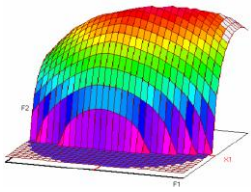
Année Universitaire : 2017-2018



Master Sciences et Techniques CAC Agiq
Chimiométrie et Analyse Chimique : Application à la gestion industrielle de la Qualité

MEMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master Sciences et Techniques

Comparaison de la régression linéaire multiple et des réseaux de neurones artificiels pour l'évaluation de la qualité chimique des eaux d'irrigation dans la région de Skhirat

Présenté par:

Melle Youssra FEKIYER

Encadré par:

- **Dr. Ahmed DOUAIK (INRA-Rabat)**
- **Pr. Adiba KANDRI RODI (FST-Fès)**

Soutenu le 13 Juin 2018 devant le jury composé de:

- Pr. A. KANDRI RODI	**FST-Fès**
- Pr. A. MELIANI	**FST-Fès**
- Pr. E. H. ALILOU	**FST-Fès**
- Pr. A. MECHAQRANE	**FST-Fès**
- Dr. A. DOUAIK	**INRA-Rabat**

Dédicaces

A la mémoire de mon défunt père.

A la plus belle créature que Dieu a créé sur terre,

A cette source de tendresse, de patience et de générosité,

A ma mère !

A mon frère qui est toujours à mes côtés

A mes chères amies : Mariem, Touraya, Fatima Ezzahra

A toute ma famille

A tous mes amis et collègues

A tous les étudiants de la promotion 2017/2018

Option : CACaqiq

A tous ceux, qui par un mot, m'ont donné la force de continuer…

Remerciements

En préambule à ce mémoire je remercie ALLAH qui m'a aidé et m'a donné la patience et le courage durant ces langues années d'étude.
Je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Mes profonds et sincères remerciements vont à **Mr. Ahmed DOUAIK** Docteur en Statistique et Géostatistique à l'URECRN. Il a conduit ce stage d'une haute compétence et grande disponibilité. J'ai particulièrement été impressionné par ses qualités scientifiques et humaines ainsi que par sa patience, sa gentillesse et la rigueur avec laquelle il a mené à bien ce travail.

Je remercie également et Chaleureusement **Mme. Adiba KANDRI RODI** professeur à la faculté de sciences et techniques de Fès. Pour l'inspiration, l'aide et le temps qu'elle a bien voulu me consacrer, ainsi pour ses conseils précieux.

Je tiens à remercier aussi les membres de jurys **Mr. A. MELLIANI**, **Mr. E. H. ALILOU** et **Mr. A. MECHAQRANE** qui m'avoir fait l'honneur d'évaluer ce travail et de participer à ce jury.

Hommages respectueux et sincères à **Mr. Mestafa EL HADRAMI** Responsable du Master CAC Agiq pour les efforts qu'il a consentis en faveur de ma formation, et pour son appréciable aide, ainsi pour la qualité de son enseignement.

Je n'oublie pas mes parents pour leur contribution, leur soutien et leur patience. Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours encouragé au cours de la réalisation de ce mémoire.

Merci à tous et à toutes.

TABLE DES MATIERES

Liste de figures

Liste des Tableaux

Introduction générale

L'eau est d'une utilité primordiale pour toutes les activités humaines y compris les utilisations domestiques, agricoles et industrielles. En agriculture, cette ressource naturelle représente un élément indispensable pour la végétation ainsi que pour le développement agricole. Le Maroc se caractérise par son climat aride et semi-aride, ces conditions climatiques ont rendu l'irrigation une exigence technique clé qui a acquis d'indéniables dimensions économiques et sociales. L'irrigation est aussi devenue un moyen de développement agricole favori et bénéficie d'une attention particulière du gouvernement. La ville de Skhirat est l'une des régions du royaume qui est connu par une énorme production de légumes. Ainsi, elle se caractérise par une importante agriculture intensive, notamment l'horticulture. Cette intensification était associée à une mauvaise utilisation des produits agrochimiques plus le pompage inconsidéré des eaux souterraines, qui deviennent de moins en moins abondantes et de mauvaise qualité. D'ailleurs la surexploitation de ces ressources, associée au phénomène de sécheresse, conduit inévitablement à la dégradation de la qualité du sol et de l'eau, ce qui entraine les problèmes de pollution par les nitrates, la salinisation et la sodification des eaux souterraines. Par conséquent, afin d'identifier ces problèmes, détecter les zones à risque, et trouver des solutions pour une gestion durable des ces ressources, une étude de la qualité de l'eau d'irrigation est nécessaire pour évaluer le degré de dégradation de cet élément primordial. C'est dans cette perspective que se situe notre travail impliquant l'étude de la qualité de l'eau d'irrigation de la région de Skhirat, en particulier dans les zones d'agriculture intensive par- ce qu'aucune étude systématique et détaillée sur la qualité de l'eau d'irrigation dans la région n'a été réalisée auparavant. Pour ce faire, le suivi de certains paramètres physico-chimiques de l'eau a été réalisé sur un nombre de puits, aussi représentatif que possible de la zone. Le traitement des données récoltées après les analyses physico-chimiques effectuées au laboratoire a été réalisé à l'aide des deux méthodes statistiques, notamment la régression linéaire multiple et les réseaux de neurones artificiels dont la deuxième constitue la méthode de base sur laquelle on s'est fondé pour évaluer le degré de dégradation de ces eaux souterraines qui sont utilisées pour l'irrigation.

Ce travail a été divisé en deux parties : La première partie est constituée de quatre chapitres. Dans le premier, nous introduirons une brève description de lieu de stage, dans le deuxième nous mettons en évidence des généralités sur l'irrigation, le troisième chapitre introduit la régression linéaire multiple et pour le dernier chapitre, nous allons aborder la théorie des réseaux de neurones artificiels. La deuxième partie comporte deux chapitres, le premier détaille les méthodes d'analyse expérimentales et statistiques utilisées dans cette étude et dans le deuxième, nous présentons les résultats des traitements statistiques des données chimiques ainsi que les discussions envisagées afin d'estimer la qualité des eaux d'irrigation.

PARTIE I : SYNTHESE BIBLIOGRAPHIQUE

CHAPITRE I : Présentation de l'entreprise

I. INRA (Institut National de Recherche Agronomique)

I.1 Histoire

L'Institut National de la Recherche Agronomique (INRA) a pour mission d'entreprendre les recherches pour le développement agricole. C'est un établissement public dont les origines remontent à 1914 avec la création des premiers services de recherche agricole officiel. L'INRA opère à travers dix centres régionaux de la recherche agronomique et 23 domaines expérimentaux répartis sur le territoire national et couvrant les divers agrosystèmes du pays [1].

II. Centre Régional de la Recherche Agronomique (CRRA) de Rabat

II.1 Présentation

La date de création : 2003

- Région : L'ex-région administrative de Rabat-Salé-Zemmour-Zaer
- Provinces : Rabat, Salé, Skhirate-Témara et Khémisset
- Nature juridique : public
- Superficie totale (ha) : 5ha
- Adresse : Avenue Mohamed Belarbi Alaoui Rabat– Instituts, 10101 – Rabat
- Boite postale : 6356
- Tél : 212 06 60 15 72 29
- Fax : 212 05 37 77 55 30
- E-mail : crra.rabat@gmail.com
- Mission : Chargé de mener les études scientifiques, techniques et économiques pour le développement de l'agriculture, de l'élevage et la conservation des ressources naturelles au niveau de la zone d'action du centre [1]. Son organisation est représentée dans la figure 1.

II.2 Organisation et axes scientifiques

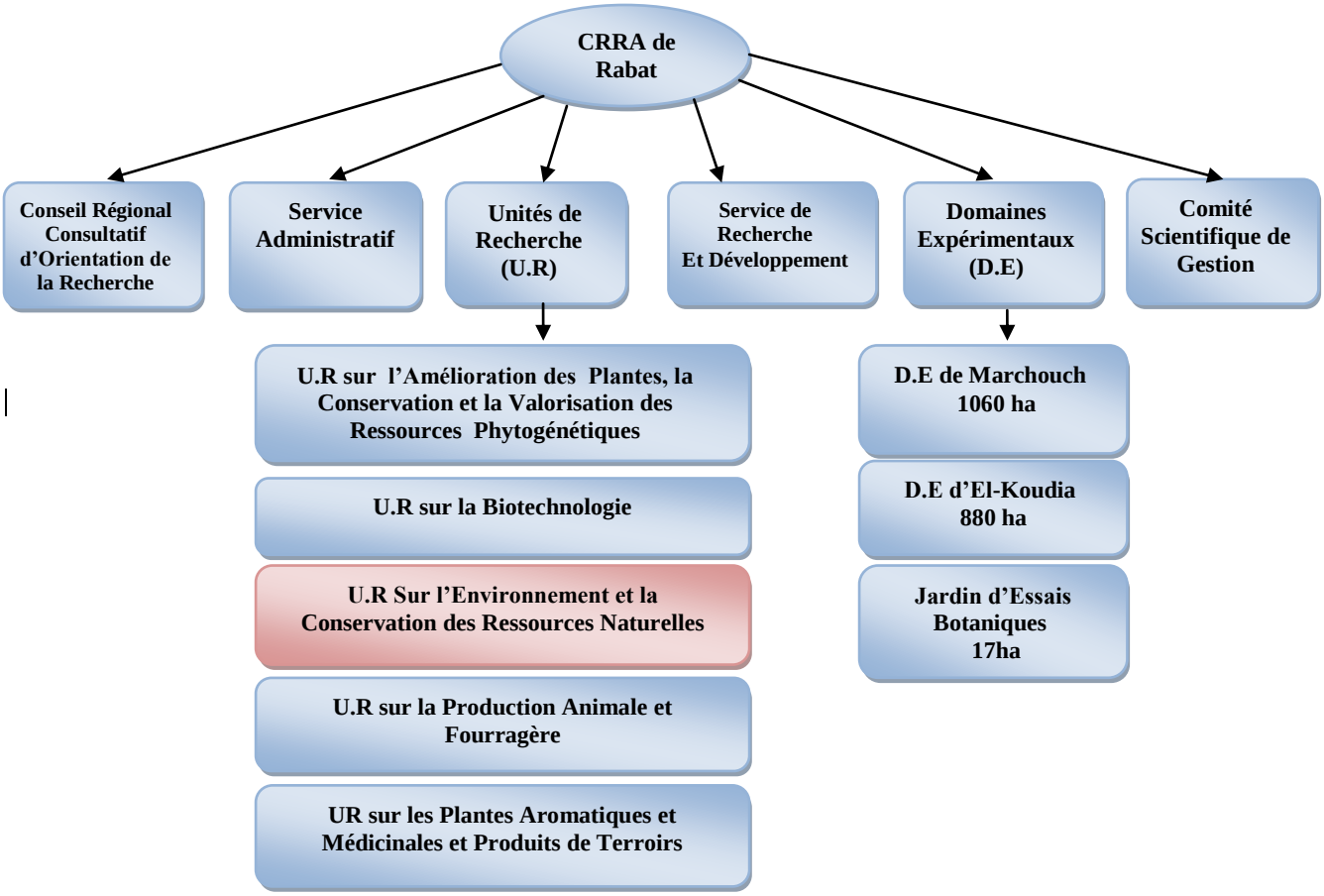


Figure 1 : Organigramme du CRRA de Rabat [2].

