus = som_bmus(sM,sD); Code_data = Code(bmus,:);
- %**************************************************************************
```

La fonction som_bmus cherche à trouver les meilleures unités de correspondance (BMU) pour un vecteur de données donné à partir d'une carte donnée. Elle renvoie les index et les erreurs de quantification correspondantes des vecteurs dans la carte qui correspondent le mieux aux vecteurs dans les données. La fonction est donnée par le programme suivant :

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Bmus,Qerrors] = som_bmus(sMap, sData, which_bmus, mask)
error(nargchk(1, 4, nargin));  % check no. of input args is correct
if isstruct(sMap),
  switch sMap.type,
   case 'som_map', M = sMap.codebook;
   case 'som_data', M = sMap.data;
   otherwise, error('Invalid 1st argument.');
  end
else
  M = sMap;
end
[munits dim] = size(M);
if any(any(isnan(M))),
  error ('Map codebook must not have missing components.');
end
if isstruct(sData),
  switch sData.type,
   case 'som_map', D = sData.codebook;
   case 'som_data', D = sData.data;
   otherwise, error('Invalid 2nd argument.');
  end
else
  D = sData;
end
[dlen ddim] = size(D);
if dim ~= ddim,
  error('Data and map dimensions do not match.')
end
if nargin < 3 | isempty(which_bmus) | any(isnan(which_bmus)),
  which_bmus = 1;
else
```

```
  if ischar(which_bmus),
    switch which_bmus,
     case 'best', which_bmus = 1;
     case 'worst', which_bmus = munits;
     case 'all', which_bmus = [1:munits];
    end
  end
end
if nargin < 4 | isempty(mask) | any(isnan(mask)),
  if isstruct(sMap) & strcmp(sMap.type,'som_map'),
    mask = sMap.mask;
  elseif isstruct(sData) & strcmp(sData.type,'som_map'),
    mask = sData.mask;
  else
    mask = ones(dim,1);
  end
end
if size(mask,1)==1, mask = mask'; end
if all(mask == 0),
  error('All components masked off. BMU search cannot be done.');
end
Bmus = zeros(dlen,length(which_bmus));
Qerrors = Bmus;
blen = min(munits,dlen);
Known = ~isnan(D);
W1 = (mask*ones(1,dlen)) .* Known';
D(find(~Known)) = 0;
unknown = find(sum(Known')==0); % completely unknown vectors
WD = 2*diag(mask)*D';
dconst = ((D.^2)*mask);
i0 = 0;
while i0+1<=dlen,
  inds = [(i0+1):min(dlen,i0+blen)]; i0 = i0+blen;
  Dist = (M.^2)*W1(:,inds) - M*WD(:,inds); % plus dconst for each sample

  if all(which_bmus==1), [Q B] = min(Dist); else [Q B] = sort(Dist); end
  if munits==1, Bmus(inds,:) = 1; else Bmus(inds,:) = B(which_bmus,:)'; end
  Qerrors(inds,:) = Q(which_bmus,:)' + dconst(inds,ones(length(which_bmus),1));
end
if ~isempty(unknown),
  Bmus(unknown,:) = NaN;
  Qerrors(unknown,:) = NaN;
end
Qerrors = sqrt(Qerrors);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
```

La fonction som_quality utilise deux critères d'évaluation : la résolution et la préservation de la topologie. Si la dimension de l'ensemble de données est supérieure à la dimension de la grille de la carte, ceux-ci deviennent généralement des objectifs contradictoires.

La première valeur renvoyée par cette fonction mesure la résolution et la seconde la préservation de la topologie.

- qe : Distance moyenne entre chaque vecteur de données et son BMU.
- te : Erreur topographique, la proportion de tous les vecteurs de données pour lesquels la première et la deuxième BMU ne sont pas des unités adjacentes.

La fonction som_quality est illustrée par le programme suivant :

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [mqe,tge] = som_quality(sMap, D)
if nargin < 2, error('Not enough input arguments.'); end
if isstruct(D), D = D.data; end
[dlen dim] = size(D);
if nargout==1, b=1; else b=[1:2]; end
[bmus qerrs]= som_bmus(sMap,D,b);
inds = find(~isnan(bmus(:,1)));
bmus = bmus(inds,:);
qerrs = qerrs(inds,:);
l = length(inds);
if ~l, error('Empty data set.'); end
mqe = mean(qerrs(:,1));
if length(b)==2,
  Ne = full(som_unit_neighs(sMap.topol));
  tge = 0;
  for i=1:l, tge = tge+(Ne(bmus(i,1),bmus(i,2)) ~= 1); end
  tge = tge / l;
else
  tge = NaN;
end
return;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

II.3. Démonstration en 2 dimensions :

Dans cette démonstration, on commence tous d'abord par un exemple par deux séries de variables aléatoires afin de montrer les propriétés de base et le comportement de la carte auto-organisatrice. En fait, chaque unité cartographique peut être considérée comme ayant deux ensembles de coordonnées :

- dans l'espace d'entrée : les vecteurs prototypes ;
- dans l'espace de sortie : la position sur la carte.

Dans les deux espaces, la carte ressemble à ceci :

Dans la phase initialisation de la carte, les points noirs affichent les positions des unités cartographiques, et les lignes grises montrent les connexions entre les unités cartographiques voisines. Les croix rouges représentent les données aléatoires

d'apprentissage. Les positions des unités cartographiques dans l'espace d'entrée sont complètement désorganisées (Fig.16).

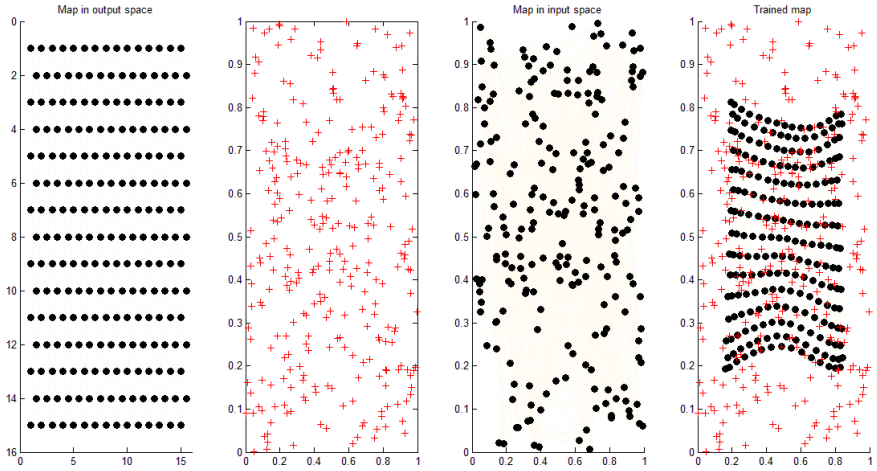


Fig.16 : - Initialisation de la carte auto-organisatrice :
a) – Carte dans l'espace de sortie ;
b) – Données aléatoires d'apprentissage ;
c) – Carte dans l'espace d'entrée ;
d) – Carte après entraînement.

La figure suivante (fig.17), montre en (a) les données et la carte originale et en (b) la carte auto-organisatrice par un apprentissage séquentiel après 300 itérations.

L'apprentissage repose sur deux principes :

- Apprentissage compétitif : le vecteur prototype le plus similaire ou semblable à un vecteur de données est modifié pour qu'il lui soit encore plus similaire. De cette façon, la carte apprend la position du nuage de données.

- Apprentissage coopératif : non seulement le vecteur prototype le plus similaire, mais aussi ses voisins sur la carte sont déplacés vers le vecteur de données. De cette façon, la carte s'auto-organise.

Pendant l'apprentissage, la carte s'organise et se replie sur les données d'entraînement. Dans ce travail, l'algorithme d'apprentissage séquentiel est utilisé.

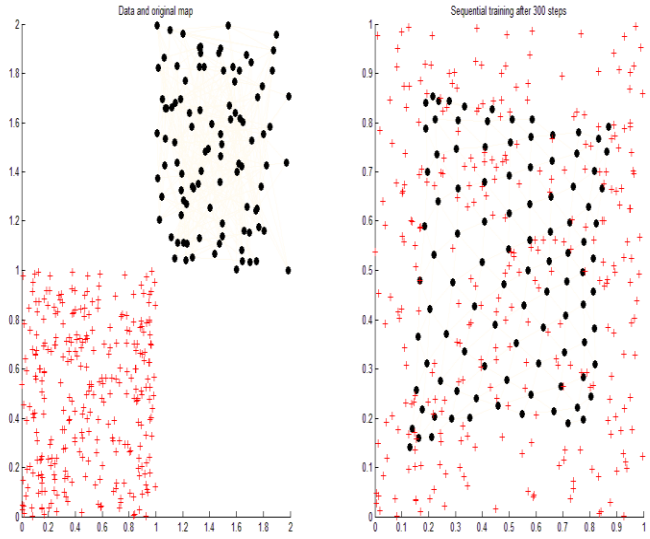


Fig.17 : - Apprentissage séquentiel après 300 itérations :
a)- Carte des données et carte auto-organisatrice originale ;
b)- Carte auto-organisatrice obtenue par apprentissage séquentiel.

Afin de voir le déploiement de la carte auto-organisatrice sur les données, on a procédé à plusieurs simulations en faisant varier le nombre d'itération de 100, 500, 1000 et 10000.

Les résultats obtenus sont présentés dans la figure 18. Elle montre que la carte auto-organisatrice se déploie mieux pour 10000 itérations que les autres.

Le BMU (pour Best-Matching Unit) d'un vecteur de données est l'unité sur la carte dont le vecteur modèle ressemble le plus au vecteur de données. En pratique, la similarité est mesurée comme la distance minimale entre le vecteur de données et chaque vecteur modèle sur la carte. Les BMU peuvent être calculés à l'aide de la fonction som_bmus comme il a été décrit précédemment. Cette fonction donne l'indice de l'unité. Pour cet exemple, le BMU est recherchée pour le point d'origine (à partir de la carte entraînée).

Ici, l'unité correspondante est représentée sur la figure 19 (c). On peut faire pivoter la figure 19 (c) pour mieux voir où se trouve le BMU.

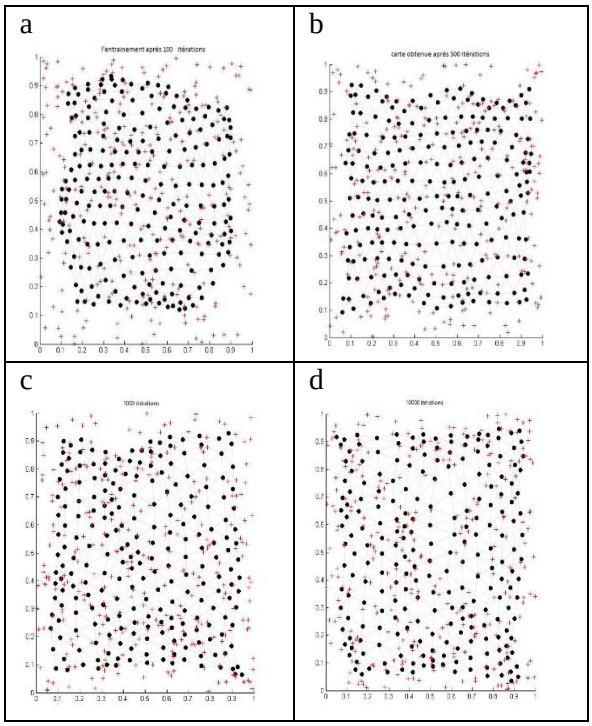


Fig. 18 : - Cartes auto-organisatrices après apprentissage pour différentes itérations :
a)- pour 100 itérations, b)- pour 500 itérations, c)- pour 1000 itérations et d)- pour 10000 itérations.

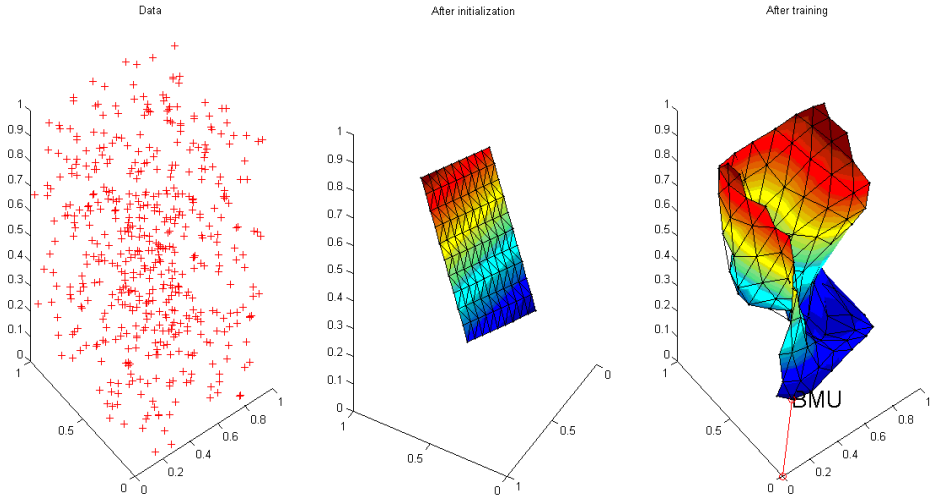


Fig.19 : Cartes 3D des données et auto-organisatrices et BMU pour le point d'origine :
a)- Cartes des données ;
b)- Carte auto-organisatrice après initialisation ;
c)- Carte auto-organisatrice après apprentissage.

II.4. Exemple d'analyse exploratoire pour les données Hydrochimiques :

Pour cet exemple, notre objectif est d'appliquer la méthode SOM aux données Hydrochimiques de la nappe alluviale de la vallée de l'Oued M'Zi. Cet exemple est surtout pour mieux maitriser l'outil SOM et tenter de nouvelles classifications et une première interprétation. Il est également l'occasion pour vérifier l'efficacité du programme de calcul.

II.4.1. Description des données :

On dispose de 35 échantillons d'eau prélevés de la nappe alluviale de l'Oued M'Zi (Bordj Senouci, El Fetha, El Assafia, Nacer Benchohra, Slisla, Taouenza, Kabeg et K'Sar El Hirane). Ces échantillons ont été prélevés lors d'une campagne de prélèvement effectuée en Janvier 2017.

Les éléments chimiques analysés sont : Ca, Mg, Na, K, Cl, SO4, HCO3 et NO3 (en mg/l) et les paramètres physico-chiques sont : Ph, CE (Conductivité Electrique) et la Température. Le tableau suivant résume les principales caractéristiques statistiques des éléments analysés :

Eléments et paramètres	Min	Max	Moyenne	Ecart-type	CV %
Ca	4.00	52.00	14.6949	9.2440	62.91
Mg	5.6	153.2	20.1794	28.9154	1.4329
Na	1.88	75.21	10.0643	13.0019	1.2919
K	0.02	1.11	0.4157	0.2509	60.35
Cl	2.8	128	14.5829	23.9058	1.6393
SO4	8.87	155.64	29.3943	30.4438	1.0357
HCO3	0.02	4.51	0.6251	0.8840	1.4141
NO3	1.8	6.12	2.7886	0.85	30.48
Ph	6.72	7.86	7.1586	0.2431	3.4
CE	1001	17500	3147.943	3118.4	99.06
T°	11	23.8	19.6714	2.4788	12.60

Tableau n°1 : Caractéristiques statistiques des propriétés physico-chimiques des échantillons d'eau de la nappe de la vallée de l'Oued M'Zi.

II.4.2. Construction de la carte auto-organisatrice :

a) – Le vecteur d'entrée de la carte présente dans ce cas une dimension de 11 correspondant au nombre de variables ;
b) – Dans ce cas aussi, les paramètres de la carte correspondent :
   - à une matrice M*N (Nombre de Neurones) M=11 et N= 35 ;
   - et à un voisinage Hexagonal.

Après l'initialisation et l'apprentissage, chaque vecteur d'entrée (individu) est associé à un neurone de la carte celui qui est le plus proche de l'individu.

Dans cette phase, la première représentation est celle des Histogrammes et de nuages de points des variables. A titre d'illustration, la figure 20 présentes les histogrammes des 5 premières variables (Conductivité, Ph, T°, Ca et Mg). Les graphiques des nuages de points montrent la relation entre les deux variables deux à deux. On constate déjà la mauvaise corrélation entre la conductivité, Ph et Température et la bonne corrélation entre la conductivité et les éléments chimiques Ca et Mg.

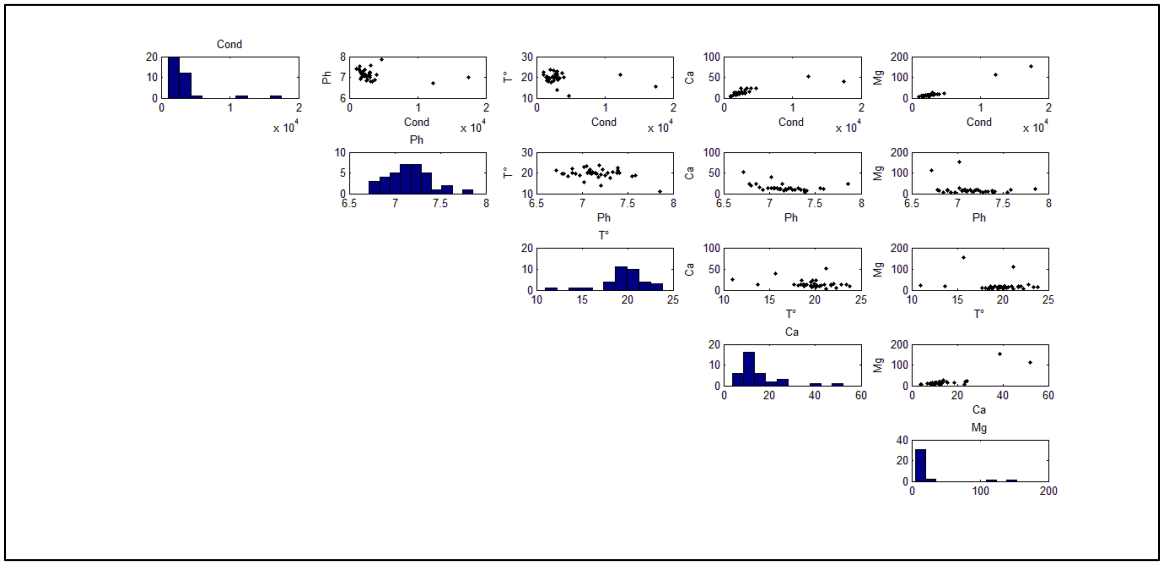


Fig.20 : - Histogrammes et nuages de points des variables : conductivité, Ph, T°, Ca et Mg

Comme il a été décrit précédemment, l'algorithme SOM est basé sur les distances euclidiennes, l'échelle des variables est très importante pour déterminer à quoi ressemblera la carte. A ce titre, les éléments de l'ensemble de données sont normalisés à l'aide de la fonction som_normalize pour que chaque composant ait une variance unitaire.