II.3 Unité de Recherche sur l'Environnement et la Conservation des Ressources Naturelles (URECRN)

L'Unité de Recherche sur l'Environnement et la Conservation des Ressources Naturelles (URECRN), affiliée au Centre Régional de la Recherche Agronomique de Rabat, a des missions régionale et nationale de gestion durable des ressources naturelles (sol, eau et climat). Elle est dotée d'infrastructure et d'équipements importants :

- ✓ Serre
- ✓ Laboratoire d'analyse
- ✓ Laboratoire de SIG et cartographie
- ✓ Station météorologique automatique, etc.

Cette unité réalise différentes activités de recherche, parmi lesquelles :

- ➢ Etude des paramètres physico-chimiques du sol afin d'identifier la qualité et l'effet du sol sur la végétation
- ➢ Etude de la qualité des eaux d'irrigation (salinité et pollution nitrique) des zones maraîchères de Skhirat, Ben Slimane, Tiflet et Khemisset
- ➢ Caractérisation de la qualité des eaux usées épurées de la STEP de Skhirat et quantification des apports azotés de ces eaux usées en irrigation.

CHAPITRE 2 : Généralités sur l'irrigation

I. Introduction

Pour permettre leur croissance végétative et leur développement, les plantes ont besoin d'eau appropriée en qualité et en quantité, à portée de leurs racines et au bon moment. La plus grande partie de l'eau absorbée par une plante sert à transporter les nutriments dissous du sol jusqu'aux organes aériens des plantes, d'où elle est libérée dans l'atmosphère par transpiration: l'utilisation de l'eau en agriculture est intrinsèquement consommatrice. Chaque culture a des besoins en eau particuliers, qui varient selon les conditions climatiques locales. A titre indicatif, la production d'un kilogramme de blé nécessite environ 1 000 litres d'eau qui retournent dans l'atmosphère, alors que le riz peut en exiger deux fois plus. D'où l'irrigation devient un moyen crucial pour la vie végétarienne et le développement agricole.

II. Différents systèmes d'irrigation adoptés au Maroc

Au Maroc, l'agriculture consomme entre 80 et 90% des ressources en eau [3]. L'irrigation gravitaire représente environ 80% de la superficie des grands périmètres irrigués du Maroc. Les systèmes d'irrigation peuvent être classés en deux grandes catégories: l'irrigation gravitaire et l'irrigation sous pression. Dans la pratique, on distingue l'irrigation gravitaire, l'irrigation goutte à goutte et l'irrigation par aspersion. Les figures 3 à 5 montre bien les trois systèmes d'irrigation appliqués à la parcelle [4].

II.1 Irrigation gravitaire ou de surface

Il s'agit de la technique d'irrigation la plus ancienne. Elle utilise un canal à ciel ouvert qui apporte l'eau par gravité à des canaux de plus en plus petits, venant irriguer les parcelles cultivées. Ce système d'irrigation utilise énormément d'eau, d'autant plus qu'une grande partie se perd par évaporation. On retrouve là les techniques les plus anciennement mises en œuvre, sur l'ensemble de la planète, qu'ils s'agissent de ruissellement (irrigation par planche ou à la raie) ou de submersion (irrigation par bassin), ou d'une combinaison de ces deux principes.

II.2 Irrigation goutte à goutte

L'irrigation goutte à goutte est une technique qui consiste à mettre l'eau au pied de la plante, directement à la disposition des racines à l'aide d'un goutteur [5]. Il y a plusieurs systèmes d'irrigation goutte à goutte qui existent sur le marché (de très cher au moins cher) avec des tailles différentes mais les principes de fonctionnement restent les mêmes. Parmi lesquels on note, le goutteur ou gaine, le diffuseur, l'orifice calibré ou ajutage et le micro-asperseur (micro-jet).

II.3) Irrigation par aspersion

C'est une irrigation qui projette l'eau en l'air pour tomber à la surface du sol sous forme de fines gouttelettes. C'est un réseau de conduites sous pression portant des asperseurs ou des buses, conçu pour projeter des jets ou pulvériser de l'eau sous forme de fines gouttes à la surface du sol.

III. Impact de la qualité chimique des eaux d'irrigation sur le sol et les cultures

Ayers et Westcot (1989) ont mis en évidence cinq critères d'évaluation de la qualité de l'eau pour irrigation qui sont la teneur en sel soluble (risque de salinité), la proportion relative des cations sodium (Na^+) par rapport aux autres (risque de sodium- effet sur la perméabilité du sol), la concentration des anions carbonates (CO_3^{2-}) et bicarbonates (HCO_3^-) en relation avec la concentration en calcium (Ca^{2+}) et en magnésium (Mg^{2+}) (alcalinité et dureté), la concentration en éléments qui peuvent être toxiques et le pH de l'eau d'irrigation [6] .

III.1) Salinité

Les principaux sels responsables de la salinité de l'eau se présentent comme des combinaisons de trois cations : calcium (Ca^{2+}), magnésium (Mg^{2+}) et sodium (Na^+), avec quatre anions : les chlorures (Cl^-), les sulfates (SO_4^{2-}), les carbonates (CO_3^{2-}) et les bicarbonates (HCO_3^-).

- Les chlorures de Na^+, Mg^{2+} et Ca^{2+} : ($NaCl$, $MgCl_2$, $CaCl_2$)
- Les sulfates de Na^+, Mg^{2+} et Ca^{2+} : (Na_2SO_4, $MgSO_4$, $CaSO_4$)
- Bicarbonates de Na^+, Mg^{2+} et Ca^{2+} : ($NaHCO_3$, $MgH_2(CO)_2$, $CaH_2(CO_3)_2$)
- Carbonates de Na^+, Mg^{2+} et Ca^{2+} : (Na_2CO_3, $MgCO_3$, $CaCO_3$)

Une concentration élevée en sels dans l'eau affectera négativement le rendement des récoltes, provoquera une dégradation des sols par la dispersion des colloïdes, comme elle peut également agir sur la croissance végétale et donc sur le développement de l'agriculture. Elle pose un problème dès l'instant où l'accumulation de sels dans la zone racinaire atteint une concentration qui provoque une baisse de rendement. Celui-ci baisse lorsque les sels sont si concentrés dans la zone racinaire que la culture ne peut plus extraire l'eau en quantité suffisante de la solution salée du sol et subit de ce fait un déficit hydrique pendant un laps de temps. En effet, les sels accumulés dans le sol, peuvent limiter ou complètement arrêter la croissance du végétal suite à une élévation de la pression osmotique du milieu.

III.2) Sodium: Proportion relative des cations sodium (Na^+) par rapport aux autres

S'il est en très grande quantité dans l'eau d'irrigation, le Na^+ devient parmi les éléments les plus indésirables. Cet élément trouve son origine dans l'altération de la roche et du sol, des intrusions d'eau de mer, des eaux traitées et des systèmes d'irrigation.

Le problème principal avec une grande quantité de sodium c'est qu'il agit sur le complexe argilo-humique du sol (Figure 2) par les cations échangeables (Na+) ce qui provoque le lessivage des

bases, la destruction des ponts calciques et comme conséquence une forte vitesse d'infiltration des eaux. En effet, le complexe adsorbant devient saturée en Na^+ et provoque la dispersion d'argiles (fraction fine) diminuant ainsi la porosité (aération) et l'emmagasinement de l'eau. Le sol devient donc dur et compact lorsqu'il est sec et imperméable à l'eau. D'autre part le remplacement du potassium (K+) par les ions sodium (Na+) prohibe la racine de la plante d'absorber cet élément indispensable pour ses activités enzymatiques (favorise la production des protéines : conversion plus rapide de N inorganique en protéines, favorise la photosynthèse, améliore l'efficacité des engrais azotés,…). Par conséquent, la croissance est freinée suite à une carence en potassium.

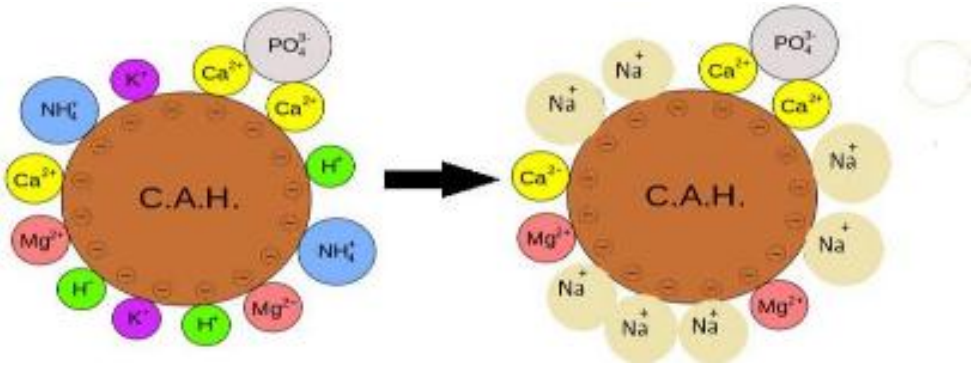


Figure 2 : Effet de sodium sur le complexe argilo-humique.