Cependant, l'interprétation des valeurs peut être plus difficile quand elles sont normalisées. Par conséquent, les opérations de normalisation peuvent être inversées avec la fonction som_denormalize pour une meilleure interprétation.

La fonction som_make est utilisée pour l'apprentissage de la carte SOM. La fonction détermine d'abord la taille de la carte, puis initialise la carte à l'aide d'une initialisation linéaire et enfin utilise un algorithme pour l'apprentissage de la carte comme il a été décrit précédemment dans le paragraphe II.2.1.

II.4.3. Visualisation de la carte auto-organisatrice SOM :

Notons que les valeurs des variables ont été dénormalisées à l'échelle d'origine avant la visualisation des cartes. La visualisation se fait à l'aide de la fonction som_show. La figure 21 montre la carte U-matrix (matrice de distance unifiée) et les cartes de chaque variable en gris.

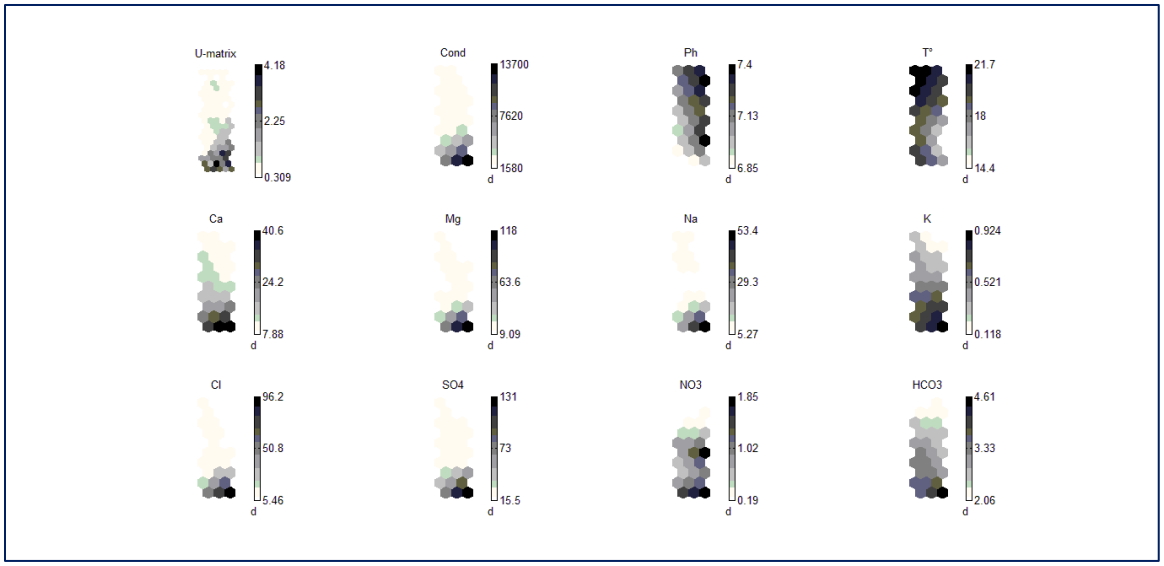


Fig.21 : - Visualisation de la carte SOM : U-matrix, Cond, Ph, T°, Ca, Mg, Na, K, Cl, SO4, HCO3 et NO3.

La U-matrix montre les distances entre les unités voisines et visualise ainsi la structure de cluster de la carte. On constate que la visualisation de la U-matrix a beaucoup plus d'hexagones que les plans des autres composants. C'est parce que les distances entre les unités de la carte sont affichées, et pas seulement les valeurs de distance aux unités de la carte. Les valeurs élevées sur la U-matrix signifient une grande distance entre les unités cartographiques voisines et indiquent ainsi les frontières des clusters. Les clusters sont généralement des zones uniformes de faibles valeurs. On se reporte à la barre de couleurs pour voir quelles couleurs signifient des valeurs élevées. Dans cette carte, il semble y avoir trois clusters.

Pour les éléments Ca, Mg, Na, K, Cl et SO4, ils présentent des structures assez homogènes indiquant ainsi leur même origine, ils sont d'ailleurs très bien corrélés à la conductivité. Par ailleurs, les autres composants comme le NO3, le Ph et la Température sembles avoir des origines diverses.

La figure 22 montre la U-matrix et la répartition régionale des échantillons afin de les mettre dans leurs clusters et les limiter en un certain nombre de groupe.

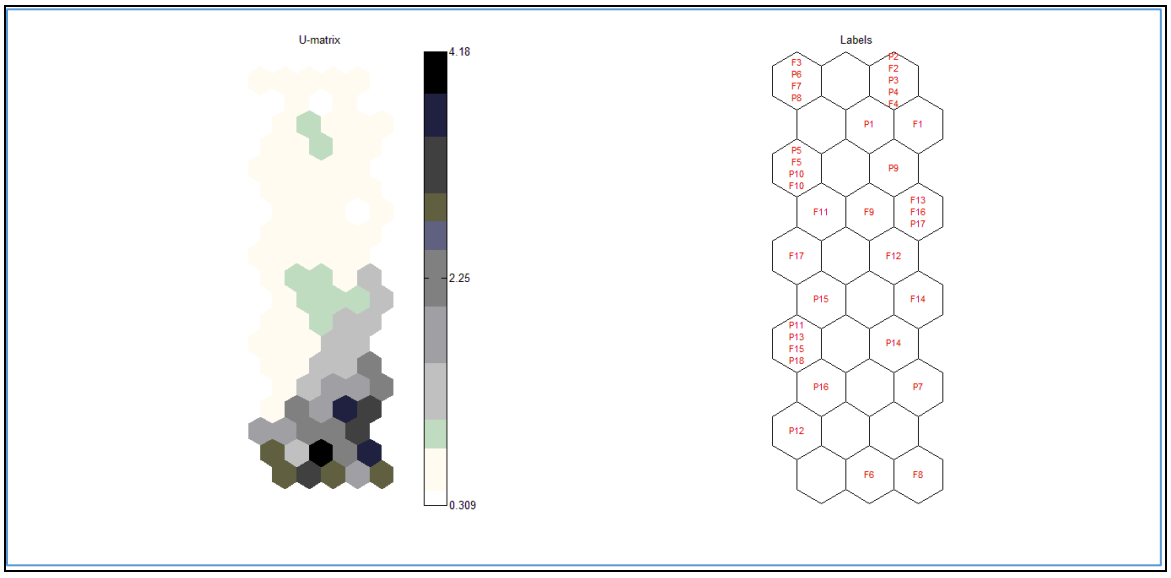


Fig.22 : - Représentation en gris de la U-matrix et de la répartition des identifiants des échantillons.

II.4.4. Clustering de la carte ;

Un outil important dans l'analyse des données à l'aide de SOM est ce qu'on appelle les histogrammes de hits. La fonction Hits permet de calculer le nombre d'occurrences de chaque valeur. Les histogrammes de hits sont formés en prenant un ensemble de données, en trouvant le BMU de chaque échantillon de données de la carte et en augmentant un compteur dans une unité de carte chaque fois qu'il s'agit du BMU. L'histogramme des résultats montre la répartition de l'ensemble de données sur la carte. Dans la figure 23, l'histogramme des hits pour l'ensemble des données est calculé et visualisé sur la U-matrix où on constate les trois clusters.

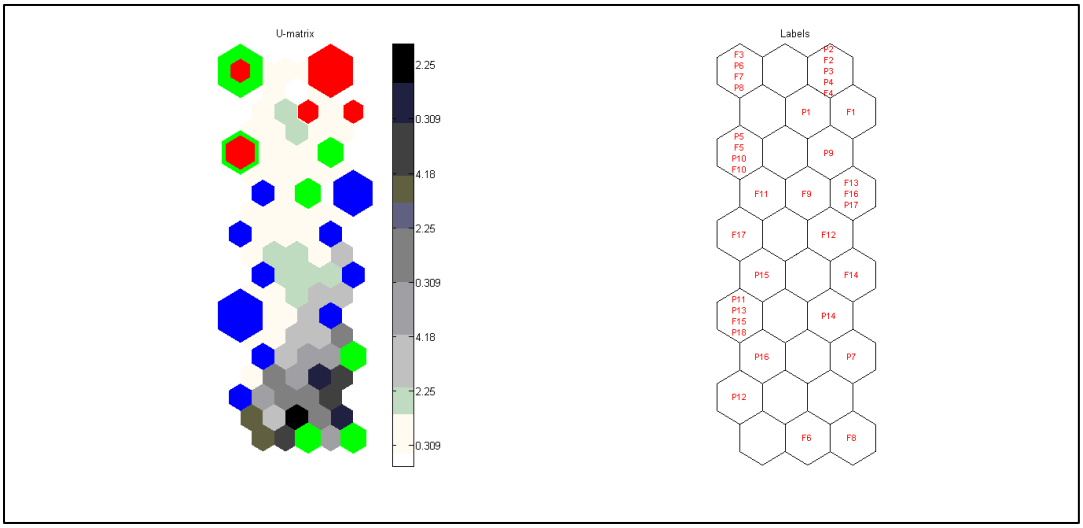


Fig.23 : - Visualisation des Histogrammes de Hits sur la U-matrix

Une autre visualisation de la carte SOM peut être également présentée en couleur avec quelques options supplémentaires (Fig.24).

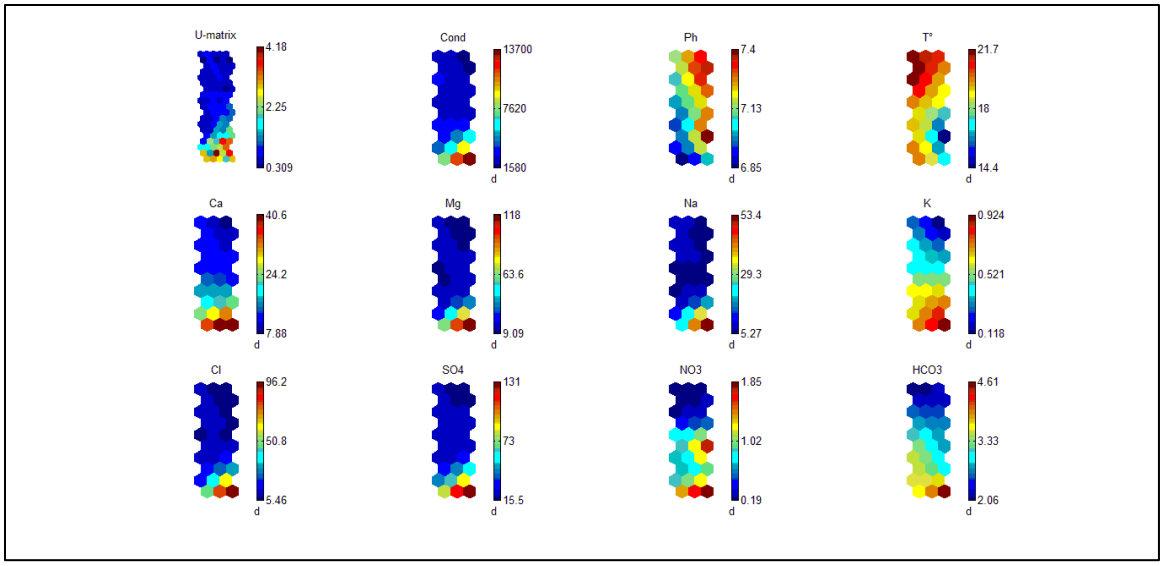


Fig.24 : Carte SOM de la U-matrix et de chaque individu ainsi que la répartition des identifiants des échantillons.

Par la suite, la projection de l'ensemble de données est étudiée et illustrée par la figure 25. Une projection des composants principaux est réalisée pour les données et appliquée à la carte (Fig.25). La palette de couleurs se fait en étalant une palette de couleurs sur la projection. Les informations de la matrice de distance sont extraites de la U-matrix, et elles sont modifiées par la connaissance des unités de zéro-Hits (interpolation). Enfin, trois visualisations sont affichées : le code couleur, avec les informations du clustering et le nombre de hits dans chaque unité, la projection et les étiquettes ou les identifiants (individus). A partir de ces figures, on peut voir clairement que la projection confirme l'existence de trois clusters différents.

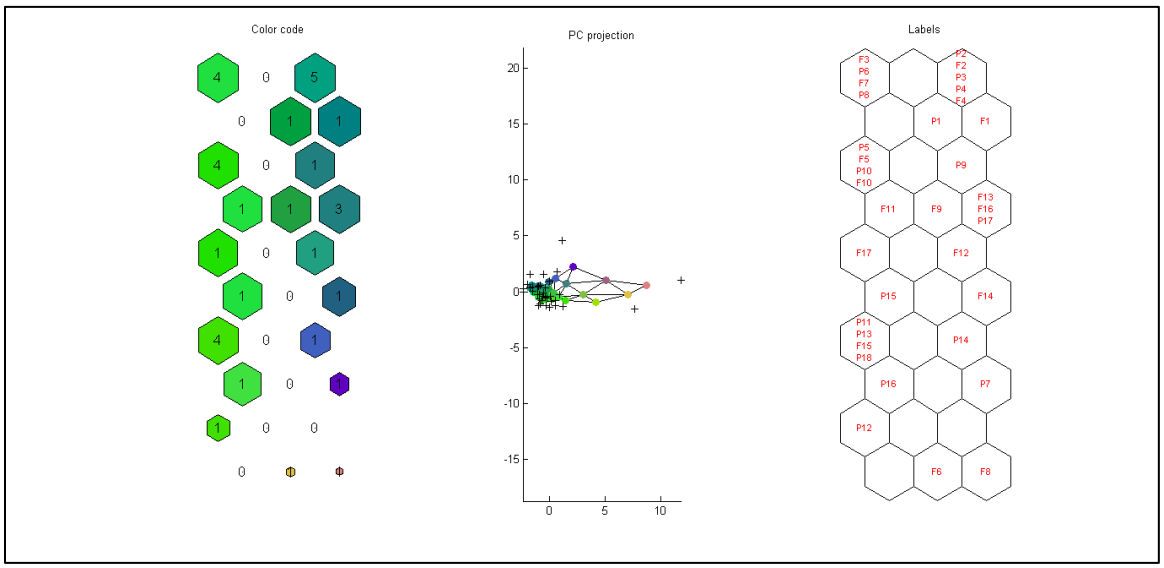


Fig.25 : - Code couleur, classement et nombre de Hits, Projection PC et Identifiants

Enfin, peut-être la figure la plus informative de toutes est celle qui présente les diagrammes de dispersion simples et les histogrammes de toutes les variables (Fig.26). Les points de données d'origine se trouvent dans le triangle supérieur, les valeurs du prototype de carte dans le triangle inférieur et les histogrammes sur la diagonale : noir pour l'ensemble de données et rouge pour les valeurs de prototype de carte. Le codage couleur des échantillons de données a été copié à partir de la carte (à partir du BMU de chaque échantillon). On note cependant là aussi, que les valeurs des variables ont été dénormalisées. Nous constatons sur les valeurs du prototype les relations existants entre les variables d'une façon nette et convaincante.

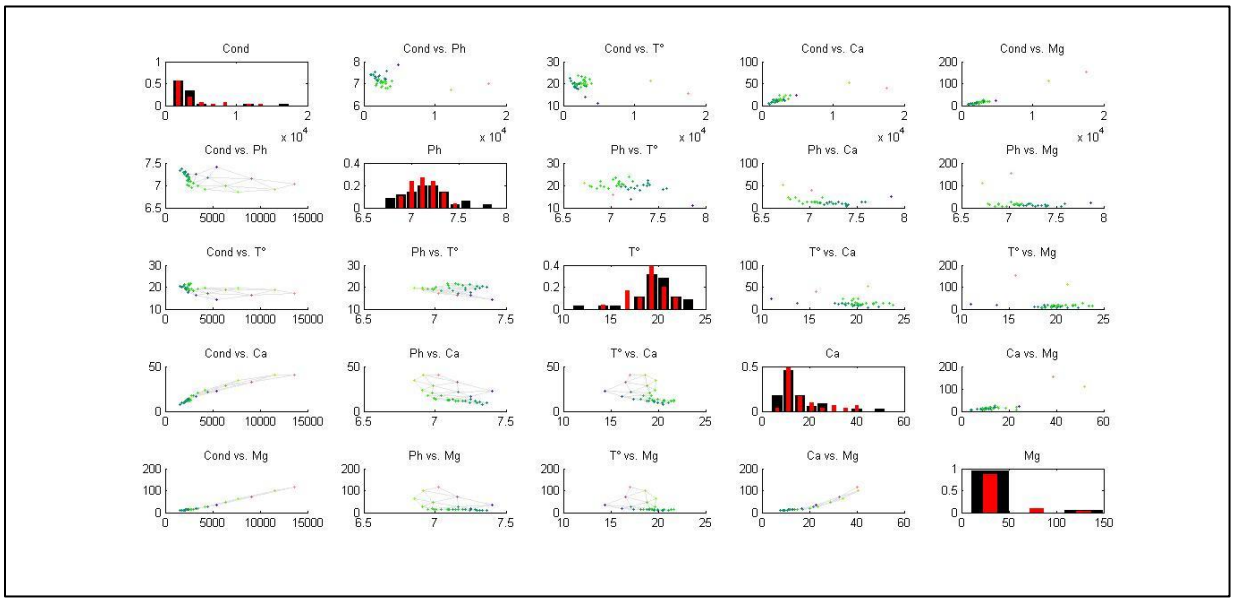


Fig.26 : - Valeurs du prototype de la carte, points de données d'origine et Histogrammes sur la diagonale.