III.3) Alcalinité et dureté

III.3.1 Alcalinité : Risque des carbonates et bicarbonates

Une forte teneur en carbonates (CO_3^{2-}) et en bicarbonates (HCO_3^-), environ 180-240mg/L, augmente la valeur du taux d'absorption de sodium (S.A.R), (environ > 180-240mg/L), Les raisons sont les suivantes:

Les ions carbonates et bicarbonates combinés au calcium ou au magnésium précipiteront sous forme de carbonates de calcium ($CaCO_3$) ou carbonates de magnésium ($MgCO_3$) dans des conditions de sécheresse.

Lorsque la concentration de Ca^{2+} et de Mg^{2+} décroît, en comparaison la teneur en sodium et le SAR deviennent plus importants. Ceci causera un effet d'alcalisation et augmentera le pH. Par conséquent, lorsqu'une analyse d'eau indique un pH élevé, ceci peut être un signe d'une teneur élevée en ions carbonates et bicarbonates. Ces conditions sont défavorables pour la stabilisation du sol ainsi que pour la croissance de la plante.

III.3.2 Dureté : calcium et magnésium

a) Calcium

Un excès de calcium sera difficile à identifier, puisqu'il entraîne le blocage d'autres éléments nutritifs, ce qui provoquera des carences en potassium, magnésium, manganèse et fer chez la plante,

ce qui ralentit sa croissance, une grande quantité en calcium engendra ainsi des brûlures au niveau des feuilles.

b Magnésium

Comme les autres cations (c'est-à-dire les éléments minéraux à charge positive), le magnésium en excès va s'opposer au prélèvement par les plantes de tous les autres éléments positifs : calcium, potassium mais aussi tous les oligo-éléments (sauf le molybdène). Ainsi l'excès de magnésie dans un sol est un facteur d'aggravation des phénomènes chlorotiques. Dans le même ordre d'idée, l'excès de MgO participe à « défloculer » le sol, et donc à dégrader sa structure (moindre prise en mottes). En effet, le magnésium prend la place du calcium sur le complexe argilo-humique, mais sans en avoir tous les rôles agglomérants [7].

III.4) Concentration en éléments toxiques

III.4.1 Chlorures et sulfates

Lorsqu'ils sont présents dans l'eau d'irrigation, ces éléments contribuent à augmenter la concentration des sels solubles. Des concentrations excessives de chlorures et de sulfates peuvent causer des brûlures sur le bout des feuilles du gazon et voire même entraîner la mort des plantes. Des concentrations de 250 à 400 ppm sont considérées comme indésirables pour l'irrigation des plantes sensibles aux sels.

III.4.2 Bore

Le bore (B) est un élément mineur essentiel à la croissance de la plante mais il n'est requis qu'en minime quantité. Le bore est soluble dans l'eau et on le retrouve dans plusieurs sources d'eau utilisées pour l'irrigation. Lorsque sa concentration dans l'eau excède 1 à 2 ppm, le bore peut être toxique pour le gazon. De plus, le bore a tendance à s'accumuler dans le sol en formant des complexes chimiques qui sont difficiles à lessiver.

III.5) pH de l'eau d'irrigation

Le pH influence la forme et la disponibilité des éléments nutritifs dans l'eau d'irrigation. Le pH de l'eau d'irrigation devrait se situer entre 5,5 et 6,5. À ces valeurs, la solubilité de la plupart des micro-éléments est optimale. Le pH de l'eau d'irrigation affecte également l'assimilation de certains éléments dans le sol, le meilleur exemple est le phosphore. Dans les sols acides, le phosphore se complexe avec le fer et devient insoluble. Dans les sols basiques, il se complexe avec le calcaire. Il peut donc y avoir du phosphore dans le sol, mais ce phosphore est non-prélevable.

CHAPITRE 3 : Régression linéaire multiple (RLM)

I. Régression Linéaire Simple (RLS)

La régression linéaire se classe parmi les méthodes d'analyses statistiques qui traitent les données quantitatives [10]. C'est une méthode d'investigation sur données d'observations, ou d'expérimentation.

Elle s'adresse à un type de problème où les 2 variables quantitatives continues x et y ont un rôle asymétrique : la variable y dépend de la variable x.

Dans le cas de la régression linéaire simple (RLS) on a une seule variable dite variable explicative x qui est en relation avec une autre variable dite variable expliquée y, cette liaison peut être modélisée par une fonction de type :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \forall_i \in \{1, \dots, n\}$$

β_0: C'est une constante (ordonnée à l'origine), c'est la valeur prédite de y quand $x = 0$

β_1 : La pente, c'est l'augmentation de la variable y lorsque la variable x augmente d'une unité.

ε_i : C'est un terme aléatoire, tient un rôle très important dans la régression. Il permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire que l'on cherche à établir entre y et x c.à.d. résumer le rôle des variables explicatives absentes.

II. Estimation du modèle par la méthode des moindres carrés ordinaires (MCO)

La première chose à faire est de dessiner le nuage des points (x_i, y_i) $\forall_i = 1, \dots\dots, n$, pour déterminer le type de liaison pouvant exister entre x et y. A priori, n'importe quel type de liaison est possible, (mais ici, on s'intéresse à une relation de type affine (Figure 3), $y = ax + b$: condition principale d'application de la régression linéaire).

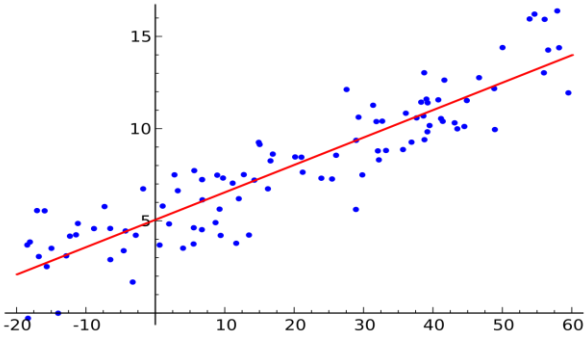


Figure 3 : Relation affine entre x et y.

Après avoir schématisé le nuage de points, on va tenter de le résumer par une droite ce qu'on appelle droite de régression de y en fonction de x et pour pouvoir tracer cette droite, il faut connaitre son équation. Autrement dit, estimer ses paramètres (β_0 et β_1).

Les coefficients du modèle β_0 et β_1 sont estimés par la méthode des moindres carrés ordinaires qui consiste à minimiser la somme des carrées des écarts (des erreurs).

On appelle estimateurs des Moindres Carrés Ordinaires (notée par MCO) $\hat{\beta}_0$ (estimateur de l'ordonnée à l'origineβ_0) et $\hat{\beta}_1$ (estimateur de la pente β_1) les valeurs qui minimisent la quantité :

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ , ou } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Expression des estimateurs ($\hat{\beta}_0$ et $\hat{\beta}_1$)

A la fin des calculs, on obtient pour expressions des estimateurs :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{cov(x,y)}{var(x)} = \frac{S_{xy}}{S_x^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

III. Test de la pente de la droite de régression

Après avoir postuler notre modèle et estimer ses paramètres, on doit forcément se demander ce que vaut ce modèle en vérifiant son fonctionnement. Autrement dit si cette pente β_1 qui représente la liaison entre les x et les y est suffisamment abrupte pour pouvoir affirmer la liaison, il va donc falloir utiliser un test statistique pour vérifier cela.

Le test de significativité de la pente consiste à vérifier l'exogène x sur l'endogène y, qu'on va le voir un peu plus en détail dans la partie RLM.

IV. Régression linéaire multiple (RLM)

La régression linéaire est une des méthodes statistiques les plus utilisées dans de nombreux domaines pour l'étude de données multidimensionnelles. Elle constitue la généralisation naturelle de la régression simple, où on cherche à expliquer les valeurs prises par la variable endogène y à l'aide de p variables exogènes $x_1, \dots \dots, x_p$**.** La différence essentielle réside dans le formalisme qui passe par des écritures matricielles des estimateurs et de leurs variances.

IV.1 Modèle de la RLM

On cherche à décrire la relation existant entre une variable quantitative y appelée variable à expliquer (ou encore, réponse, exogène ou dépendante) et plusieurs variables quantitatives $x_1, \dots \dots, x_p$ dites variables explicatives (ou encore de contrôle, endogènes, indépendantes ou régresseurs).

Le modèle de régression linéaire multiple, noté par RLM, est défini par :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \dots + \beta_p x_{ip} + \varepsilon_i \qquad \forall_i \in \{1, \dots, n\}$$

où β_0, β_1, β_2, ……, β_p sont appelés les paramètres ou les coefficients inconnus du modèle que l'on veut estimer à partir des données

On a dans ce modèle de RLM :

- $i = 1, \ldots\ldots, n$ correspond au numéro des observations (des individus).
- y_i est la $i - éme$ observation de la variable y.
- x_{ip} est la $i - éme$ observation de la $p - éme$ variable.
- ε_i est l'erreur du modèle (bruit). Il représente la déviation entre ce que le modèle prédit et la réalité.

IV.2 Notation matricielle

Ce modèle s'écrit sous la forme des équations comme suit :

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

ou sous la forme matricielle :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

De façon équivalente, on écrit :

$$Y = \beta X + \varepsilon$$

Où

Y : est le vecteur à expliquer de taille $n \times 1$

β : est le vecteur des coefficients à estimer de taille $p \times 1$

X : est la matrice, de taille $n \times (1 + p)$, qui contient l'ensemble des observations sur les exogènes , avec une première colonne formée par la valeur 1 indiquant que l'on intègre la constante β_0 dans l'équation où p étant le nombre de variables explicatives réelles.

ε : est le vecteur des erreurs de taille $n \times 1$

Remarque

1. Le coefficient β_0 est un paramètre appelé ordonnée à l'origine qui représente la moyenne des y_i lorsque la valeur de chaque variable explicative est égale à 0.

2. Les coefficients $\beta_j (j = 1, \dots\dots, p)$ représentent le changement subi par $E(y_i)$ correspondant à un changement unitaire dans la valeur de la $j - ième$ variable explicative, lorsque les autres variables explicatives demeurent inchangées.

IV.3 Estimation du modèle par la méthode des moindres carrés ordinaires (MCO)

Conditionnellement à la connaissance des valeurs des $X_j (j = 1, \dots\dots, p)$, les paramètres inconnus du modèle : le vecteur $\beta = (\beta_0, \dots\dots, \beta_p)^t$ et σ_ε^2, sont estimés par minimisation du critère des moindres carrés ordinaires (MCO).

Le principe des moindres carrés choisit le vecteur $\hat{\beta}$ minimisant la fonction de la somme des carrés des résidus, notée par SCE_r.

Proposition

L'estimateur par moindres carrés ordinaires (MCO) $\hat{\beta}$ de β dans le modèle de régression linéaire multiple est :

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

IV.4 Estimation de la variance des résidus

On peut construire l'estimateur sans biais pour σ_ε^2 qui est :

$$S_{\hat{\varepsilon}}^2 = \hat{\sigma}_{\hat{\varepsilon}}^2 = \frac{\hat{\varepsilon}^t \hat{\varepsilon}}{n-p-1} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-p-1} = \frac{SCR}{n-p-1} \quad \text{où} \quad \hat{\varepsilon} = Y - X\hat{\beta}$$

La variance résiduelle mérite d'être examinée car elle sert à deux choses :

- Un intermédiaire de calcul (grâce à la variance résiduelle, on peut faire des tests, calculer les intervalles de confiance,…..)
- Elle constitue également une incertitude qu'on aura sur chaque mesure.

IV.5 Intervalle de confiance

Après avoir obtenu l'estimateur, il ne reste plus qu'a construire son intervalle de confiance.

Proposition

L'intervalle de confiance permet d'estimer la valeur d'une grandeur statistique dans la population (toujours inconnue).

On construit l'intervalle de confiance à partir des données de notre échantillon pour la pente β_j.

1. Pour tout $j \in \{1, \dots, p\}$, un intervalle de confiance de niveau de $(1-\alpha)$ pour β_j est :

$$\left[\hat{\beta}_j - t_{n-p-1}^{1-\frac{\alpha}{2}} S_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p-1}^{1-\frac{\alpha}{2}} S_{\hat{\beta}_j}\right]$$

où $t_{n-2}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ d'une loi Student T_{n-p-1}.

IV.6 Décomposition de la variabilité

Quand on s'intéresse à un phénomène, on dit souvent qu'il est du à tant pourcent à ceci, et tant pourcent à cela. C'est le même principe qui s'applique pour l'approche de régression.

Lorsqu'on réalise plusieurs répétitions d'une expérience, on n'obtient pratiquement pas le même résultat, les valeurs obtenues s'écartent de leur moyenne (centre de gravité) et c'est cet écart ou cette variabilité des résultats qu'on l'appel écart total ou variabilité totale qu'on va essayer d'expliquer par le modèle de régression, mais le modèle postulé explique juste une portion de cette variabilité, l'autre portion reste inexpliquée

Pour mieux connaitre le pouvoir explicatif du modèle, on effectue un calcul de la décomposition de variabilité suivante (Figure 4):

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

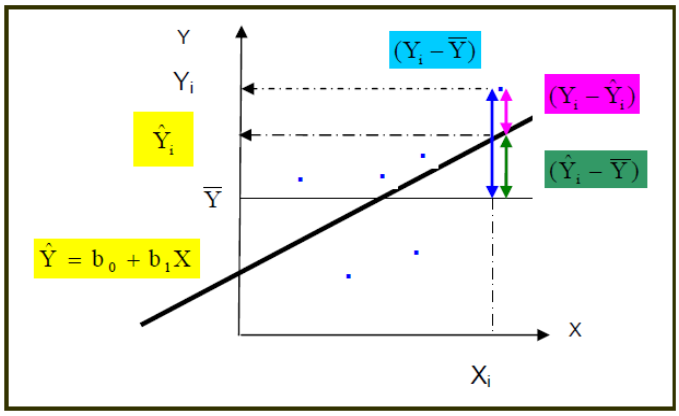


Figure 4 : Explication géométrique de la décomposition de la variabilité.

Ecart total $(Y_i - \overline{Y})$ = écart dû au model $(\hat{Y}_i - \overline{Y})$ + écart résiduel $(Y_i - \hat{Y}_i)$

On peut obtenir la décomposition :

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}$$

$$SCE_t = SCE_{reg} + SCE_r$$

où

- ✓ SCE_t : désigne la somme des carrés totaux (centrés), autrement dit l'écart total des résultats par rapport à leur moyenne (variabilité totale).
- ✓ SCE_{reg} : la somme des carrés expliqués (centrés), c'est la portion de variabilité expliquée par le modèle.
- ✓ SCE_r : la somme des carrés des résidus, c'est la portion inexpliquée par le modèle, elle représente l'écart entre ce qu'on observe et ce que le modèle a donné.

IV.7 Eléments d'évaluation de la qualité du modèle

Un modèle est une représentation de la réalité. Une fois défini, on est amené à évaluer sa qualité explicative et prédictive, en examinant entre autres :

IV.7.1 Test de signification globale du modèle

Lorsqu'on réalise une régression multiple, on souhaite savoir si le lien entre la variable dépendante et les variables explicatives est significatif. Dans ces conditions, on pose l'hypothèse :

$$\begin{cases} H_0 \ : \ \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_1 \ : \ \exists! \ \beta_j \neq 0; (j = 1, \dots, p) \end{cases}$$

et on utilise l'analyse de la variance pour la tester.

C'est l'hypothèse que tous les coefficients de régression, sauf la constante, son nuls. L'hypothèse nulle signifie que le coefficient de détermination multiple théorique R^2 est nul et que le R^2 obtenu par le calcul ne représente qu'une valeur due aux termes aléatoires.

En effet, la variation totale (SCE_t), comme nous avons déjà vu se décompose en une variation due à la régression (SCE_{reg}) et une variation résiduelle (SCE_r).

Cette relation permet de dresser le tableau d'analyse de la variance (Tableau 1), en utilisant les nombres de degrés de liberté appropriés.

Tableau 1 : Tableau d'analyse de la variance pour un modèle de régression.

Source de variation	***ddl***	***SCE***	***CM***	***F_{obs}***	***Prob***
Régression	$p-1$	SCE_{reg}	$\mathbf{CM_{reg}}$	$F_{obs} = \dfrac{SCE_{reg}/p}{SCE_r/(n-p-1)} = \dfrac{CM_{reg}}{CM_r}$	p
Résidu	$n-p$	SCE_r	$\hat{\sigma}^2 = CM_r$		
Total	$n-1$	SCE_t			

On rejette l'hypothèse H_0 lorsque $\mathbf{F_{obs}} \geq F_{p,n-p-1}^{1-\alpha}$, ou encore lorsque la probabilité de signification (p) est inférieure ou égale au seuil de signification α.

La statistique $\mathbf{F_{obs}}$ indique si la variance expliquée est significativement supérieure à la variance résiduelle. Dans ce cas, on peut considérer que l'explication emmenée par la régression traduit une relation qui existe réellement dans la population.

IV.7.2 Coefficients de détermination R^2 et R^2_{adj} ajusté

Le rapport entre SCE_{reg} et SCE_t représente la proportion de variance expliquée par le modèle et porte le nom de coefficient de détermination, noté par R^2 :

$$R^2 = \frac{SCE_{reg}}{SCE_t} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{SCE_{reg}}{SCE_t}$$

Ce coefficient R^2 est compris entre 0 et 1 : plus il est proche de 1 et plus grande est la part expliquée, autrement dit meilleure est la régression. Inversement, un coefficient R^2 proche de 0 indique que la quantité SCE_r est élevée.

Le coefficient R^2 est un indicateur de la qualité de l'ajustement des valeurs observées par le modèle mais il a le défaut de ne pas tenir compte du nombre de variables explicatives utilisés dans le modèle. On ne peut pas l'utiliser pour comparer plusieurs modèles entre eux car, si on ajoute une variable explicative à un modèle, la part des erreurs diminue forcément et donc le coefficient R^2 augmente : cela signifie que plus il y a de variables explicatives et plus le R^2 est élevé. Or un modèle n'est pas nécessairement meilleur parce qu.il a plus de variables explicatives.

On définit donc un coefficient R^2 ajusté qui tient compte des degrés de liberté. Ce coefficient, noté par R^2_{adj}, est défini comme suit :

$$R^2_{adj} = 1 - \frac{\frac{SCR}{n-p-1}}{\frac{SCT}{n-1}} = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2)$$

IV.7.4 Test de Student de signification du paramètre du modèle

L'objectif du test de Student est d'évaluer l'influence de la variable X_j $(j = 1, \dots, p)$ sur Y, on considère les hypothèses :

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

où β_j est le paramètre associé à la variable explicative X_j

L'hypothèse H_0 de nullité d'un paramètre du modèle peut être testée au moyen de la statistique de

$$T_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

1. Si $\left|T_{\hat{\beta}_j}\right| \geq t_{n-p-1}^{1-\frac{\alpha}{2}}$, on rejette l'hypothèse H_0

où $t_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ d'une loi Student à $(n-p-1)$ ddl.

Si l'hypothèse H_0 est rejetée, donc le paramètre β_j est significativement différent de 0, autrement dit la variable X_j $(j=1,\dots,p)$ à une influence sur la variable Y.

IV.7.5 La qualité des résidus

Dans ce qui précède, nous avons souligné qu'un coefficient de détermination élevé et des coefficients de régression partielle significativement différents de zéro n'entrainent nullement que le modèle est bien ajusté. Il est conseillé de juger a posteriori la qualité du modèle construit en analysant, entre autres, la structure des résidus, car cela permettra de :

- ✓ Détecter d'éventuelles valeurs aberrantes ou atypiques.
- ✓ Déceler une chronologie particulière dans les données (autocorrélation)
- ✓ Vérifier certaines des hypothèses que nous avons posées, d'une part pour obtenir des coefficients de régression sans biais et de variance minimum et d'autre part pour réaliser des inductions statistiques (intervalles de confiance et tests d'hypothèses).