L'inspection visuelle a déjà laissé entendre qu'il y a trois clusters dans les données, et que les propriétés des clusters sont différentes les unes des autres. Pour une enquête plus approfondie, la carte doit être partitionnée. Cependant, la fonction kmeans_clusters permet de retrouver un partitionnement initial. Le graphique montre l'indice de clustering Davies-Boulding (Fig.27), qui est minimisé avec le meilleur clustering. L'indice Davies-Bouldin semble indiquer qu'il y a deux clusters sur la carte. Les informations de clustering calculées précédemment et le résultat du partitionnement sont présentés dans la figure 27. On peut également utiliser la fonction som_select pour créer ou modifier manuellement le partitionnement.

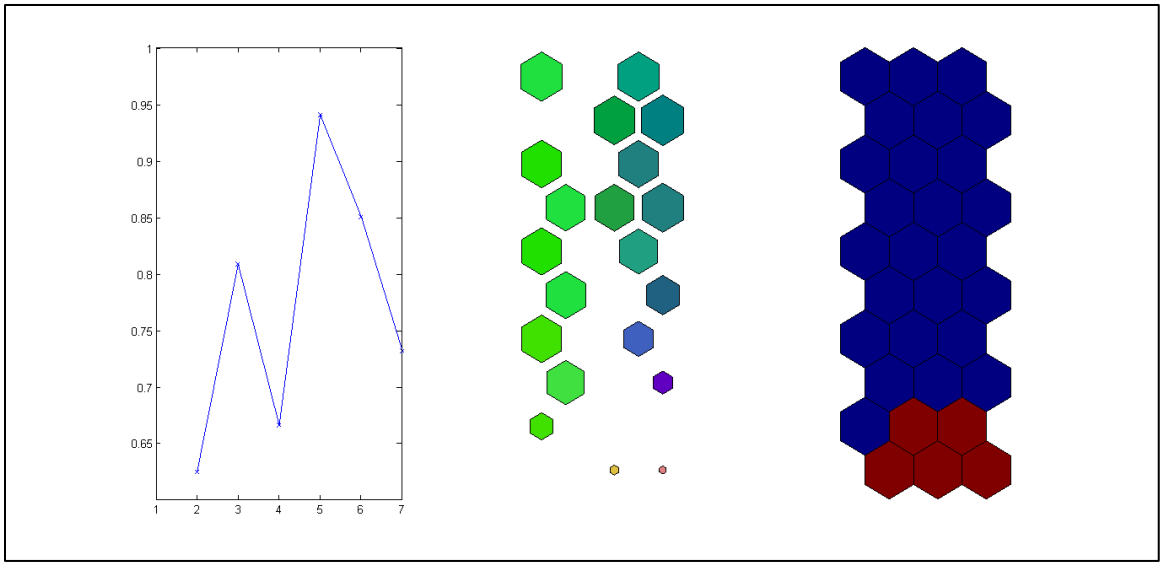


Fig.27 : - Indice de clustering Davies-Boulding et classification

Après cela, l'analyse procéderait à la synthèse des résultats et à l'analyse de chaque groupe un à la fois.

II.4.5. Modélisation :

On peut également construire des modèles à partir de la carte auto-organisatrice SOM. Typiquement, ces modèles sont de simples modèles locaux ou du plus proche voisin. Dans ce cas, SOM est utilisé pour l'estimation de la densité de probabilité.

Chaque prototype de carte est le centre d'un noyau gaussien dont les paramètres sont estimés à partir des données. Le modèle de mélange gaussien est estimé avec la fonction som_estimate_gmm et les probabilités peuvent être calculées avec la fonction som_probability_gmm.

A titre indicatif et pour illustration, la valeur de la fonction de densité de probabilité pour le premier échantillon de données (Conductivité) en termes de chaque unité de carte est donnée par la figure 28 ci-contre.

L'algorithme SOM développé comme décrit précédemment, est utilisé pour la classification. Bien que l'algorithme puisse être utilisé pour la classification en tant que tel, il faut se rappeler qu'il n'utilise pas du tout les informations de classe, et donc ses résultats sont intrinsèquement sous-optimaux.

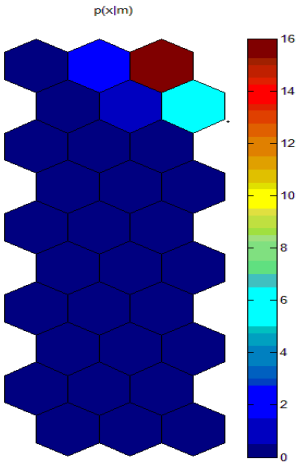


Fig.28 : - Fonction de densité de probabilité en termes d'unité de carte de la conductivité

Cependant, avec de petites modifications, le réseau peut prendre en compte la classe. La fonction som_supervised fait cette option.

L'apprentissage de la quantification vectorielle (LVQ pour Learning Vector Quantization) est un algorithme très similaire à la méthode SOM à de nombreux aspects. Cependant, il est spécialement conçu pour la classification. On a les fonctions lvq1 et lvq3 qui implémentent deux versions de cet algorithme.

La fonction som_supervised est utilisée pour créer un classificateur pour l'ensemble de données de l'exemple, la figure 29 donne cette classification à travers la carte U-matrix.

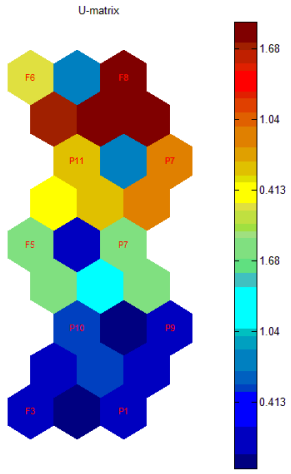


Fig.29 : - Clustering de tous les échantillons sur la U-matrix

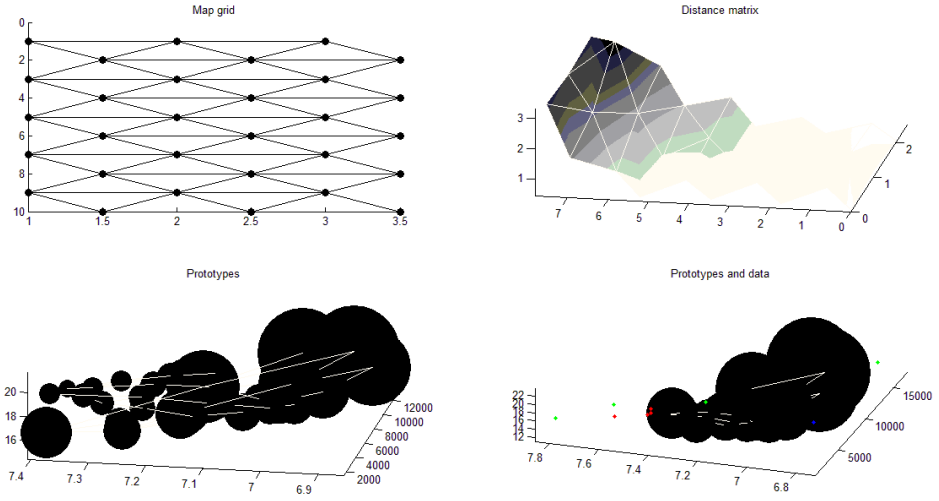


Fig. 30 : Matrice des distance, Prototype et prototype & données.

II.3. Conclusion :

Dans ce chapitre, nous avons décrit les différents organigrammes permettant d'utiliser la méthode SOM, en présentant les fonctions nécessaires aux calculs. Dans une première étape l'application de la méthode a été uniquement sur des données hypothétiques. Par la suite, nous avons appliqué la méthode sur l'exemple des données hydrochimiques de la nappe alluviale de la vallée de l'Oued M'Zi, afin de mieux maitriser l'outil SOM et tester la méthode sur un ensemble de données connues au préalable et tenter de nouvelles classifications et une première interprétation. Ceci, bien-entendu dans le but de vérifier l'efficacité du programme de calcul et sa validité. Le résultat obtenu montre l'efficacité et la flexibilité du code pour l'obtention des cartes auto-adaptatives pour une classification plus rationnelle des données.

Le chapitre 3 suivant, présente l'application de la méthode SOM sur les données hydrochimiques des eaux souterraines du Continental Intercalaire en l'Algérie.

Chapitre 3

Classification des eaux souterraines du Continental Intercalaire par la méthode SOM

III.1. Introduction

Occupant une superficie de plus de 600 000 km^2 en Algérie, le Continental Intercalaire est l'un des plus grands réservoirs d'eau souterraine de la région. La configuration lithologique et structurale de l'aquifère et le climat de la région font que les réserves d'eau souterraine de cet aquifère se renouvellent très peu : ce sont des réserves fossiles dont les exutoires naturels : sources et foggaras, ont permis le développement d'oasis où les modes de vie séculaires sont restés longtemps en parfaite symbiose avec l'écosystème saharien.

Cependant, depuis plusieurs années, l'exploitation par forages a sévèrement entamé cette réserve d'eau souterraine. Cette intensification de l'exploitation engendre un certain nombre de problèmes dont principalement la baisse régulière du niveau d'eau, l'augmentation du coût du pompage, l'affaiblissement de l'artésianisme, le tarissement des exutoires naturels et un risque de plus en plus grand de l'évolution hydrochimique des eaux par salinisation. A ce titre, on propose dans cette étude dans cette étude d'appréhender la caractérisation hydrochimique des eaux de la nappe à l'aide des cartes auto-organisatrices de Kohonen.

III.2. Présentation du la région d'étude :

III.2.1. Caractéristiques générales du Sahara :

Avec un ciel clair, quasiment sans nébulosité, le Sahara est avant tout "le pays du soleil". Il reçoit de 50 à 100 mm de pluie par an. Les vents desséchants érodent les roches jusqu'à les réduire en sable. Leurs effets se conjuguent avec ceux d'une insolation violente pour accroître l'intensité de l'évaporation (OULD BABA SY, 2005).

Sans eau en surface et sans terre arable, le désert saharien est également soumis à des alternances de froid et de chaud : l'amplitude diurne des températures, en effet, dépasse couramment 35°C, tandis que les hivers rigoureux succèdent aux étés torrides. Hormis les régions du cercle polaire et les hauts sommets enneigés, nul autre milieu ne semble aussi hostile à toute forme de vie (LEMIRE et al, 2003 ; OULD BABA SY, 2005).

a) - Précipitations :

La faiblesse de la pluviosité est le caractère fondamental du climat saharien (Fig.31). On note ainsi des précipitations annuelles très faibles dans certaines localités : 33 mm à Béni Abbès ; 13 mm à Adrar ; 10 mm à Ain Salah, en Algérie (Meddi, M et Meddi, H. 1998).

Toutefois, des pluies diluviennes peuvent aussi se produire au Sahara. En septembre 1950, Tamanrasset a reçu 44 mm en trois heures (Sahara, 2003) alors que sa moyenne annuelle est de 25 mm (Meddi, M et Meddi, H. 1998).

b) - Températures

Le climat thermique est assez uniforme ; les étés du Sahara septentrional ne sont donc guère moins torrides que ceux de la zone centrale ou même de la région soudanaise. Juin, juillet et août sont les mois les plus chauds des zones septentrionale et centrale. Mais vers le Sud, cette période estivale se trouve décalée, recouvrant avril, mai et juin. Juillet est, dans le premier cas, le mois le plus chaud avec, en année normale, une moyenne des maxima quotidiens comprise entre 40° et 46°, selon les localités. Les plus hautes températures ont été observées à Ain Salah avec 56° et à Tindouf avec 59°.

En hiver, il gèle presque partout. Les températures les plus basses enregistrées atteignaient -10° dans le Tibesti, -7° à Tamanrasset, -6° à Béchar et à Béni Abbès. Il existe donc de grands écarts de température entre l'hiver et l'été. L'amplitude des variations thermiques annuelles, qui est l'une des particularités du climat des déserts chauds, peut dépasser 55° au Sahara. En outre, la variation diurne, c'est-à-dire la différence entre le maximum diurne et le minimum nocturne, dépasse souvent 35° (OULD BABA SY, 2005).

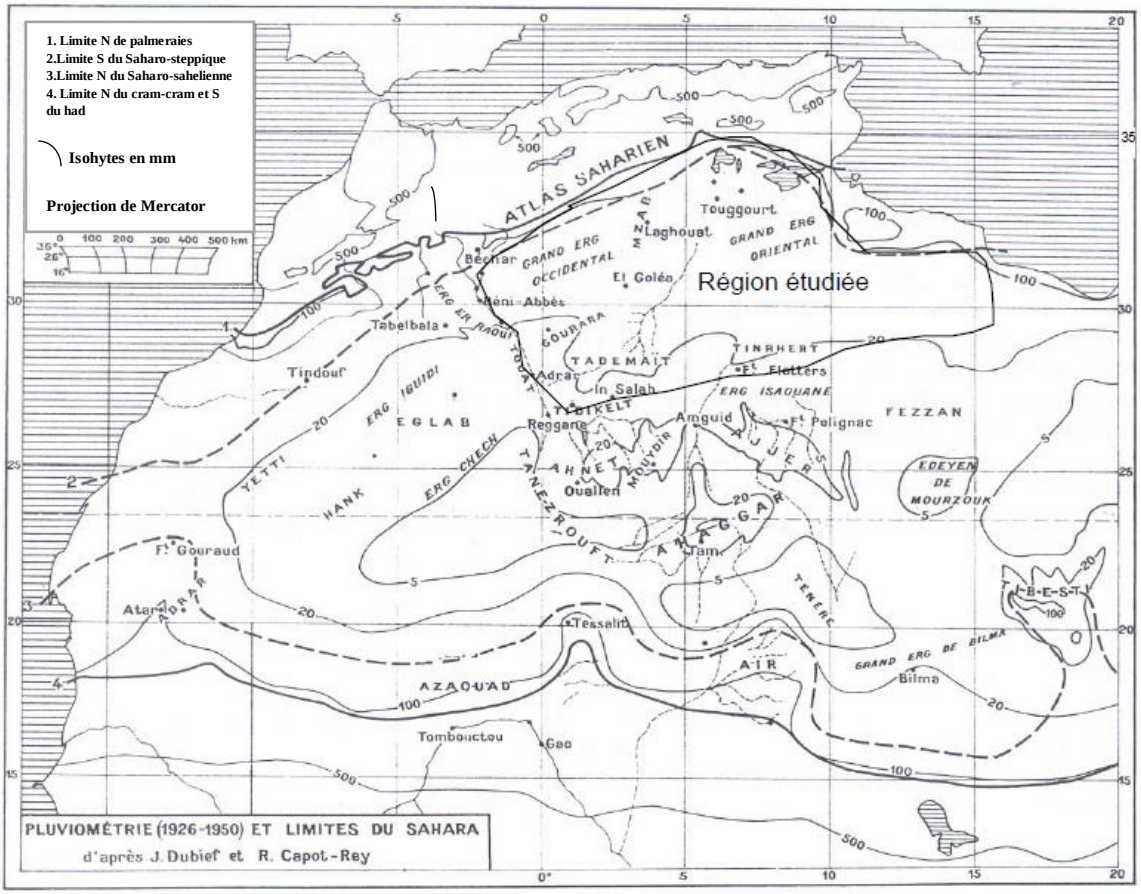


Fig. 31 : Pluviométrie et limites du Sahara (*in* CONRAD, 1969)

III.2.2. Système aquifère du Sahara :

Occupant une superficie de plus d'un million de km^2 (Fig. 32), le Système Aquifère du Sahara Septentrional (SASS), partagé par l'Algérie, la Tunisie et la Libye, est formée de dépôts continentaux renfermant deux grandes nappes souterraines : le Continental Intercalaire [CI] et le Complexe Terminal [CT]. Ce système représente des réserves considérables plus de 40000 milliards de mètre cube partagées en : 60% en Algérie ; 30% en Libye et 10% en Tunisie (Taibi, 2017). Dans cette étude, on s'intéresse uniquement au Continental Intercalaire de l'Algérie. Le terme du « Continental intercalaire » désigne un épisode continental localisé entre deux cycles sédimentaires marins :

- à la base, le cycle du Paléozoïque qui achève l'orogenèse hercynienne,
- au sommet, le cycle du Crétacé supérieur.