L'un des moyens d'analyse des résidus, que nous proposons, est l'examen graphique (Figure 5). Si certaines anomalies apparaissent en examinant ces graphiques, certaines propositions seront faites, selon les cas, pour améliorer la qualité du modèle. On peut citer, entre autres :

- ✓ Des transformations de variables,
- ✓ Un ajout de variables explicatives
- ✓ Une élimination de certaines observations atypiques, et/ou
- ✓ Une modification du modèle proposé.

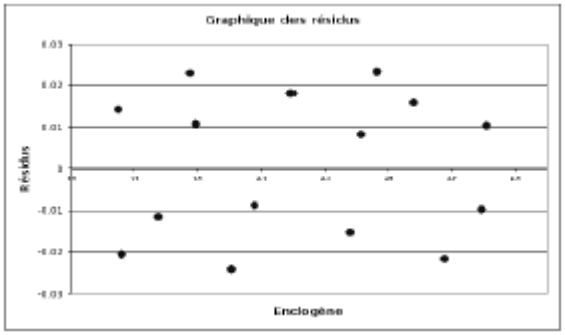


Figure 5 : Représentation des résidus standardisés en fonction des valeurs estimées de la variable dépendante (données fictives).

IV.7.6 Etude de la colinéarité entre les variables indépendantes

On dit qu'il y a colinéarité exacte lorsqu'il existe des relations linéaires entre les variables explicatives. Le vecteur des paramètres inconnus $\hat{\beta}$ est alors indéterminé, puisque le déterminant de la matrice (X^tX) est nul et on ne peut pas calculer son inverse. Dans ce cas, l'élimination d'une ou de plusieurs variables explicatives qui sont colinéaires s'impose.

Certains éléments sont des indicateurs de la colinéarité :

- De fortes corrélations entre les variables explicatives ;
- Des coefficients de régression de grandeurs importantes en valeur absolue et/ou de signes opposés à ceux auxquels on s'attendait ;
- Une assez grande instabilité des coefficients de régression ;

IV.8 Prévision

Comme dans le cadre du modèle de régression linéaire multiple, les buts de la régression est de proposer des prédictions pour la variable à expliquer Y lorsqu'on a :

$x_{n+1} = (x_{n+11} + x_{n+1p})^t$ une nouvelle valeur pour laquelle, on veut prédire y_{n+1} qui est définie par :

$$y_{n+1} = x_{n+1}^t \beta + \varepsilon_{n+1}$$

A partir des n observations précédentes, on a pu calculer un estimateur $\hat{\beta}$ *de* β . Donc, il est naturel de prédire la valeur correspondante via le modèle ajusté $\hat{y}_{n+1}$ définie par :

$$\hat{y}_{n+1} = x_{n+1}^t \hat{\beta}$$

Ce qui permet de construire l'intervalle de confiance pour y_{n+1} :

$$\left[\hat{y}_{n+1} - t_{(n-p-1)}^{1-\frac{\alpha}{2}} S_{\hat{\varepsilon}} \sqrt{x_{n+1}^t (X^t X)^{-1} x_{n+1}}, \hat{y}_{n+1} + t_{(n-p-1)}^{1-\frac{\alpha}{2}} S_{\hat{\varepsilon}} \sqrt{x_{n+1}^t (X^t X)^{-1} x_{n+1}}\right]$$

Cela signifie, que pour une observation n dans la population qui a pour valeur de variable explicative x_{n+1}, sa réponse correspondante y_{n+1} se trouve dans l'intervalle de confiance construit sur l'ensemble d'échantillon.

CHAPITRE 5 : Théorie des réseaux de neurones artificiels

I. Introduction

Les réseaux de neurones artificiels (RNA) sont des modèles inspirés de la neurobiologie qui imitent le fonctionnement du cerveau. Ils sont basés sur la fonction neuronale, parce que les neurones sont identifiés comme éléments cellulaires responsables du traitement de l'information dans le cerveau humain. Les réseaux de neurones artificiels se sont donc basés sur l'hypothèse disant que le raisonnement intelligent des êtres humains a pour origine la structure de système nerveux et donc on peut l'inculquer à un ordinateur en lui implémentant un réseau de neurones artificiels pour le rendre « Intelligent», tout en sauvegardant sa puissance et sa rapidité d'exécution [11].

II. Historique

1943 : Le neurone formel (McCulloch & Pitts)

1949 : Première règle d'apprentissage (Hebb)

1958 : Le perceptron (Rosenblatt)

1960 : L'adaline (Widrow & Hoff)

1969 : Limite Perceptrons (Minsky & Papert)

⟶ les limites du Perceptron

⟶ besoin d'architectures plus complexes,

⟶ Comment effectuer l'apprentissage ? On ne sait pas !

--------: ~ 17 ans de sub-stagnation

1986 : Rétropropagation (Rumelhart & McClelland)

Actuellement : Multitudes d'applications.

III. Domaines d 'application

Vu l'utilité majeur des RNA dans le traitement des données, ainsi que leur fiabilité et leur performance en classification, leur domaine d'application devient de plus en plus vaste, on trouve :

- **L'industrie :** contrôle qualité, diagnostic de panne, corrélations entre les données fournies par différents capteurs, analyse de signature ou d'écriture manuscrite.
- **La finance :** prévision et modélisation du marché (cours de monnaies...), sélection d'investissements, attribution de crédits.
- **La télécommunication et l'informatique :** analyse du signal, reconnaissance de formes (bruits, images, paroles), compression de données.
- **L'environnement :** évaluation des risques, analyse chimique, prévisions et modélisation météorologiques et hydrologiques, gestion des ressources.

IV. Réseau de neurones: fondement biologique

Les modèles neuronaux ont été développés par analogie avec le neurone biologique, il est donc important de rappeler au préalable son fonctionnement.

IV.1 Neurone

C'est l'élément de base du système nerveux. Il reçoit des signaux en provenance des neurones voisins d'un côté par des chevelures appelées dendrites, les traite dans le corps cellulaire puis engendre, conduit et transmet l'influx nerveux de l'autre côté à d'autres neurones via un long prolongement appelé axone. La communication entre deux neurones voisins s'effectue au niveau de la terminaison de l'axone, cet endroit est appelé synapse (Figure 6).

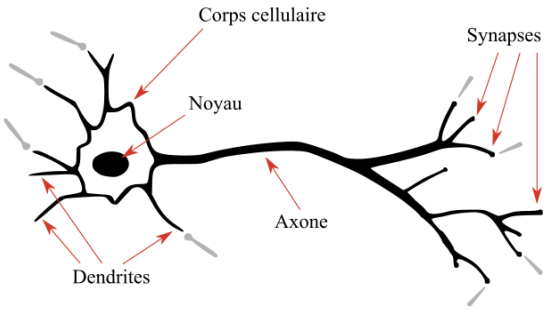


Figure 6 : Neurone biologique.

V. Réseaux de neurones artificiels

V.1 Modélisation

V.1.1 Principe général

L'archétype d'un neurone vu précédemment nous permet maintenant d'extraire un modèle épuré qu'on appellera le 'neurone artificiel'. Nous partons d'un schéma simplifié d'un neurone biologique (Figure 7).

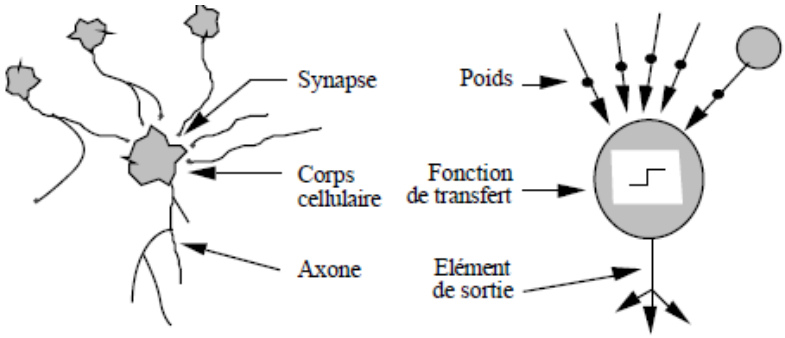


Figure 7 : Mise en correspondance neurone biologique / neurone artificiel.

A partir du schéma présenté ci-dessus, il est bien clair que les deux neurones fonctionnent similairement. L'influx nerveux (dans l'approche biologique) ou l'information (dans l'approche artificielle) se propagent de la même façon tout au long du neurone. Voyons plus en détail comment tout cela s'organise dans un neurone artificiel.

V.1.2 Modèle mathématique du neurone artificiel (formel) de MacCuloch et Pitts (1943).

Un neurone formel simple peut réaliser des fonctions logiques, arithmétiques et symboliques complexes. Alors qu'est-ce qu'un neurone formel ?

Un neurone artificiel formel est, comme on l'a vu, l'unité élémentaire de traitement d'un réseau de neurones. Il est connecté à des sources d'information en entrée (d'autres neurones par exemple) et renvoie une information en sortie.

a. Entrées

On note $(x_i)\ 1 \leq i < n$ les n informations parvenant au neurone. De plus, chacune sera plus ou moins valorisée vis à vis du neurone par le biais d'un poids. Un poids est simplement un coefficient w_i lié à l'information x_i. La $i^{éme}$ information qui parviendra au neurone sera donc en fait $w_i * x_i$. Il y a toutefois un "poids" supplémentaire, qui va représenter ce que l'on appelle le coefficient de biais. Nous le noterons w_0 et le supposerons lié à une information $x_0 = -1$. Nous verrons plus tard son utilité, dans la section Fonction d'activation.

Le neurone artificiel (qui est une modélisation des neurones du cerveau) va effectuer une somme pondérée de ses entrées plutôt que de considérer séparément chacune des informations. On définit une nouvelle donnée, S, par :

$$S = \sum_{i=0}^{n} w_i x_i = \left(\sum_{i=1}^{n} w_i x_i\right) - w_0$$

C'est en fait cette donnée-là que va traiter le neurone. Cette donnée est passée à la fonction d'activation, qui fait l'objet de la prochaine section. C'est d'ailleurs pour ça que l'on peut parfois appeler un neurone une unité de traitement (Figure 8).