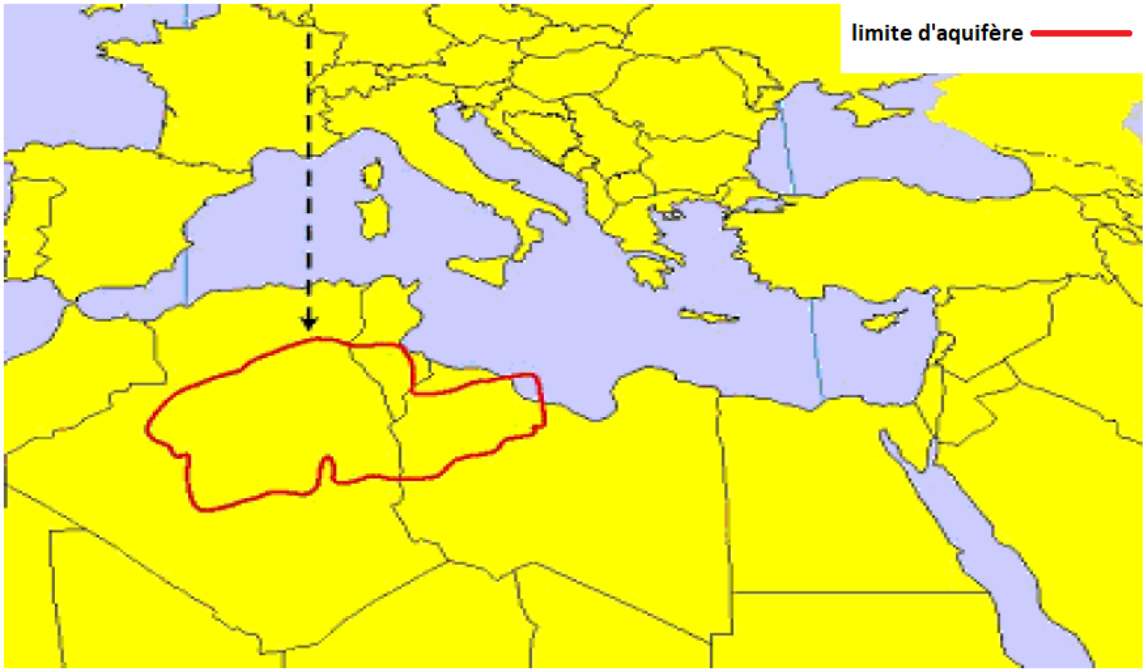


Fig.32 : Extension du Système Aquifère du Sahara Septentrional
(D'après Taibi, 2017)

Le Complexe Terminal est un ensemble assez peu homogène incluant des formations carbonatées du Crétacé supérieur et des épisodes détritiques du Tertiaire et principalement du Miocène (SASS ,2003). Les limites du CI et du CT sont représentés dans la figure 33.

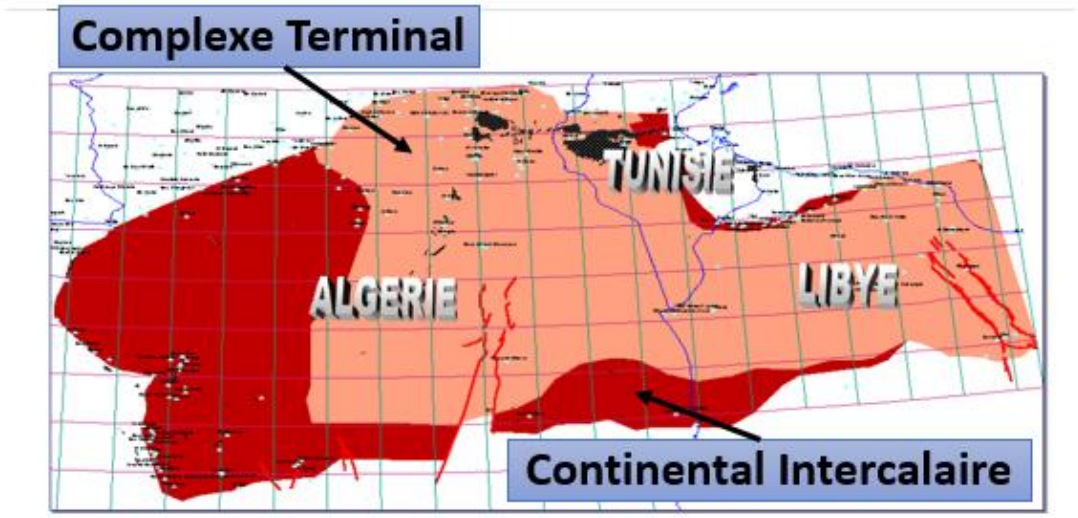


Fig.33 : Limites d'extension du CI et du CT
(D'après Taibi, 2017)

La figure 34 représente la carte géologique du Sahara. Elle montre l'extension des réservoirs CT et CI et leurs substratums.

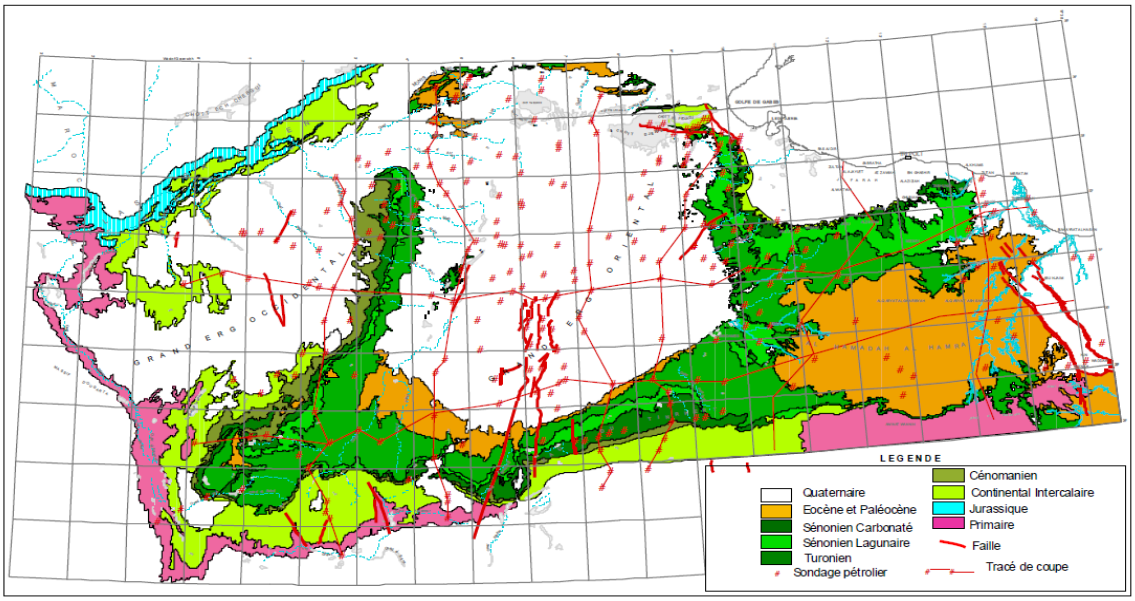


Fig. 34 : Carte géologique du Sahara
(D'après SASS, 2003)

Les prélèvements du CI, utilisés autant pour des fins agricoles que pour l'alimentation en eau potable et pour l'industrie, ont passés ces dernières années à 2,5 milliards de m^3/an à travers des forages d'eau dont le nombre a atteint aujourd'hui plus de 5000 forages (Fig. 35).

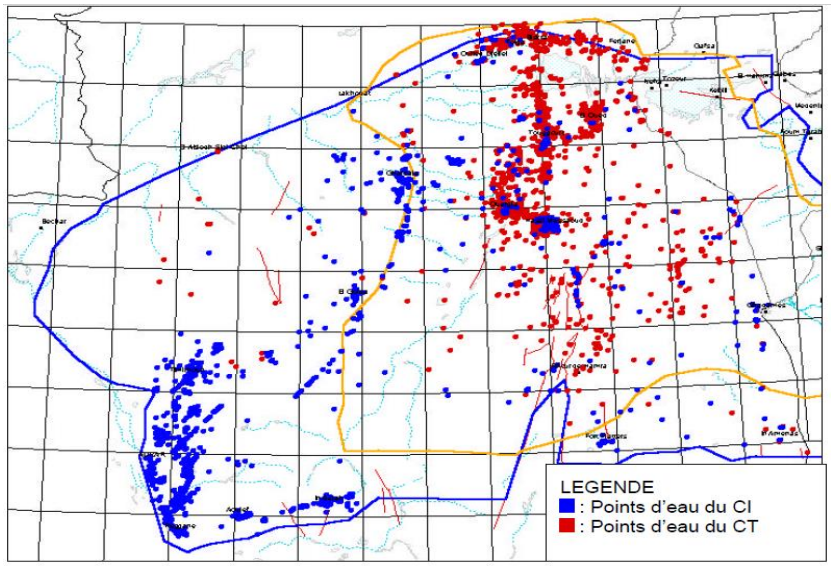


Fig. 35 : Répartition des points d'eau captant le CT et le CI
(D'après SASS, 2003)

III.3. Données hydrochimiques :

III.3.1. Données et statistiques :

Dans cette étude, on a utilisé les données Hydrochimiques dont les prélèvements ont été effectué pour la plupart par l'ANRH et par Mr. Dedjell pour les échantillons prélevés dans la Wilaya de Ghardaïa et Mr. Chettih pour les échantillons prélevés dans les Wilayas de Laghouat, Djelfa et Ouargla en 2015. Ainsi, 61 échantillons ont été prélevés et analysés (Annexe 1). Le tableau n°2 regroupe les caractéristiques statistiques des résultats d'analyses.

Eléments et paramètres	Min	Max	Moyenne	Ecart-type	CV %
Ca (mg/l)	22.00	433.87	182.8210	84.2772	46.0982
Mg (mg/l)	28.00	221.13	76.5682	32.8521	42.9056
Na (mg/l)	30.00	779.36	262.69	143.5644	54.6516
K (mg/l)	5.00	83.28	31.2125	22.0610	70.6801
Cl (mg/l)	110.0	1078.4	420.9697	168.6593	40.0645
SO_4 (mg/l)	10.00	1290.6	657.5779	266.9919	40.6023
HCO_3 (mg/l)	24.00	250	166.8591	39.0894	23.4266
NO_3 (mg/l)	0.00	108.61	3.3023	13.8670	419.9186
Ph	6.80	8.31	7.02	0.3388	4.8262
CE µSiemens	1144.7	5750	2175.8	951.7476	35.0454
T°	15.00	71	41.4	15.9722	38.5803

Tableau n°2 : Caractéristiques statistiques des propriétés physico-chimiques des échantillons d'eau de la nappe du Continental Intercalaire.

III.3.2. Diagrammes :

Afin d'interpréter les résultats d'analyses, on a utilisé différents diagrammes :

a) **Diagramme de Piper :** Le diagramme de Piper utilise les éléments majeurs pour représenter les différents faciès des eaux . Il permet également de voir l'évolution d'une eau, passant d'un faciès à un autre, grâce à des analyses espacées dans le temps ou des analyses d'échantillons pris à des endroits différents. En regroupant les échantillons par régions, le diagramme de Piper donne le résultat représenté par la figure ci-dessous.

   Le diagramme de Piper montre deux classes d'eau, une classe Chlorurée et Sulfatée Calcique et Magnésienne et une classe Chlorurée Sodique et Potassique et Sulfatée Sodique.

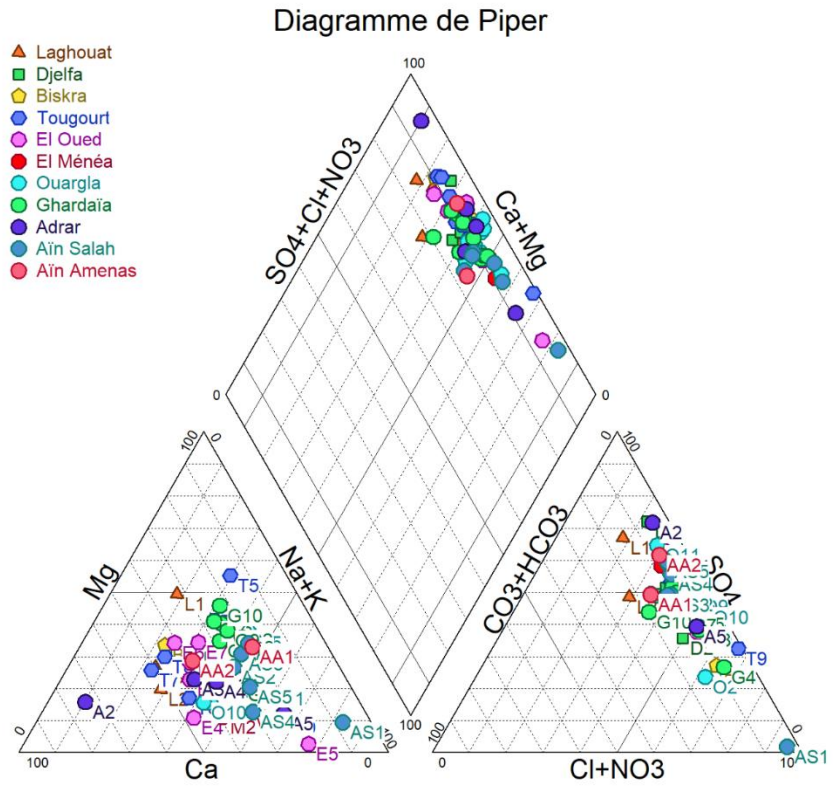


Fig. 36 : Diagramme de Piper des eaux du Continental Intercalaire

b) Diagramme de Schoeller-Berkaloff : Le diagramme de Schoeller permet entre autres de reconnaître simplement le faciès d'une eau souterraine, en utilisant les concentrations des éléments majeurs et en les reportant sur un graphique en colonnes à échelles logarithmiques.

Les résultats sont illustrés dans la figure 37 pour les quatorze premiers échantillons et pour le reste des échantillons, sont reportés en Annexes n°2. D'après les diagrammes de Schoeller-Berkaloff, les faciès chimiques sont Sulfatés Magnésiens et Chlorurés Sodiques pour certains et Sulfatés Sodiques et Chlorurés Magnésiens pour d'autres. Un autre groupe minoritaire présente un faciès Chloruré Sodique et Sulfaté Calcique des eaux de la région de Touggourt.

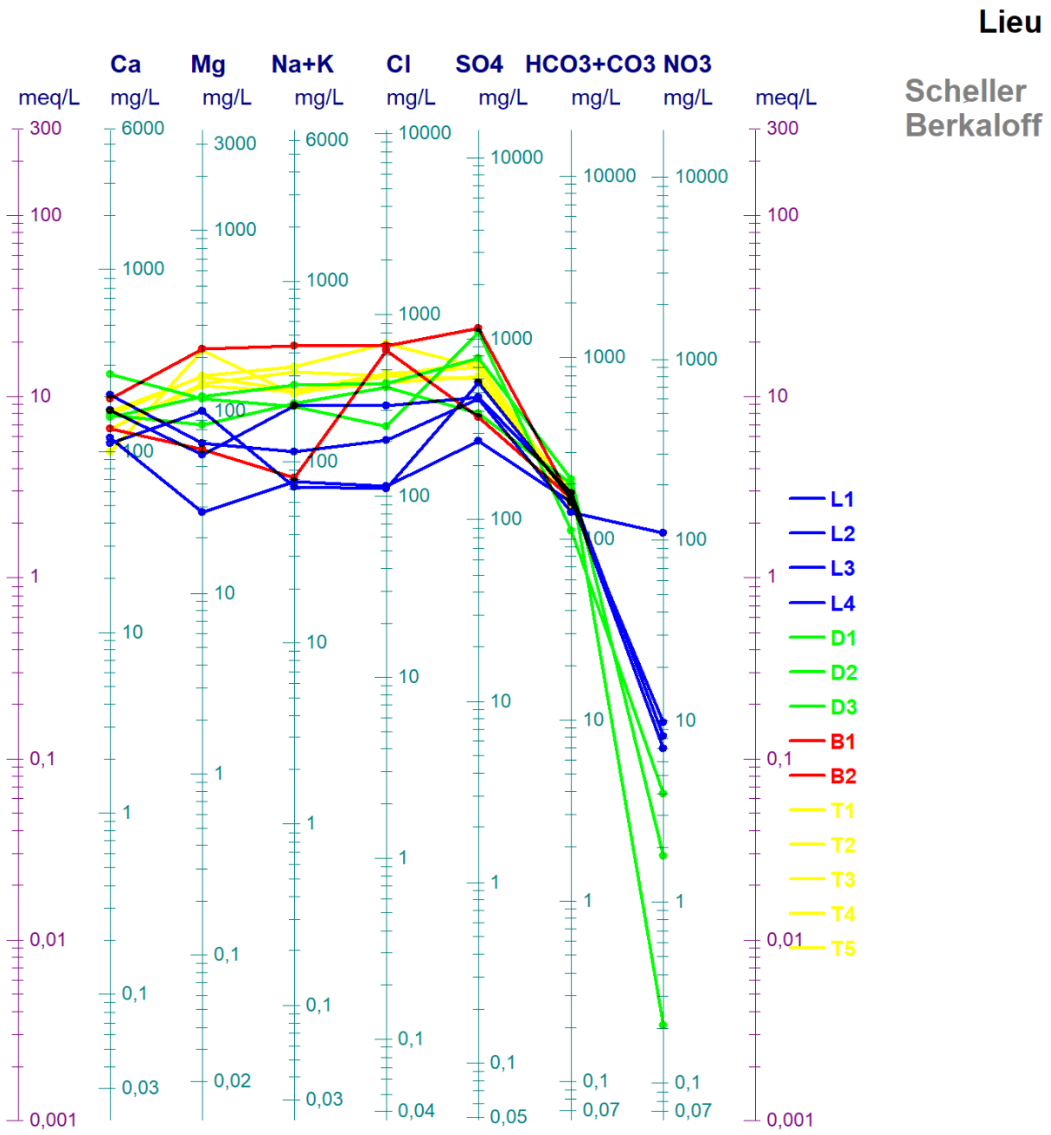


Fig. 37 : Diagramme de Schoeller-Berkaloff des eaux du CI

c) Diagramme de Wilcox : Ce diagramme est essentiellement utilisé pour évaluer le risque de salinisation des sols. Il utilise pour cela la conductivité électrique (CE) ou la charge totale dissoute, toutes deux relatives à la salinité de l'eau, et l'indice d'adsorption du sodium (SAR en anglais) aussi appelé "pouvoir alcalisant".

Le diagramme montre une très faible quantité d'échantillons classés dans la catégorie des eaux de bonne qualité, alors que la magorité des échantillons sont classés dans la catégorie médicre et mauvaise. Un seul échantillon A5, celui de la région d'Adrar est classé comme eau de qualité admissible.

Cette classification, montre bien que la magorité des échantillons ne sont pas recommandables pour l'irrigation à cause de leur fort risque d'alcalinisation des sols.

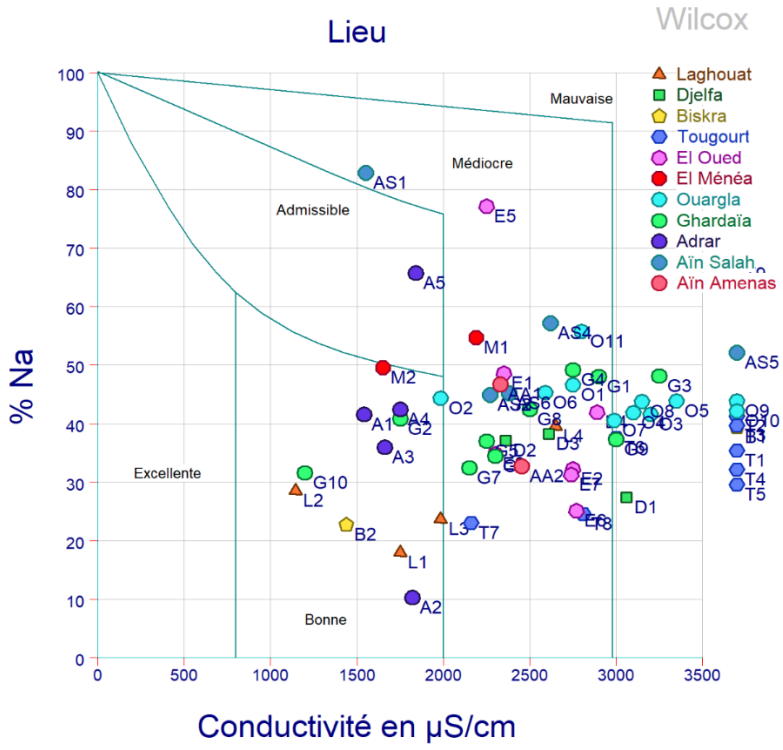


Fig. 38 : Diagramme de Wilcox des eaux du CI

III.3.3. Matrice de corrélation :

La matrice de corrélation est une matrice superposée qui représente les coefficients de corrélation entre les éléments chimiques. Le traitement statistique donne les résultats indiqués dans le tableau n°3.

Elements	Ca	Mg	Na	K	Cl	SO_4	HCO_3	NO_3	CE	Ph	T°
Ca	1										
Mg	0.07	1									
Na	0.38	0.04	1								
K	0.54	0.20	0.48	1							
Cl	0.34	0.15	0.42	0.51	1						
SO_4	0.70	0.25	0.52	0.42	0.38	1					
HCO_3	-0.12	0.33	-0.15	-0.07	-0.22	-0.01	1				
NO_3	0.03	-0.11	-0.16	-0.20	-0.21	-0.10	-0.03	1			
CE	0.68	0.30	0.77	0.66	0.74	0.82	-0.09	-0.18	1		
Ph	0.25	-0.05	-0.05	-0.07	-0.08	0.17	0.06	-0.18	0.04	1	
T°	0.35	0.32	0.26	0.69	0.39	0.26	0.08	-0.23	0.45	-0.08	1

Tableau n°3 : Matrice de corrélation

La matrice de corrélation montre une forte corrélation Ca-SO_4 qui pourrait être liée au gypse. La conductivité est assez bien corrélée aux éléments majeurs Ca, Na, Cl, SO_4, ceci est liée à la forte minéralisation des eaux. Les différents éléments chimiques semblent être uniquement d'origine naturelle.

III.4. Analyse des données par la méthode SOM :

III.4.1. Création des données :

Comme il a été décrit dans le chapitre 2, on commence tous d'abord la création et la structuration de la base de données.

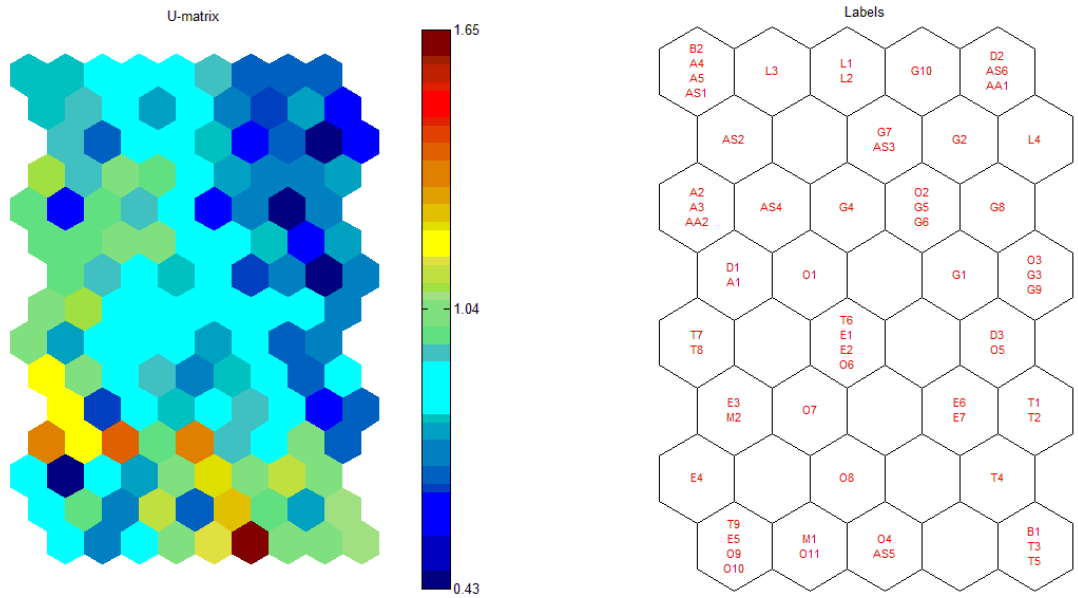
Après la construction et normalisation des données de l'annexe 1, le programme détermine la taille de la carte SOM automatiquement et l'erreur dans cette étape d'apprentissage. Pour l'ensemble des données, l'erreur de quantification finale q_e est de 1.861 qui représente la distance moyenne entre chaque vecteur de données et son BMU. Cette valeur semble faible et acceptable.

Cependant, l'erreur topographique qui représente la proportion de tous les vecteurs de données pour lesquels la première valeur et la deuxième valeur du BMU ne sont pas des unités adjacentes a été calculée à l'aide de la fonction som_quality, et la valeur de l'erreur topographique finale calculée est nulle ($t_e = 0.000$). Ceci montre que la carte auto-organisatrice est parfaite pour représenter l'ensemble des données.

III.4.1. Création des données :

III.4.2. Visualisation des cartes auto-organisatrices

Le résultat obtenu après visualisation est représenté par l'ensemble des cartes suivantes. La figure 39 (a), visualise la carte U-matrix qui représente les distances euclidiennes entre deux neurones voisins dans la carte auto-organisatrice. Une couleur rouge indique une zone frontalière à forte valeur. Une couleur bleue indique que les distance entre les neurones sont faibles et caractérise une zone homogène dans la carte c'est-à-dire un ensemble de neurones peu différents. En général, les zones frontalières coïncident avec les unités vides de la carte.



(a) (b)

Fig. 39 : a) – Carte des distances (U-Matrix), b) – Carte des individus

La figure 39 (b), visualise la carte des individus, où on liste sur la carte les observations classées par le neurone figuré par un hexagone. La figure 39 (b), visualise la carte des individus, où on liste sur la carte les observations classées par le neurone figuré par un hexagone. La carte peut être vue comme une extension non linéaire de l'Analyse en Composantes Principales (ACP). Pour l'exemple étudié, on constate une répartition plus uniforme des échantillons sur la carte de Kohonen.

La visualisation illustrée par la figure 40, montre que la taille de l'hexagone est proportionnelle au nombre d'individus classés par le neurone.

La visualisation ci-dessous (Fig. 41), montre l'intérêt de la carte des variables (physico-chimiques). Chaque carte est représentative de la variable correspondante, cela permet d'étudier les corrélations entre les variables au sens de la carte des individus.

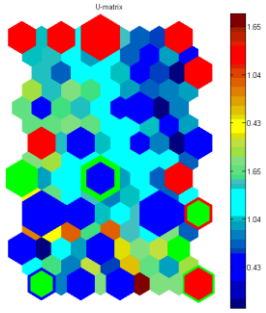


Fig. 40 : Carte U-matrix et nombre d'individus classés.

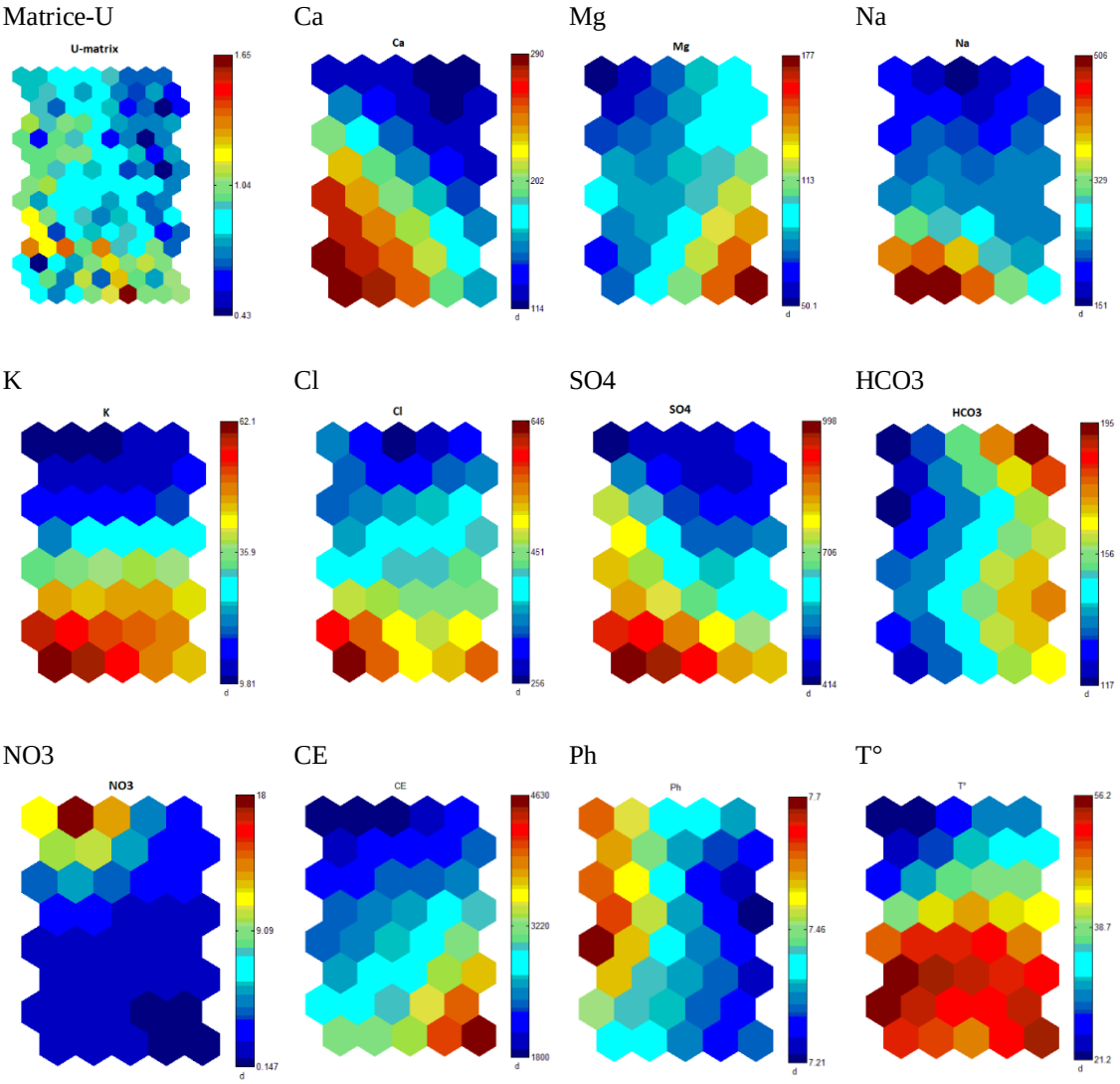


Fig. 41 : - Visualisation de la carte SOM :
U-matrix, Ca, Mg, Na, K, Cl, SO_4, HCO_3, NO_3, Conductivité, Ph, et T°.

On constate sur la figure 41 que pour la majorité des éléments majeurs, leurs cartes sont homogènes et régulières et évoluent d'une façon significative, ceci peut indiquer déjà une même origine naturelle. On constate que le Calcium et le Sodium se comportent d'une façon analogue, alors que le Potassium, le Chlore et le Sulfate se comportent eux semblablement. Par contre le bicarbonate se comporte différemment. On constate nettement la bonne corrélation entre la conductivité et le Magnésium, ceci montre que c'est le Magnésium responsable de la forte dureté des eaux qui contrôle la conductivité électrique de l'eau. Cependant, le Nitrate se présente sur la carte d'une façon complètement différente des autres éléments ce qui prouve que sa provenance est d'origine anthropique. Quant à la température, elle apparait liée à la profondeur des forages, elle se corrèle bien avec la profondeur, d'où l'origine géothermale. On constate également un comportement contradictoire entre le Ph et le Bicarbonate, ceci est dû à l'environnement acide des eaux. Les associations minérales les plus importantes sont l'Halite, Gypse, Dolomite et Sylvite.

Une autre visualisation également importante car elle permet de dégager déjà les principales calasses, qui est celle de la projection des composants principaux réalisés pour les données et appliquée à la carte (Fig.42). Là aussi, la palette de couleurs se fait en étalant une palette de couleurs sur la projection. Elle montre déjà une classification primaire par groupement d'individus.

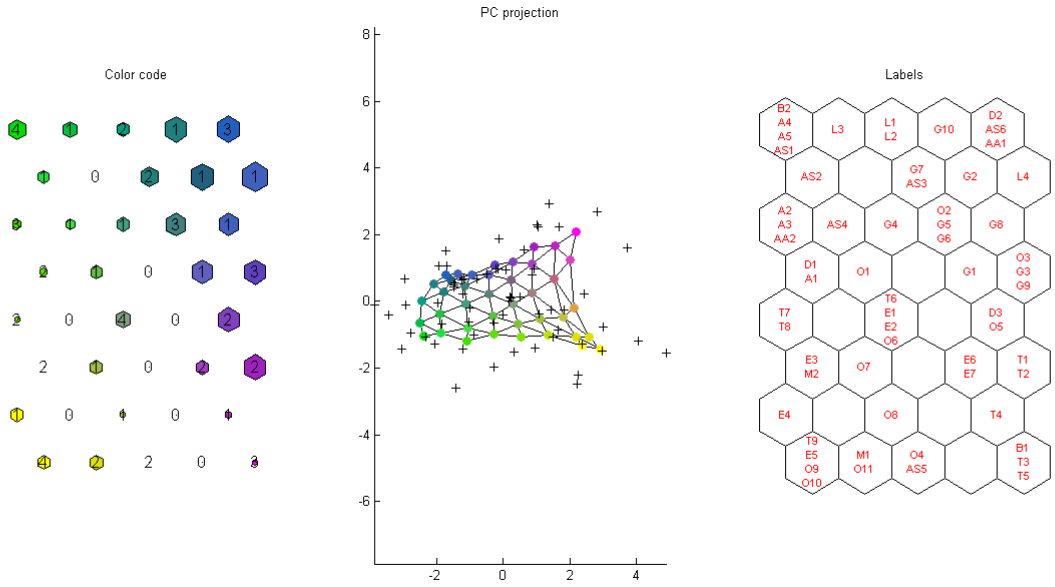
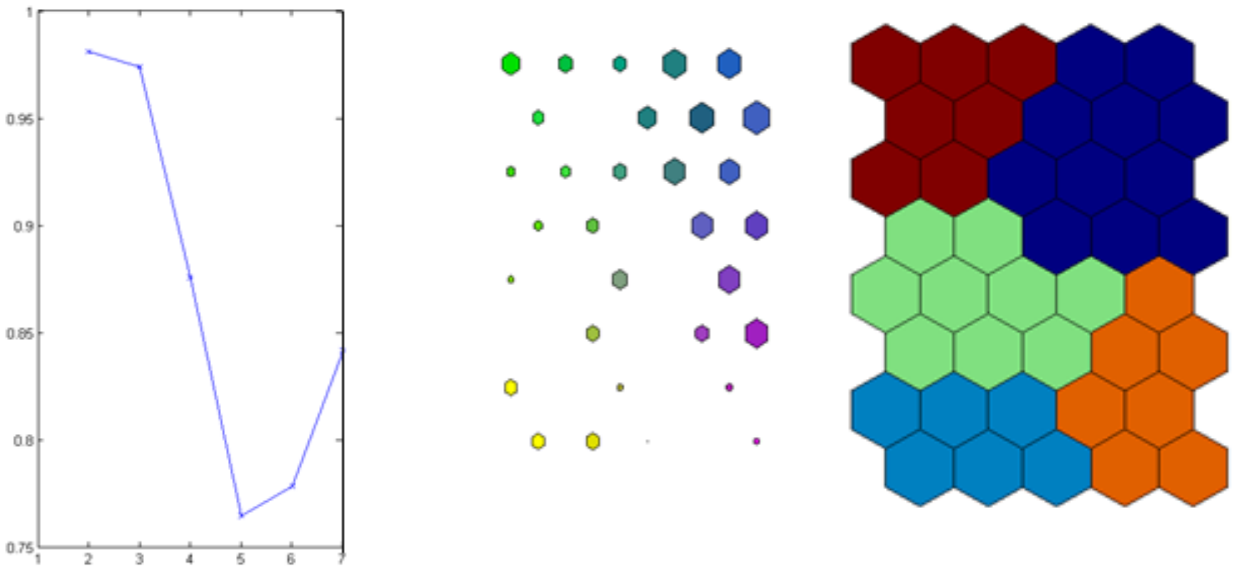


Fig. 42 : - Code couleur, classement et nombre de Hits, Projection PC et Carte des individus

Cette classification primaire basée uniquement sur l'inspection visuelle a déjà laissé entendre qu'il y a au moins deux ou trois clusters dans les données, et que les propriétés des clusters sont différentes les unes des autres.