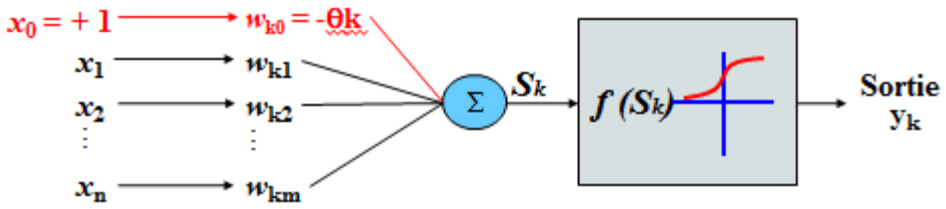


Figure 8 : Modélisation du Neurone artificiel.

b. Fonction d'activation (fonction de transfert)

Le réseau renvoie des réels grâce à un seuil θ qui se fixe à priori par l'utilisateur à l'aide de la fonction d'activation, il renvoie par exemple une valeur de 1 si la somme pondérée calculée par le neurone est supérieure au seuil fixé et il renvoi 0 sinon. On utilise généralement des fonctions d'activation à valeurs dans l'intervalle réel [0,1]. Quand le réel est proche de 1, on dit que l'unité (le neurone) est **active** alors que quand le réel est proche de 0, on dit que l'unité est **inactive.**

Le réel en question est appelé la **sortie** du neurone et sera notée y_c . Si la fonction d'activation est linéaire, le réseau de neurones se réduirait à une simple fonction linéaire.

En notant f la fonction d'activation, on obtient donc la formule donnant la sortie d'un neurone :

$$y_c = f(s) = f\left(\sum_{i=0}^{n} w_i x_i\right)$$

On remarque que le coefficient de biais est inclus dans la somme, d'où la formule plus explicite :

$$y_c = f(s) = f\left(\left(\sum_{i=1}^{n} w_i x_i\right) - w_0\right)$$

Il y a bien sûr beaucoup de fonctions d'activations possibles, c'est à dire répondant aux critères que nous avons donnés, toutefois dans la pratique il y en a principalement 2 qui sont utilisées (Figure 9).

- La fonction de Heaviside
- La fonction sigmoïde

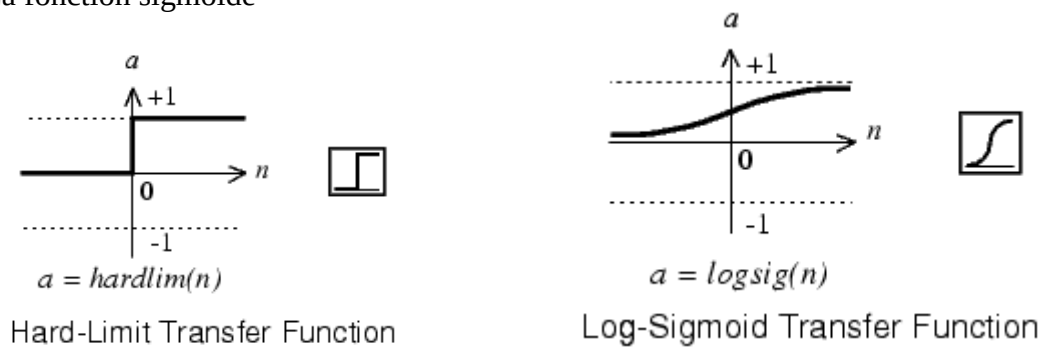


Figure 9 : Fonctions de transfert fréquemment utilisées.

V.2 Différents modèles de RNA

En fait, il existe deux grandes principales catégories des RNA qu'on va souligner dans ce paragraphe.

V.2.1 Réseaux de neurones directs (feedforward networks)

Les réseaux directs (feedforward) (Figure 10) : propagation vers l'avant de l'information, sont connectés dans un seul sens où chaque neurone d'une couche $(n-1)$ est connecté à tous les neurones de la couche n. Ce type de neurones comporte à son tour d'autres sous-types :

- Perceptron monocouche (une seule couche cachée)
- Perceptron multicouche (plusieurs couches cachées)
- Neurone "à base radiale" ou distance (dont La fonction de transfert de la couche cachée est une gaussienne).

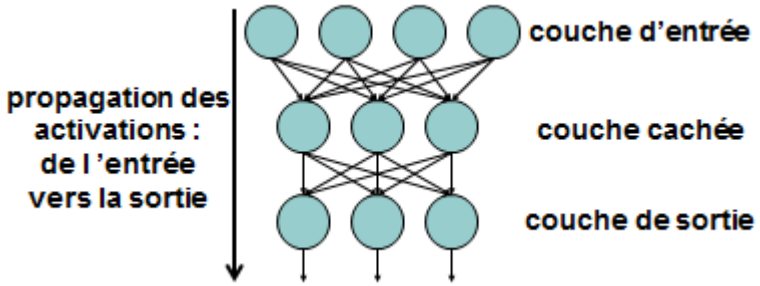


Figure 10: Réseaux directs (feedforward networks).

V.2.2 Réseaux de neurones récurrents (feedback networks)

Dans ce type de réseaux la sortie d'un neurone d'une couche n est utilisée comme entrée d'un neurone d'une couche précédente $(n-1)$ ou d'un neurone de la même couche n. Il comporte également quatre sous-types:

- Réseaux compétitifs
- Cartes de Kohonen
- Réseaux de hopfield
- Modèles ART.

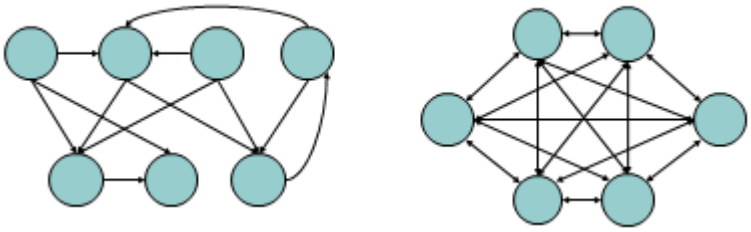


Figure 11: Réseaux récurrents (feedback networks).

V.3 Méthodes d'apprentissage ou d'entrainement

Comme le cerveau humain, les réseaux de neurones artificiels (RNA) peuvent apprendre par expérience.

L'apprentissage est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré.

Trois grandes classes d'apprentissage existent :

V.3.1 Apprentissage supervisé (back propagation)

Cet algorithme d'apprentissage (Figure 12) ne peut être utilisé que lorsque les combinaisons d'entrées-sorties désirés sont connues à priori. L'apprentissage est alors facilité et par là, beaucoup plus rapide que pour les deux autres algorithmes puisque l'ajustement des poids est fait directement à partir de l'erreur, soit la différence entre la sortie obtenue par le RNA et la sortie désirée.

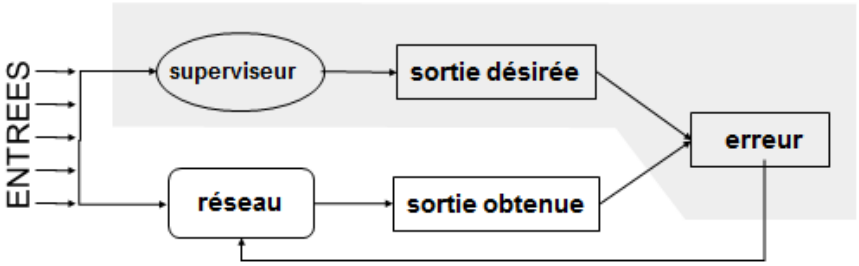


Figure 12: Processus d'apprentissage supervisé.

V.3.2 Apprentissage non-supervisé

Il n'y pas de connaissances à priori des sorties désirés pour des entrées données (Figure 13). En fait, c'est de l'apprentissage par exploration où l'algorithme d'apprentissage ajuste les poids des liens entre neurones de façon à maximiser la qualité de classification des entrées.

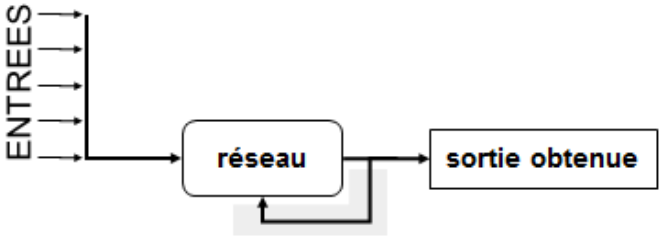


Figure 13: Processus d'apprentissage non-supervisé.

V.3.3 Apprentissage par renforcement

Dans ce cas, bien que les sorties idéales ne soient pas connues directement, il y a un moyen quelconque de connaître si les sorties du RNA s'approchent ou s'éloignent du but visé. Ainsi, les poids sont ajustés de façon plus ou moins aléatoire et la modification est conservée si l'impact est positif ou rejetée sinon.

V.4) Réseau de neurones monocouche : perceptron de Rosenblatt (1958)

Un perceptron est un neurone formel muni d'une règle d'apprentissage qui permet de déterminer automatiquement les poids synaptiques de manière à séparer un problème d'apprentissage supervisé. Le perceptron possède un nombre variable de neurones alignés verticalement constituant une seule couche, on parle alors d'un réseau de neurones monocouche (Figure 14).

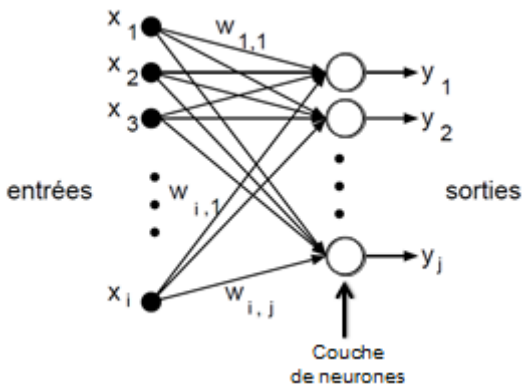


Figure 14: Schéma du perceptron de Rosenblatt.

V.4.1 Lois d'apprentissage

Durant la phase d'apprentissage d'un réseau de neurone on essaie d'améliorer la performance du RNA pour le rendre plus puissant et capable de résoudre le maximum de fonctions possibles, alors c'est dans cette perspective que réside le rôle des algorithmes d'apprentissage, qui consistent à modifier la valeur des poids w_{ij} entre les neurones ainsi que la valeur des biais.