A ce titre, nous avons eu recours à la fonction kmeans_clusters qui permet de retrouver un partitionnement initial en se basant sur l'indice de clustering de Davies-Boulding qui est minimisé avec le meilleur clustering. La figure 43 montre que pour un nombre de partitionnement égal jusqu'à 7, l'indices de Davies-Boulding est minimum pour un nombre de classe égal à 5 (Fig. 43 (a)). On constate que la carte du code de couleur (Fig. 43 (b)) est partitionnée en cinq classes (Fig. 43 (c)).



(a) (b) (c)

Fig. 43 : - (a) – Indice de Davies-Boulding, (b) – Carte de code de couleur, (c) – Carte de classification par groupement d'individus.

Une autre visualisation, la plus informative de toutes est celle qui présente les diagrammes de dispersion simples et les histogrammes de toutes les variables (Fig.44). Cette figure montre une partie des données traitées pour illustration. Les points de données d'origine se trouvent dans le triangle supérieur, les valeurs du prototype de carte dans le triangle inférieur et les histogrammes sur la diagonale : noir pour l'ensemble de données et rouge pour les valeurs de prototype de carte. Le codage couleur des échantillons de données a été copié à partir de la carte (à partir du BMU de chaque échantillon). On note cependant là aussi, que les valeurs des variables ont été dénormalisées. Nous constatons sur les valeurs du prototype les relations existants entre les variables d'une façon plus nette et plus convaincante. La figure 45 donne les diagrammes de l'ensemble des données traitées.

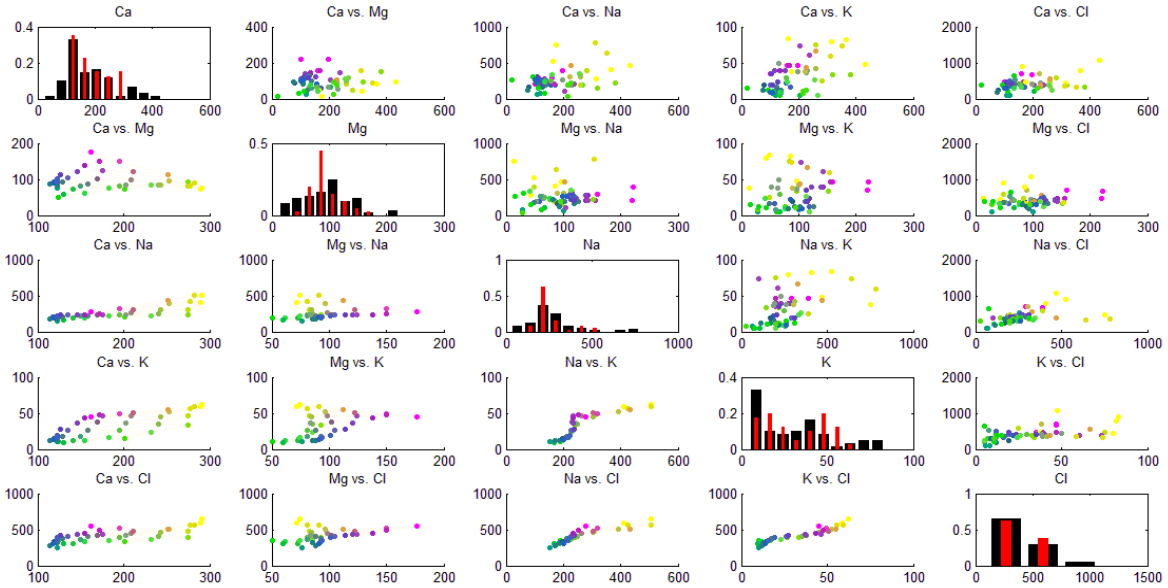


Fig.44 : - Valeurs du prototype de la carte, points de données d'origine et Histogrammes sur la diagonale.

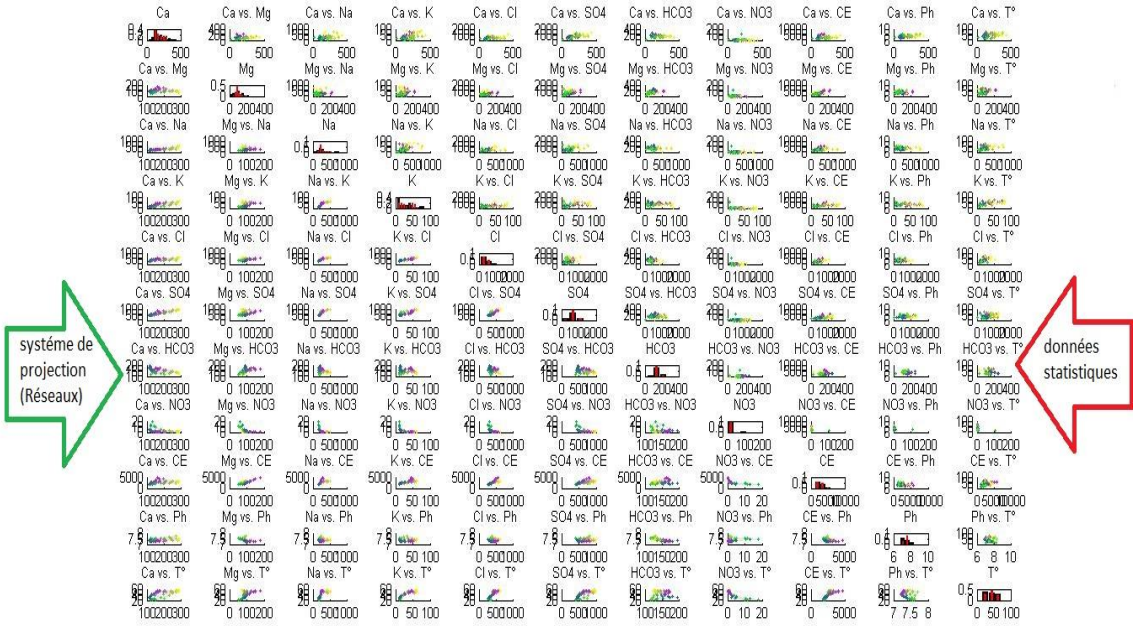


Fig. 45 : - Valeurs du prototype de la carte, points de données d'origine et Histogrammes sur la diagonale de l'ensemble des données hydrochimiques traitées.

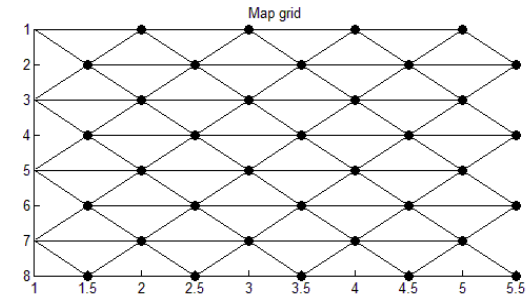
Dans ce travail, nous avons utilisé une autre fonction de visualisation : *som_grid*. Elle permet de visualiser la carte SOM dans des coordonnées librement spécifiées, par exemple l'espace d'entrée (bien sûr, uniquement jusqu'à l'espace 3D). Cette fonction a beaucoup d'options et est assez flexible.

On présentera ci-dessous (Fig. 46) quatre visualisations réalisées avec cette fonction.

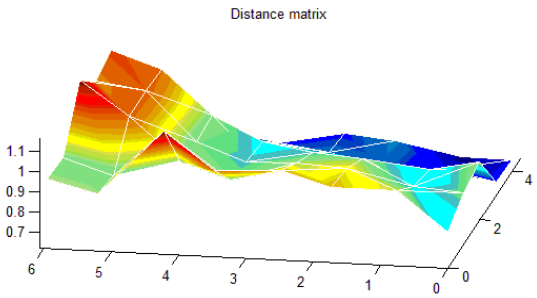
1. La carte dans l'espace de sortie (Fig. 46 (a)).
2. Le tracé de surface de la matrice de distance (Fig. 46 (b)): à la fois couleur et la coordonnée z indiquent la distance moyenne par rapport aux unités cartographiques voisines. Ceci est étroitement lié à la matrice U.
3. La carte dans l'espace de sortie (Fig. 46 (c)) où les trois premiers composants déterminent les coordonnées 3D de l'unité cartographique et la taille du marqueur est déterminé par le quatrième composant. Notez que les valeurs ont été dénormalisées.
4. La carte comme ci-dessus (Fig. 46 (d)), mais les données d'origine ont également été tracées : les coordonnées indiquent les valeurs des trois premiers composants et la couleur indique l'espèce de chaque échantillon. Les autres composants ne sont représentés.

Afin de construire des modèles sur la base de la méthode SOM avec les données hydtochimiques traitées précédemment, nous avons utilisés les fonctions décrites dans le chapitre 2 : la fonction som_estimate_gmm et la fonction som_probability_gmm. Ces modèles sont de simples modèles locaux ou du plus proche voisin. Ces fonctions ont été utilisées pour l'estimation de la densité de probabilité. Chaque prototype de carte est le centre d'un noyau gaussien dont les paramètres sont estimés à partir des données. Le modèle de mélange gaussien est estimé avec la fonction som_estimate et les probabilités peuvent être calculées avec som_probability.

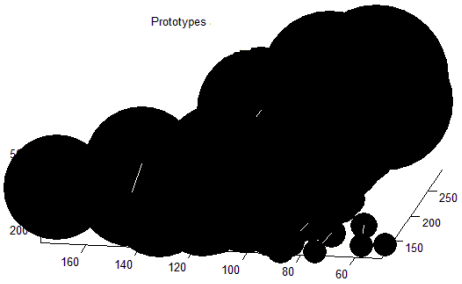
La figure 47 présente les cartes de densité de probabilité de chaque paramètre.



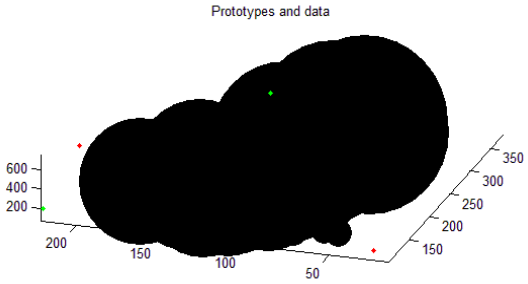
(a)



(b)



(c)



(d)

Fig. 46 : - (a) – Carte dans l'espace de sortie, (b) - Tracé de surface de la matrice de distance, (c) - Carte dans l'espace de sortie pour les trois premiers composants, (d) - Carte dans l'espace de sortie avec les données d'origine.

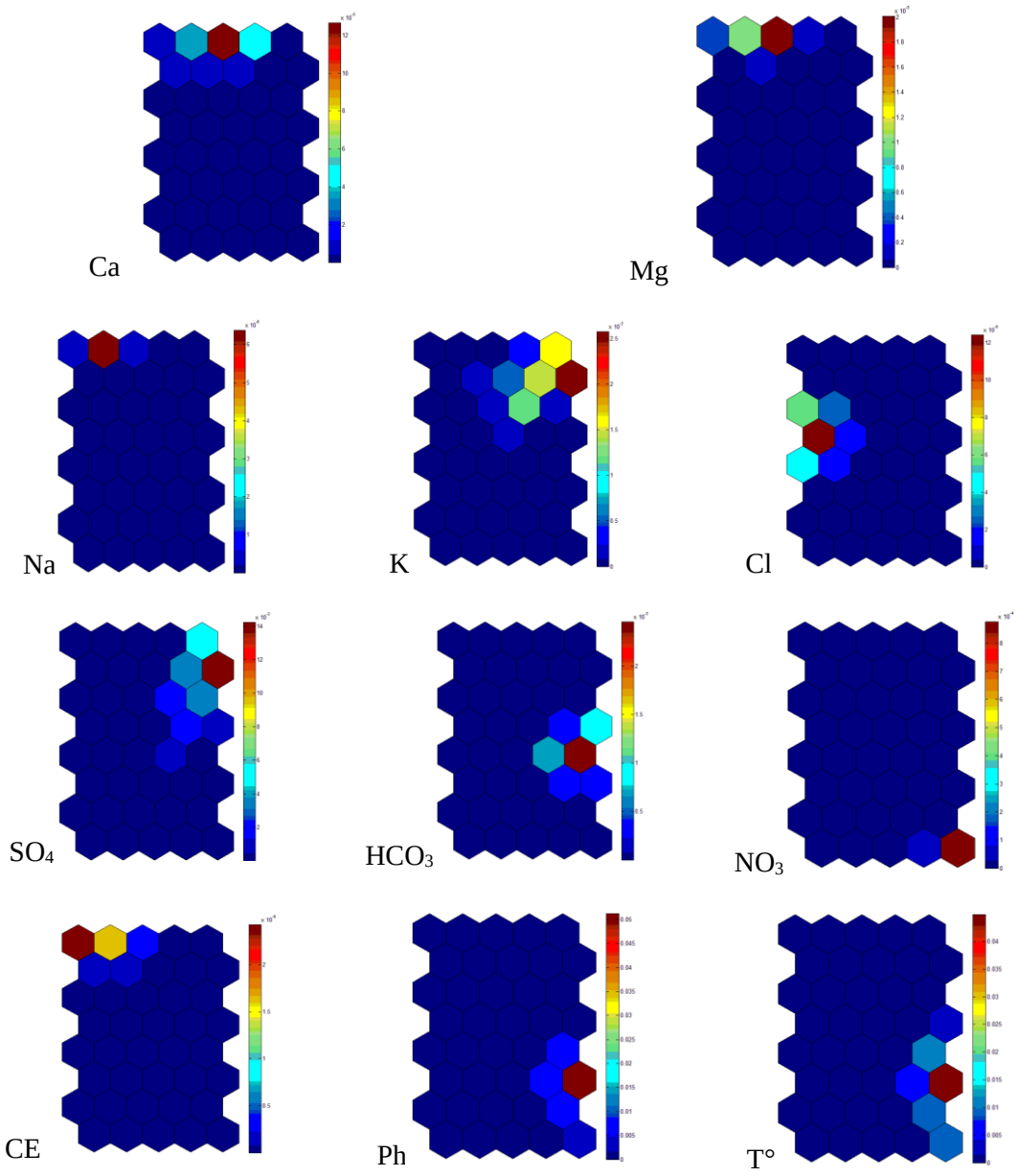


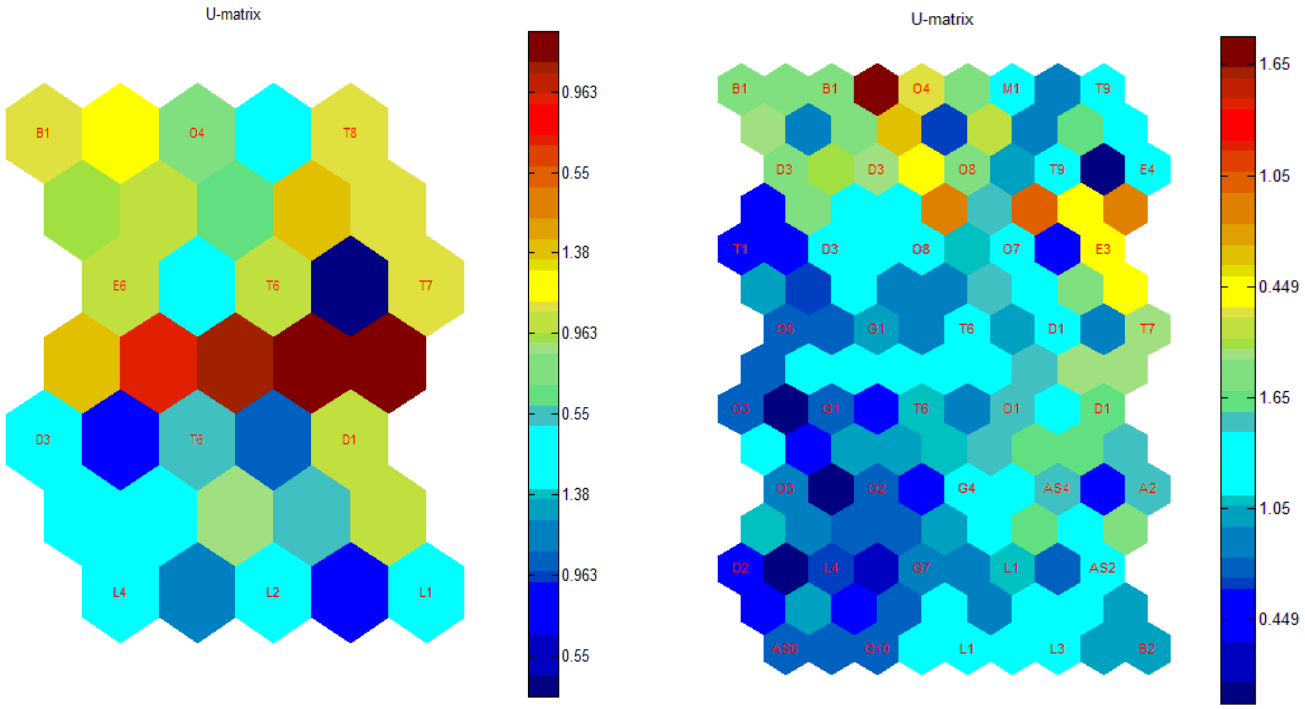
Fig. 47 : - Cartes de Densité de Probabilité des paramètres physico-chimiques des eaux du Continental Intercalaire.