En général, le poids d'une connexion i est modifié comme suit :

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n)$$

a. Loi de Hebb

Cette règle d'apprentissage supervisé très simple émet l'hypothèse que lorsqu'un neurone « i » est excité par un neurone « j » de façon répétitive ou persistante, le poids de leur connexion $w_{ij}(n)$ se modifie, soit il augmente soit il diminue par une quantité $\Delta w_{ij}(n)$: c'est une constante positive qui représente la force d'apprentissage (Learning rate).

$$w_i(k+1) = w_i(k) + \Delta w_i(k) \quad \text{où}$$
$$\Delta w_i(k) = \mu y_d(k) x_i(k)$$

μ : Représente le taux d'apprentissage

$y_d(k)$: La réponse désirée ou attendue par le réseau

$x_i(k)$: L'entrée du neurone i

b. Loi de Rosenblatt ou du perceptron

La règle de Hebb diverge dans certains cas, bien qu'une solution existe.

Une autre règle d'apprentissage a été proposée. Cette règle tient compte de l'erreur observée $e(k)$en sortie. C'est la règle du perceptron qui s'écrit sous la forme suivante :

$$w_i(k+1) = w_i(k) + \Delta w_i(k) \quad \text{où}$$
$$\Delta w_i(k) = \mu[y_{désirée}(k) - y_{calculée}(k)] x_i(k)$$
$$\Delta w_i(k) = \mu e(k) x_i(k)$$

c. Règle de Widrow-Hoff (1960) (delta-rule ou méthode des moindres carrées (LMS: Least Mean Square))

Cette loi est aussi une version modifiée de la loi de Hebb. Les poids des liens entre les neurones sont continuellement modifiés de façon à réduire la différence (le delta) entre la sortie désirée et la valeur calculée de la sortie du neurone.

$$w_i(k+1) = w_i(k) + \mu[d(k) - y(k)] x_i(k)$$

La loi de Widrow-Hoff est semblable à la loi d'apprentissage du perceptron.

V.4.2 Limites du perceptron

a. Séparation linéaire et condition d'approximabilité

Un perceptron prend en entrée un vecteur à plusieurs dimensions (1 par neurone) et opère une séparation entre ces données pour fournir une sortie. Grâce à cette séparation qu'il a construite entre les données, il sait, pour un nouvel exemple, quelle doit être la réponse.

Par exemple, si on a 2 classes en sortie (chat ou chien), on va entraîner le réseau à comprendre la différence entre les deux à partir des entrées. Le réseau se comportera ensuite comme une fonction « affine » et tracera une droite séparant les chats des chiens (Figure 15). Pour tout nouveau point, il aura juste à regarder de quel côté il est de la droite.

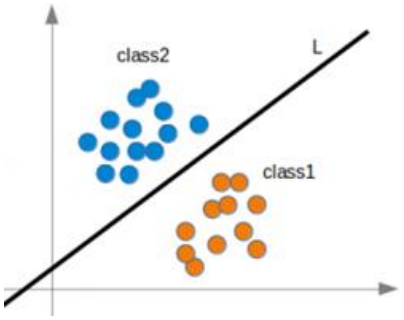


Figure 15 : Plan mettant une évidence une séparation linéaire entre deux classes.

b. Problème de la fonction XOR

Maintenant, on essaie de reproduire cette démarche de séparation linéaire pour une fonction appelée XOR. La fonction booléenne XOR renvoie 1 quand l'une de ses deux entrées vaut 1, 0 sinon. On essaie donc de placer les points pour lesquels x XOR y vaut 0 en orange et ceux pour lesquels x XOR y vaut 1 en bleu. On essaie maintenant de tracer une droite séparant les points oranges des points bleus par une seule couche de neurones. Nous verrons alors que c'est impossible. On voit donc les limitations du perceptron, il ne peut jamais apprendre une fonction qui n'est pas linéairement séparable mais si on ajoute plus de neurones au réseau, ce problème de séparation peut être surmonté (Figure 16).

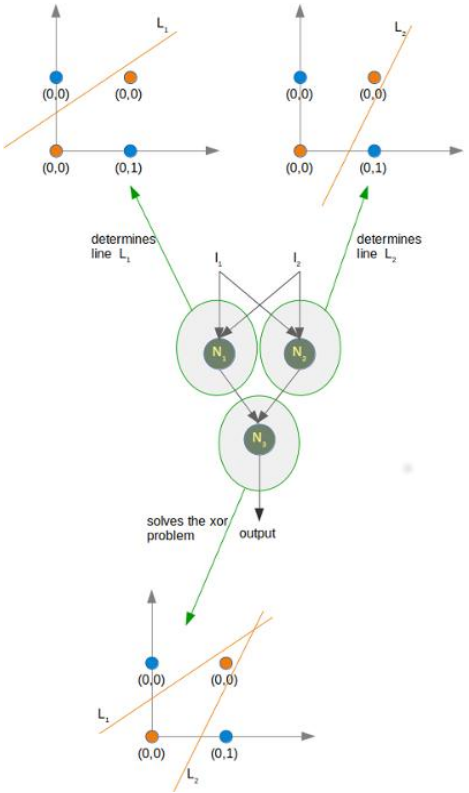


Figure 16: Solution simple du XOR : deux neurones identifient un cas particulier puis un autre neurone assemble leurs réponses.

V.5 Réseau de neurones multicouche : perceptron multicouches (MultiLayers Perceptron(MLP))

V.5.1 Architecture d'un réseau MLP

Comme nous avons déjà vu, le perceptron monocouche ne peut apprendre que dans le cas où les catégories à apprendre seraient linéairement séparables. Face à ces limites, vient l'idée de postuler un réseau MLP.

Le perceptron multicouche est la forme de réseau de neurones la plus couramment utilisée. Un tel réseau est constitué d'un minimum de trois couches de neurones. Les neurones d'une couche donnée ne sont connectés qu'aux neurones de la couche suivante (figure 17), ainsi l'activation des neurones est propagée à travers les différentes couches, de l'entrée vers la sortie.

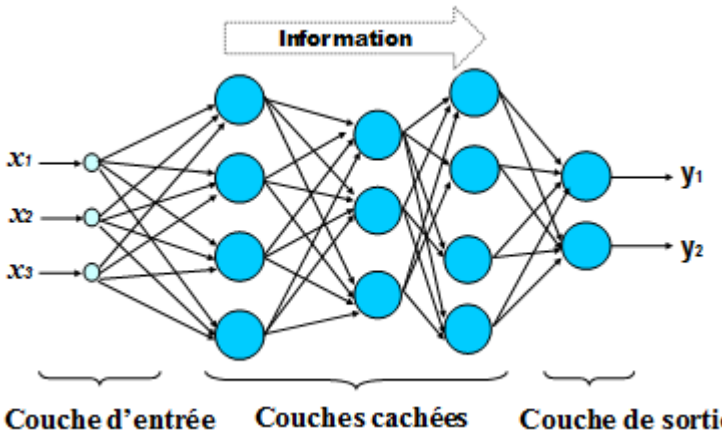


Figure 17: Perceptron multicouches (MultiLayers Perceptron (MLP)).

V.5.2 Processus d'apprentissage dans les réseaux MLP

a. Algorithme d'apprentissage

Un RNA apprend son fonctionnement et acquiert son savoir-faire à travers l'adaptation de ses paramètres (poids et biais) en appliquant un algorithme d'apprentissage approprié.

b. Algorithme de RétroPropagation du Gradient (RPG) (Backpropagation algorithm)

L'algorithme RPG un est algorithme d'apprentissage supervisé. Il se base sur une séquence de mesures des entrées-sorties (Figure 18).

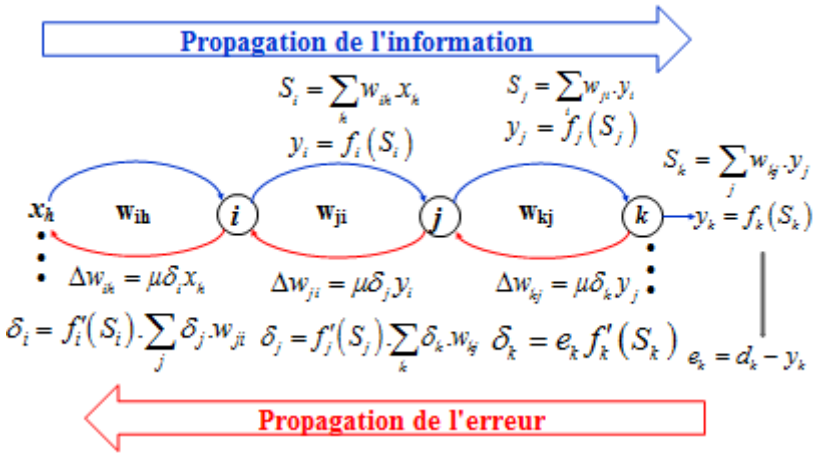


Figure 18 : Schéma explicatif de l'algorithme RPG.

L'algorithme RPG consiste à **mesurer l'erreur** entre les sorties désirées *d(n)* et les sorties observées *y(n).*
Il nécessite des **fonctions d'activations dérivables** des différentes couches de neurones.

L'algorithme RPG consiste à ajuster les poids et biais liés à chaque neurone par la rétropropagation de l'erreur calculée sur le neurone de sortie, tout en commençant par la dernière couche (couche de sortie) vers les couches inférieures jusqu'à arriver à la première (couche d'entrée) Alors ce processus d'apprentissage supervisé se diffère selon l'emplacement du neurone dans le réseau. On peut donc citer deux cas distincts :

- **Evaluation du gradient pour un neurone dans la couche de sortie**

Considérons un neurone de sortie *k* :

L'ajustement des poids se fait toujours à l'aide de l'équation :

$$w_{kj}(n+1) = w_{kj}(n) + \Delta w_{kj}(n)$$

La correction à appliquer peut s'écrire sous la forme:

$$\Delta w_{kj}(n) = \mu \delta_k(n) y_j(n)$$

où $\delta_j(n)$ est le gradient local défini par:

$$\delta_k(n) = e_k(n) f'_k(S_k(n))$$

Avec :

$y_j(n)$: sortie calculée par le neurone k.