La méthode SOM a été utilisé dans ce travail pour la classification des eaux du Continental Intercalaire, bien que cette méthode puisse être utilisé pour la classification, il faut se rappeler qu'elle n'utilise pas toutes les informations de la classe, et donc ses résultats sont intrinsèquement sous-optimaux. Cependant, avec de petites modifications, le réseau peut prendre en compte les informations pertinentes de la classe. La fonction som_supervised

fait cela. Cependant, la quantification vectorielle d'apprentissage (LVQ) est un algorithme très similaire au SOM à bien des égards. Cependant, il est spécialement conçu pour la classification. Dans la SOM Toolbox, il existe des fonctions LVQ1 et LVQ3 qui implémentent deux versions de cet algorithme.

En faisant appelle à cette fonction, nous avons construit les données d'apprentissage en ajoutant une matrice codée aux données d'origine en fonction des informations de classe dans le champ des individus. La dimension des vecteurs après le processus est (l'ancienne dimension + un nombre de classes différentes). Dans chaque vecteur, l'un des nouveaux composants a la valeur '1' (cela dépend de la classe du vecteur), et les autres '0'.

On appelle la fonction som_make pour construire la carte. Ensuite, la classe de chaque unité cartographique est déterminée en prenant au maximum ces composants ajoutés, et une étiquette (identifiant d'individu) est attribuée en conséquence. Enfin, les composants supplémentaires sont supprimés. Le résultat obtenu est illustré par la figure ci-dessous (Fig. 48). La figure ci-dessous donne la matrice finale des distances en visualisation « small » avec quelques individus représentatifs des groupements (Fig.48 (a)), et la matrice finale des distances en visualisation Normale avec les individus.



(a) **(b)**

Fig. 48 : - Les matrices des distances euclidiennes (U-Matrix) après modification à l'aide de la fonction som_supervised

III.5. Conclusion :

En hydrochimie, pour classer et caractériser les eaux souterraines, plusieurs méthodes sont utilisées pour la mesure de la qualité de l'eau. Ces méthodes de classification classiques deviennent insuffisantes lorsque les eaux ont une très forte variabilité, c'est le cas des eaux souterraines du Continental Intercalaire. La méthode SOM utilisée dans cette étude pour classer les eaux du CI constitue une nouvelle approche utilisant le concept des réseaux de neurones pour cette tâche. Les résultats obtenus, ont permis de classer ces eaux en cinq classes en mettant en évidence leur origine. Elle a permis également de mettre en évidence les associations géochimiques qui expliquent leur chimisme.

Conclusion Générale

Au terme de cette étude, et après avoir donné dans un premier temps, un aperçu général sur les méthodes usuelles de la classification automatique, nous avons détaillé plus longuement les cartes auto-adaptives. Dans ce contexte, nous avons évoqué les principes de la méthode SOM avec les formules mathématiques en mode séquentielle et différé où nous avons résumé la méthode par un organigramme général.

Par la suite, nous avons décrit les différentes étapes de calcul de l'algorithme de la méthode et nous avons donné une description détaillée des fonctions associées au programme. Nous avons également testé l'efficacité du code de calcul, sa validité, sa flexibilité à travers les données hydrochimiques de la nappe alluviale de la vallée de l'Oued M'Zi où nous avons pu dégager les principales classes des eaux de la nappe.

Dans la dernière étape, nous avons appliqué la méthode SOM aux données hydrochimiques des eaux souterraines du Continental Intercalaire du Sahara. Pour l'ensemble de ces données, l'erreur de quantification finale est de 1.861, une valeur qui semble relativement faible et acceptable, et l'erreur topographique finale calculée est nulle. Par ailleurs, l'indice de clustering de Davies-Boulding optimal a permis de partitionner la carte auto-organisée en cinq classes. Les principaux résultats exploitables obtenus à l'aide de la méthode SOM peuvent être résumé comme suit:

- Les cartes de la majorité des éléments majeurs sont homogènes et régulières et évoluent d'une façon significative, ceci peut indiquer déjà une même origine naturelle.
- Le calcium et le sodium se comportent d'une façon analogue, alors que le potassium, le chlore et le sulfate se comportent eux semblablement.
- Une bonne corrélation entre la conductivité et le magnésium, ceci montre que c'est le magnésium responsable de la forte dureté des eaux qui contrôle la conductivité électrique de l'eau.
- Le nitrate se présente sur la carte d'une façon complètement différente des autres éléments ce qui prouve que sa provenance est d'origine anthropique.
- La température apparait liée à la profondeur des forages, elle se corrèle bien avec la profondeur, d'où l'origine géothermale.
- Un comportement contradictoire entre le Ph et le bicarbonate, ceci est dû à l'environnement acide des eaux.
- Les associations minérales les plus importantes sont l'Halite, Calcite, Gypse, Dolomite et Sylvite.

Afin de construire des modèles sur la base de la méthode SOM avec les données hydrochimiques traitées, nous avons utilisés d'autres fonctions pour l'estimation de la densité de probabilité et le modèle de mélange gaussien.

Ces modèles sont de simples modèles locaux ou du plus proche voisin où chaque prototype de carte est le centre d'un noyau gaussien dont les paramètres sont estimés à partir des données.

En introduisant la quantification vectorielle et en faisant de petites modifications, le réseau a pu prendre en compte les informations pertinentes de la classe et par la suite construire la carte. Le modèle donne cependant, une carte U-matrix finale des distances avec les individus représentatives des groupements.

Les perspectives de ce travail sont quasiment inépuisables, dans la mesure où la méthode self-organizing map constitue un outil puissant qu'il faut diffuser largement afin de le confronter à d'autres méthodes de classification. Il serait également intéressant d'accompagner les analyses hydrodynamiques par les résultats obtenus par la méthode SOM afin de mieux interpréter le fonctionnement des systèmes hydrogéologiques.

Références bibliooographique:

Ait-Aoudia, S. (1994). *Modélisation géométrique par contraintes: quelques méthodes de résolution* (Doctoral dissertation, Ecole Nationale Supérieure des Mines de Saint-Etienne; Université Jean Monnet-Saint-Etienne).

Anifah, L., Purnama, I. K. E., Hariadi, M., & Purnomo, M. H. (2013). Osteoarthritis classification using self organizing map based on gabor kernel and contrast-limited adaptive histogram equalization. *The open biomedical engineering journal*, *7*, 18.

Augustson, J. G., & Minker, J. (1970). An analysis of some graph theoretical cluster techniques. *Journal of the ACM (JACM)*, *17*(4), 571-588.

Bação, F., Lobo, V., & Painho, M. (2005, May). Self-organizing maps as substitutes for k-means clustering. In *International Conference on Computational Science* (pp. 476-483). Springer, Berlin, Heidelberg

Bauer, H. U., & Pawelzik, K. R. (1992). Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on neural networks*, *3*(4), 570-579.

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. *Journa of machine learning research*, *2*(Dec), 125-137.

Berkhin, P. (2002). Survey of clustering data mining techniques. Accrue Software. *Inc. TR, San Jose, USA*.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (pp. 144-152).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cheng, Y. (1997). Convergence and ordering of Kohonen's batch map. *Neural Computation*, *9*(8), 1667-1676.

Cherel, J. P. (2010). Traitement d'images-classification d'images de télédétection. *Support de cours*.

Clark, S. (2018). *Advances in self-organizing maps for spatiotemporal and nonlinear systems* (Doctoral dissertation, University of New South Wales, Sydney, Australia).

Conrad, G. (1969). *L'évolution continentale post-hercynienne du Sahara algérien (Saoura, erg Chech-Tanezrouft, Ahnet-Mouydir)* (No. 10). Éditions du Centre national de la recherche scientifique

Cornuejols, A., & Miclet, L. (2002). L'apprentissage artificiel: Methodes et concepts. *Paris: Eyrolles*.

Danho, D. (2016). Modèle de mélange et classification. Daphine université Paris sous la direction de madame Angelina, R.

Dong, G., & Xie, M. (2005). Color clustering and learning for image segmentation based on neural networks. *IEEE transactions on neural networks*, *16*(4), 925-936.

Dreyfus, H. L. (2002). Intelligence without representation–Merleau-Ponty's critique of mental representation The relevance of phenomenology to scientific explanation. *Phenomenology and the cognitive sciences*, *1*(4), 367-383.[2222222222222222]

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, *21*, 768-769.

Hajjar, C. (2014). *Cartes auto-organisatrices pour la classification de données symboliques mixtes, de données de type intervalle et de données discrétisées* (Doctoral dissertation, Supélec).

Hebb, D. O. (1949). The organization of behavior; a neuropsycholocigal theory. *A Wiley Book in Clinical Psychology*, *62*, 78.

Heskes, T. (1999). Energy functions for self-organizing maps. In *Kohonen maps* (pp. 303-315). Elsevier Science BV.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

Kiviluoto, K. (1996, June). Topology preservation in self-organizing maps. In *Proceedings of International Conference on Neural Networks (ICNN'96)* (Vol. 1, pp. 294-299). IEEE.

Kohonen, T. (1972). Correlation matrix memories. *IEEE transactions on computers*, *100*(4), 353-359.

Kohonen, T. (1984). Self-Organization and Associative Memory. *Springer Series in Information Sciences*, *8*.

Kohonen, T. (1998). The self-organizing map. Neurocomputing 21, 1e6.

Kohonen T. (2001): Self-Organizing Maps, Third Edition. Springer Science & Business Media.

Labroche, N. (2012). *Méthodes d'apprentissage automatique pour l'analyse des interactions utilisateurs* (Doctoral dissertation, Université Pierre et Marie Curie, Paris 6).

Le Cun, Y. (1986). Learning process in an asymmetric threshold network. In *Disordered systems and biological organization* (pp. 233-240). Springer, Berlin, Heidelberg.

Lemaire, V. (2006). Cartes auto-organisatrices pour l'analyse de données. *RIAS2006*.

Lemire, M., Meurgues, G., & Petter, F. (2003): Désert saharien. Muséum de l'histoire naturelle.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing* (Vol. 2, pp. 20-21). Cambridge, MA: MIT press.

Meddi, M., & Meddi, H. (1998). Etude des pluies annuelles et journalières dans le Sahara algérien. *Science et changements planétaires/Sécheresse*, *9*(3), 193-199.

Minsky, M., & Papert, S. (1969). An introduction to computational geometry. *Cambridge tiass., HIT*.

Observation du Sahara et du sahel (OSS). (2002). Système aquifère du Sahara septentrional. *Hydrologie*, volume II, 322.

Ould Baba Sy, M. (2005). Recharge et paléo recharge du système aquifère du Sahara septentrional. *Faculte des Sciences de Tunis, Departement de Geologie*, *271*.

Parker, D. B. (1985). Learnins logic. *Technical Report*.

Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements*, *146*(2), 04020022.

Raghavan, V. V., & Birchard, K. (1979, September). A clustering strategy based on a formalism of the reproductive process in natural systems. In *Proceedings of the 2nd annual international ACM SIGIR conference on Information storage and retrieval: information implications into the eighties* (pp. 10-22).

Ritter, H. J., & Schulten, K. (1988, July). Kohonen's self-organizing maps: exploring their computational capabilities. In *ICNN* (pp. 109-116).

Ritter, H., Martinetz, T., & Schulten, K. (1992). *Neural computation and self-organizing maps: an introduction* (pp. 141-161). Reading, MA: Addison-Wesley.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rosenblatt, F. (1962). Principles of Neurodynamics (New York: Spartan).

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Soubiran, T. (2016) Cartes auto-organisatrices de Kohonen et typologisation de territoires Ceraps - UMR 8026 du CNRS Semin-R, 10 juin 2016.

Taibi, R. (2017). Le Système aquifère transfrontalier du Sahara septentrional

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions technip.

Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, *316*(5), 533-534.

Ultsch, A. (1990). Kohonen's self organizing feature maps for exploratory data analysis. *Proc. INNC90*, 305-308. Alfred Ultsch, H. Peter Siemon, Bernard Widrow (éditeur) et Bernard Angeniol (éditeur)

Victorri, B. (2006). Le connexionnisme. *Traité de neuropsychologie clinique. Bruxelles: De Boeck, coll.«Neurosciences et Cognition»,* à paraître – Version préliminaire.

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

Wang, W., Yang, J., & Muntz, R. (1997, August). STING: A statistical information grid approach to spatial data mining. In *VLDB* (Vol. 97, pp. 186-195).

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130

Zrehen, S., & Blayo, F. (1992). A geometric organization measure for Kohonen's map. In *Neuro-Nîmes 92: neural networks & their applications, Nîmes, November 2-6, 1992* (pp. 603-610).

Webographie:

https://eric.univ-lyon2.fr Classification Supervisée Julien JACQUES 26/09/2018

http://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html Cours Cnam RCP209

https://dataanalyticspost.com/Lexique/svm (2020) DAP

https://analyticsinsights.io/apprentissage-supervise-vs-non-sup Zakariyaa Ismaili

http://larmarange.github.io/analyse-R/classification-ascendante-hierarchique.html

http://eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles.pdf

Géographie du Sahara. http://membres .lycos.fr/fgeo2.htm ? SAHARA (2003)

Annexes

Annexe n°1: tableau des analyses chimiques

Régions	ident	Ca^{2+}	Mg^{2+}	Na^{+}	K^{+}	Cl^{-}	$SO4^{-}$	HCO3	$NO3^{-}$	CE	PH	T°
Laghouat	L1	110	101	65	8	110	575	165	7	1750	7.9	19
	L2	119.01	27.63	71.98	5.68	113.21	270.49	158.6	9.86	1144.7	7.02	19.5
	L3	203.13	66.37	108.33	5.27	202.71	470.04	139.64	108.61	1984.3	7.00	20.0
	L4	169.9	57.51	189.37	16.4	316.21	471.67	178.86	8.26	2647.6	7.15	31
Djelfa	D1	266	119	197	5	243	1088	111	4	3062	7.25	33
	D2	158	85	189	22	399	387	201	0.21	2360	7.49	36
	D3	153.8	121.6	228.8	40	418	780	213.5	1.8	2610	7.5	37
Biskra	B1	196.39	221.13	390.83	46.92	673.55	1147.92	164.75	0.00	5200	7.84	53
	B2	133.00	62.00	75.80	5.50	639.00	370.00	167.00	0.00	1440	8.22	15
Touggourt	T1	134.27	139.73	206.91	39.10	425.40	614.78	183.06	0.00	4500	7.25	55.5
	T2	134.27	143.37	275.88	39.10	464.40	600.38	176.96	0.00	4000	7.00	53
	T3	158.32	159.17	289.67	46.92	698.37	696.44	158.65	0.00	5750	7,02	57
	T4	166.33	155.52	202.31	46.92	478.58	696.44	152.55	0.00	4520	7.00	56.2
	T5	100.20	219.92	202.31	35.19	464.40	749.27	164.75	0.00	4200	7.30	56
	T6	230.46	52,25	195.42	39.10	350.96	624.39	146.45	0.00	3000	7.05	51
	T7	320.00	97.00	145.00	33.00	399.00	760.00	159.00	0.00	2160.00	8.31	49
	T8	380.00	152.00	215.00	34.00	324.00	800.00	134.00	0.00	2810.00	8.11	70
	T9	162.32	48.60	521.87	83.28	893.34	590.77	24.41	0.00	3700	7.10	51
El Oued	E1	132.00	65.00	230.00	50.00	450.00	434.00	139.00	0.00	2350.00	7.40	71
	E2	226.00	97.00	190.00	36.00	440.00	570.00	170.00	0.00	2750.00	7.70	62
	E3	258.00	84.00	200.00	75.00	476.00	662.00	89.00	0.00	2300.00	7.45	65
	E4	316.00	44.00	275.00	80.00	440.00	920.00	148.00	0.00	2890.00	7.50	69
	E5	176.00	14.00	750.00	38.00	462.00	1280.00	152.00	0.00	2251.00	7.72	70
	E6	204.40	104.51	102.00	73.39	397.03	570.00	198.86	1.5	2770.00	7.51	65
	E7	236.47	143.39	212.18	60.16	389.98	530.00	173.24	1.7	2740.00	7.49	51
El Ménéa	M1	350.00	87.00	640.00	74.00	332.00	800.00	155.60	0.5	2190.40	7.90	48
	M2	280.00	48.00	390.00	25.00	589.00	938.00	180.20	2.5	1650.00	7.40	55

Annexes

Ouargla	O1	156.00	68.00	250.00	30.00	417.00	550.00	138.30	5.5	2750.00	7.90	51
	O2	125.00	48.00	176.00	17.00	370.00	190.00	147.65	1.0	1986.23	7.05	63
	O3	106.21	113.00	218.41	35.19	414.77	513.92	164,75	1.5	3200	6.80	46
	O4	260.52	114.21	331.06	66.47	556.57	950.99	158,65	0.00	3100	7.10	55
	O5	116.23	138.51	285.08	39.10	450.22	638.80	164,75	0.5	3350	7.20	49
	O6	196.39	63.18	262.09	38.32	409.09	600.38	166,58	1.5	2590	7.20	56
	O7	260.41	101.22	310.10	41.50	360.00	835.60	123.50	0.5	2990	7.23	48
	O8	222.55	88.57	305.45	38.10	700.00	950.20	145.30	0.0	3150	7.32	52
	O9	367.73	85.41	406.92	81.72	797.63	1128.71	129.97	2.5	4000	7.30	44
	O10	433.87	96.96	469.00	47.70	1078.39	1290.57	115.94	3.5	4500	7.30	44
	O11	310.62	154.31	779.36	58.65	347.10	1066.27	145.45	2.5	2800	7.20	44
Ghardaïa	G1	100.20	111.78	289.67	19.55	425.40	499.51	146.45	0.0	2900	7.35	42
	G2	80.16	88.70	167.83	19.55	248.15	451.48	170.86	1.0	1750	7.00	40
	G3	110.22	115.43	308.07	19.55	478.58	533.13	140.35	4.5	3250	7.20	40
	G4	138.28	59.54	257.49	7.82	538.84	297.79	109.84	2.5	2750	7.30	41
	G5	136.27	102.06	197.71	11.73	407.68	489.91	115.94	1.5	2250	7.20	35
	G6	130.26	126.36	197.71	11.73	326.14	585.97	134.24	3.5	2300	7.20	32
	G7	126.25	117.86	170.13	11.73	368.68	456.29	140.35	1.5	2150	7.40	32.2
	G8	118.24	119.07	259.79	11.73	510.48	499.51	146.45	0.5	2500	7.30	32
	G9	126.25	117.86	204.61	23.46	418.31	691.63	158.65	1.5	3000	7.30	41.5
	G10	76.15	93.56	114.95	11.73	194.98	317.00	176.96	1.0	1200	7.40	30.5
Adrar	A1	256	113	346	26	388	815	111.20	2.5	1540	7.9	22
	A2	216	28	30	7.20	300	1225	98.50	1.5	1820	7.9	22
	A3	226	76	220	12.00	355	750	101.45	5.5	1660	7.35	23
	A4	114	43	150	10.50	220	376	125.80	0.0	1750	7.70	23
	A5	94	31	312	12	387	399	115.40	0.0	1840	7.60	23
Aïn Salah	AS1	22.00	16.00	260.00	15.00	400.00	10.00	24.00	2.0	1550	7.20	22
	AS2	136.00	77.00	240.00	10.00	300.00	500.00	140.00	1.0	2270	7.8	21
	AS3	107.00	82.00	220.00	10.00	295.00	500.00	153.00	0.5	2270	7.58	22
	AS4	161.00	41.00	335.00	24.00	330.00	675.00	119.00	1.5	2620	7.28	23

Annexes

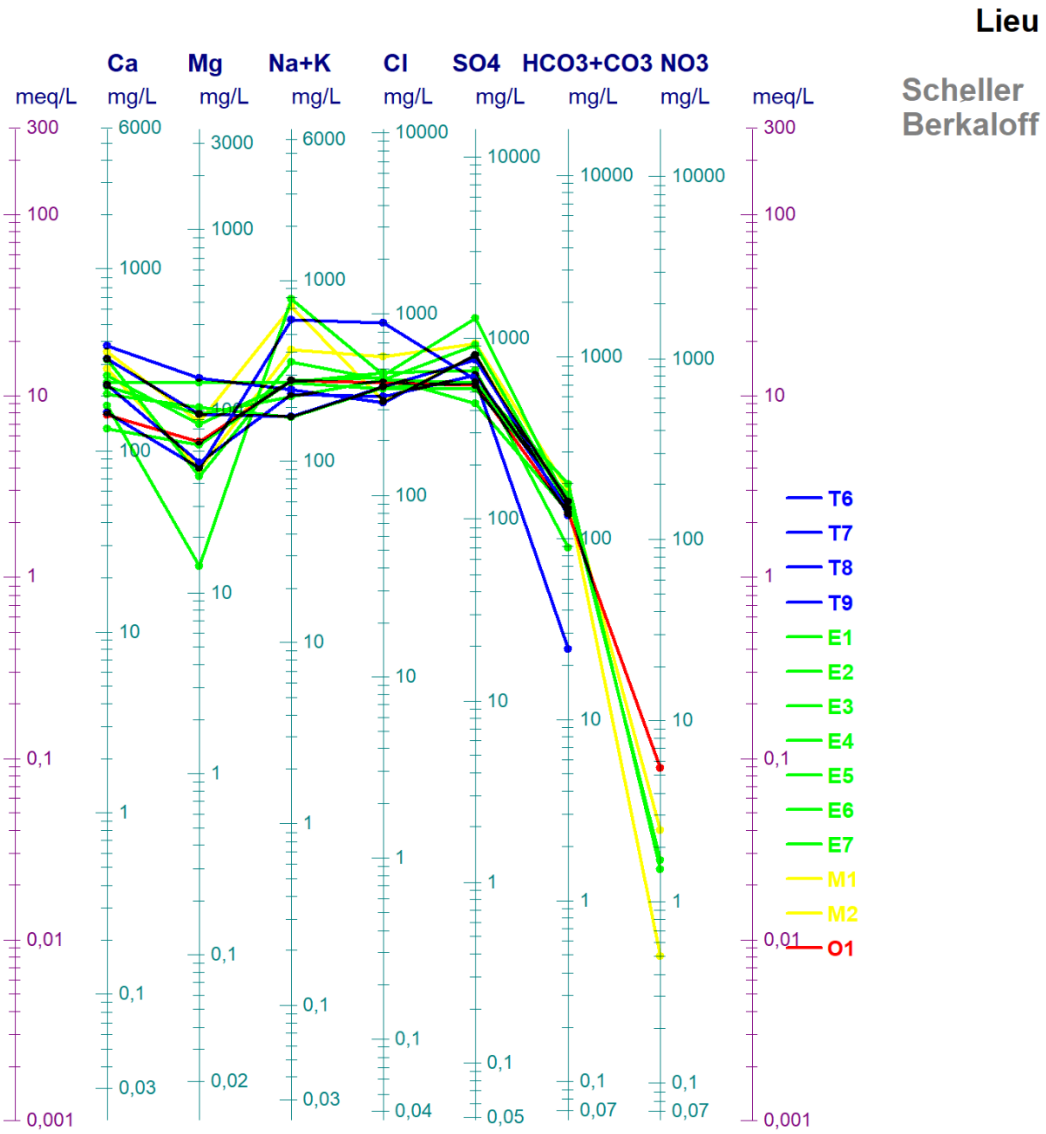
	AS5	226.00	103.00	470.00	43.00	400.00	960.00	140.00	1.0	3980	7.28	24
	AS6	107.00	106.00	260.00	9.00	307.00	600.00	250.00	2.0	2380	7.7	21
Aïn Amenas	AA1	101	99	260	9	300	580	245	1.0	2330	7.2	22
	AA2	214	95	200	11	290	788	125	0.5	2455	7.4	21

Remarque:

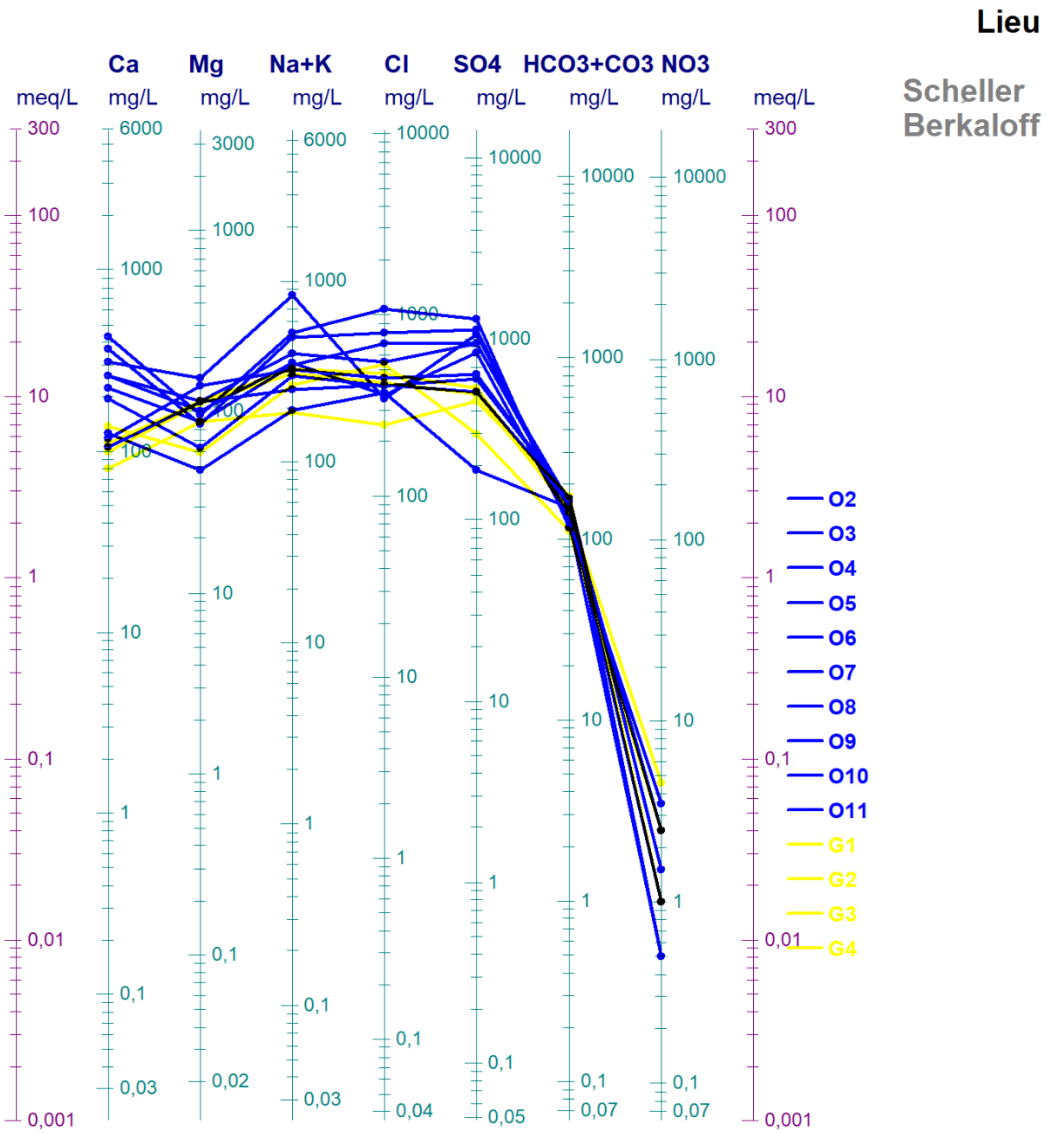
Les unités sont exprimées en **mg/l**

Annexes

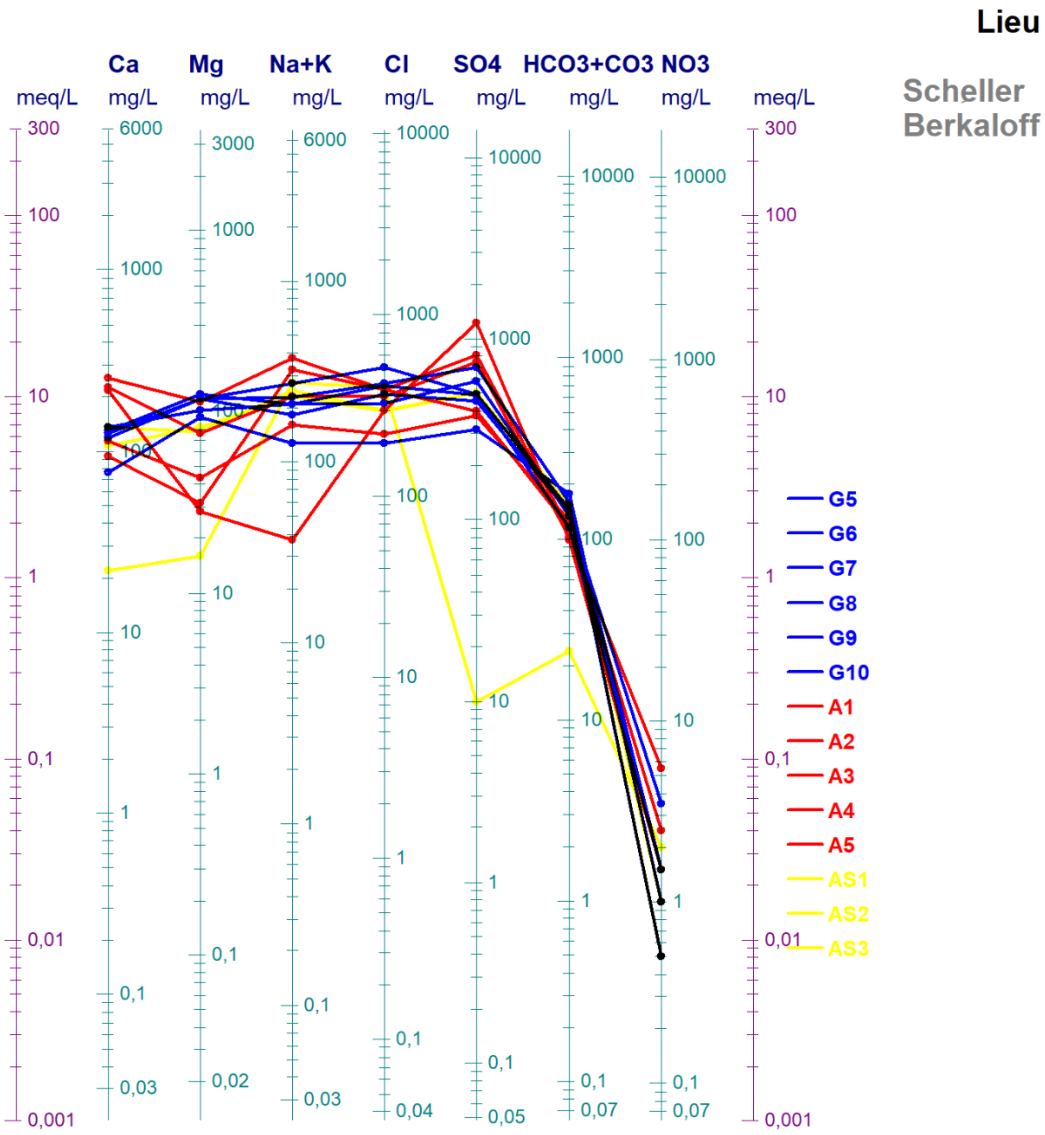
Annexe n°2: les graphes de Scheller Berkaloff :



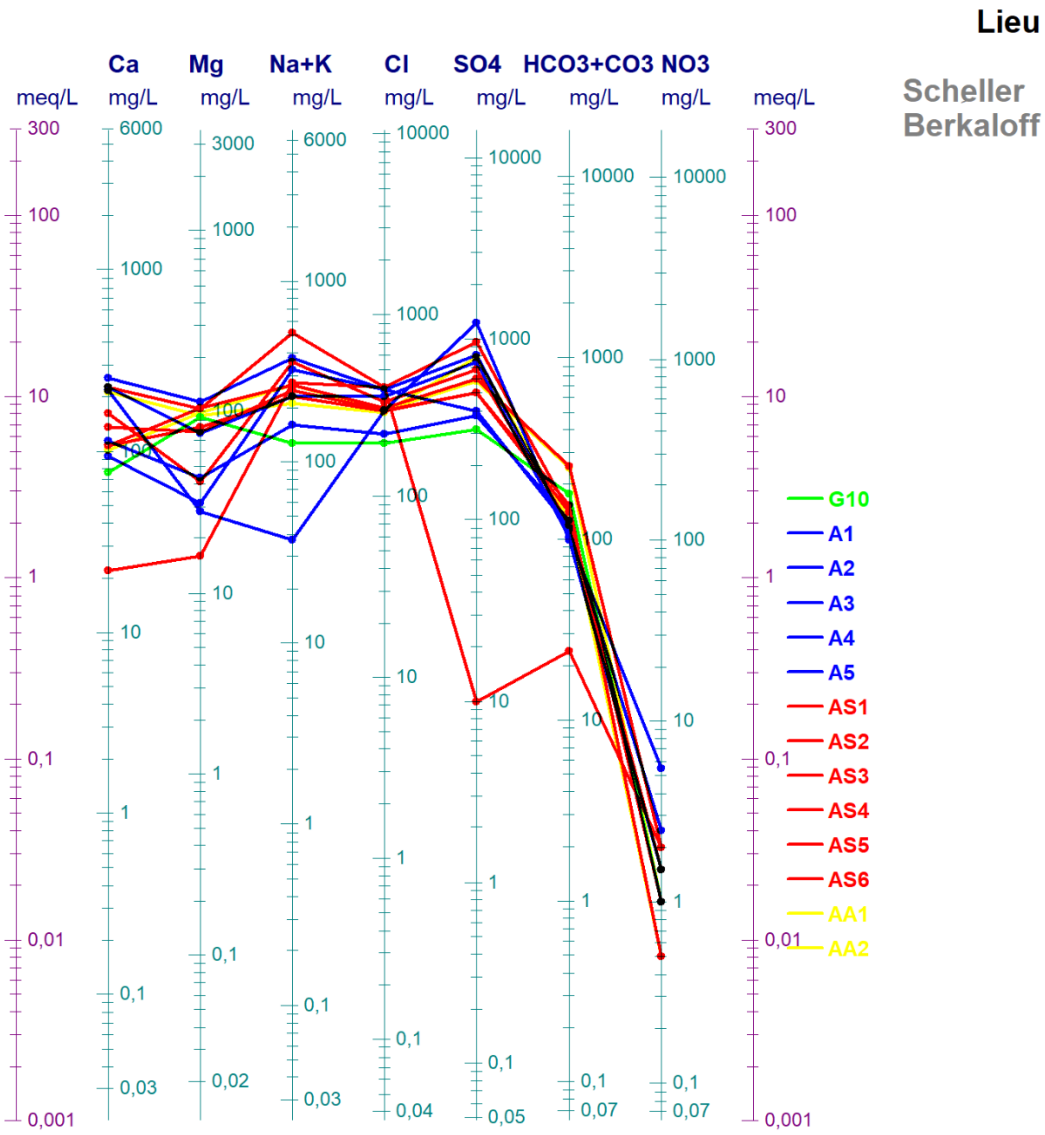
Annexes


Lieu
Scheller
Berkaloff
Ca
Mg
Na+K
Cl
SO4
HCO3+CO3
NO3
meq/L
mg/L
O2
O3
O4
O5
O6
O7
O8
O9
O10
O11
G1
G2
G3
G4

Annexes


Lieu
Scheller
Berkaloff
Ca
Mg
Na+K
Cl
SO4
HCO3+CO3
NO3
meq/L
mg/L
G5
G6
G7
G8
G9
G10
A1
A2
A3
A4
A5
AS1
AS2
AS3

Annexes


Lieu
Scheller
Berkaloff
Ca
Mg
Na+K
Cl
SO4
HCO3+CO3
NO3
meq/L
mg/L
G10
A1
A2
A3
A4
A5
AS1
AS2
AS3
AS4
AS5
AS6
AA1
AA2