μ : le taux d'apprentissage

$e_k = d_k - y_k$: erreur de sortie

$f'_k(S_k(n))$: dérivée de la fonction d'activation sur la sortie du neurone.

- **Evaluation du gradient pour un neurone dans la couche cachée**

Si le neurone *j* est dans une couche cachée du réseau, on n'a pas de sortie désirée, donc la formule d'ajustement des poids s'écrit sous la forme suivante :

$$\Delta w_{ji}(n) = \mu \delta_j(n) y_i(n)$$

où:

$$\delta_j(n) = f'_j\left(S_j(n)\right).\sum_k \delta_k(n).w_{kj}(n)$$

Notez bien, que dans le cas de la première couche cachée du réseau, puisqu'il n'y a pas de couche précédente de neurones, il faut substituer la variable $y_i(n)$ par l'entrée $x_i(n)$.

- Pour une unité *k* de la couche de sortie O:

$$\begin{aligned}\delta_k^{(O)}(n) &= f'_k\left(S_k(n)\right).e_k(n)\\ &= f'_k\left(S_k(n)\right).\left[d_k(n) - y_k(n)\right]\end{aligned}$$

- Pour une unité *i* de la couche cachée $(l-1)$:

$$\delta_i^{(l-1)}(n) = f'_i\left(S_i(n)\right).\sum_j \delta_j^{(l)}(n).w_{ji}(n)$$

V.5.3 Optimisation de l'algorithme

En effet, il existe plusieurs démarches de choix des paramètres ou de procédures qui permettent d'améliorer de façon significative la performance de l'algorithme de back-propagation,

a. Modes d'entrainement séquentiel et "batch"

- **Mode séquentiel (en-ligne (on line) ou stochastique)**

Dans ce mode, l'actualisation des poids est faite après la présentation de chaque exemple d'entraînement.

- plus rapide particulièrement lorsque l'ensemble de données d'entraînement est très grand et très redondant
- problèmes de convergence.

- **Mode "batch"**

Dans ce mode, l'actualisation des poids est faite après la présentation de l'ensemble des exemples d'entraînement.

- calculs et stockage plus lourds si trop d'exemples.

b. Correction des oscillations

La formule de correction des poids:

$$w_{ji}(n+1) = w_{ji}(n) + \mu\delta_j(n)y_i(n)$$

Ne tient compte que de l'erreur commise à l'instant n et ne tient pas compte des corrections précédentes. L´apprentissage ainsi conduit peut donner lieu à des oscillations des poids et ne se termine donc pas.

Pour corriger ces oscillations, on introduit un terme proportionnel à la dernière variation des poids:

$$w_{ji}(n+1) = w_{ji}(n) + \alpha\,\Delta w_{ji}(n-1) + \mu\delta_j(n)y_i(n)$$

où α, compris entre 0 et 1, est le **moment "momentum**"

La relation de correction des poids :

$$\Delta w_{ji}(n) = \alpha\,\Delta w_{ji}(n-1) + \mu\delta_j(n)y_i(n)$$

est appelé la "**règle delta généralisée**" (generalized delta rule).

Pour $\alpha=0$, on obtient la "**règle delta**".

V.5.4 Critères d'arrêt du processus d'apprentissage

Différents critères permettant de décider quand stopper le processus d'entraînement ou d'apprentissage:

- quand le **nombre d'itérations** fixé est terminé (c'est le critère le plus utilisé),
- quand la **norme du gradient** atteint un seuil faible fixé
- quand le **taux de variation de l'erreur quadratique moyenne par itération** est suffisamment faible.

V.5.7 Problème de généralisation

La généralisation consiste à tester le MLP, construit en minimisant une fonction de coût (erreur) sur un ensemble de données (ensemble de validation), qui n'ont pas été utilisées dans le processus d'apprentissage.

L'erreur de généralisation dépend de trois paramètres :

- Le nombre d'exemples utilisés pour l'apprentissage,
- Le nombre d'itérations utilisées
- L'architecture du réseau (le nombre de couches cachées et le nombre de neurones dans chaque couche "cachée"). Ceci dépend de la complexité du problème sous-jacent,

- **Nombre d'exemples utilisé pour l'apprentissage**

Pour une taille du RNA fixée, augmenter le nombre d'exemples n'améliorera la généralisation que jusqu'à une valeur asymptotique (Figure 19).

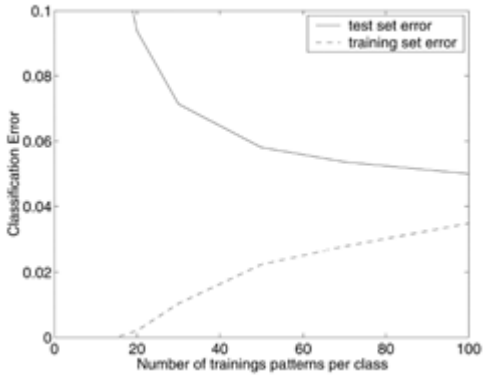


Figure 19: Influence du nombre de données sur l'erreur de généralisation.

- **nombre d'itérations utilisées**

Pour un nombre d'exemples d'apprentissage fixé *N* et une taille du MLP fixe, en augmentant le nombre d'itérations d'entraînement l'erreur de généralisation décroît jusqu'à une valeur critique puis, il commence à augmenter de nouveau (Figure 20).

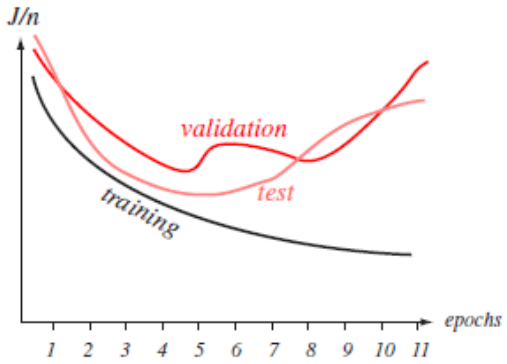


Figure 20: Influence du nombre d'itération (epochs) sur l'erreur de généralisation.

- **Architecture du réseau (nombre de neurones)**

Pour un nombre d'exemples d'apprentissage fixé *N*, si on augmente progressivement la taille du MLP (en augmentant le nombre de neurones), l'erreur de généralisation décroît jusqu'à une valeur critique puis, il commence à augmenter de nouveau (Figure 21).

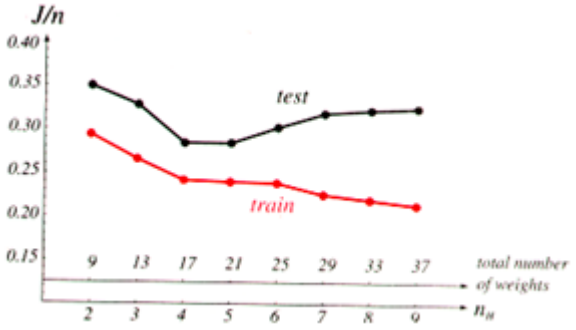


Figure 21: Influence du nombre de neurones.

V.5.8 Problème de sur- et sous-ajustement (Over and under-fitting)

- **Problème de sous-ajustement**

Un réseau trop simplifié (en terme de neurones et de connections) peut échouer dans l'apprentissage de la relation entrée-sortie dans le cas d'un processus complexe : c'est le problème de sous-ajustement (underfitting) (Figure 22).

- **Problème de sur-ajustement**

Un réseau trop complexe peut apprendre non seulement le signal entrée-sortie mais aussi les bruits présents dans les mesures : c'est le problème de sur-ajustement (overfitting) (Figure 22).

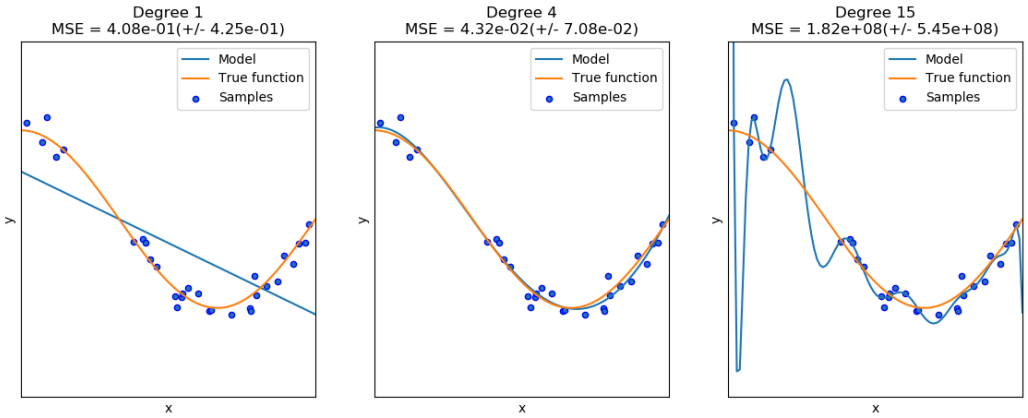


Figure 22: Problème de sur- et sous-ajustement.

Avec :

Degree1 : Problème de sous-ajustement

Degree15 : Problème de sur-ajustement

Degree4 : Bon ajustement (la fonction à résoudre est bien approchée par le modèle).

Remarque:

Le problème de sur-ajustement engendre deux erreurs :

- Une faible erreur sur les données d'entraînement
- Une forte erreur sur les données de validation

Alors, pour éviter ce problème, il faut utiliser un réseau de dimension juste suffisante et l'entraîner juste suffisamment pour bien apprendre et bien généraliser.

V.5.9 Approches constructive et destructive

Dans la pratique, deux approches peuvent être utilisées pour construire un RNA de dimension juste suffisante pour bien apprendre et bien généraliser :

- **Approche constructive :**

La dimension du RNA, minimale au départ, est augmentée, petit à petit, en ajoutant des neurones dans la couche cachée, jusqu'à ce que le RNA devienne capable d'apprendre sa tâche correctement (avec une erreur acceptable).

- **Approche destructive (pruning method)**

On commence par un RNA de dimension élevée et on élimine les connections moins importantes (une par une ou en %) jusqu'à ce que le réseau devienne capable d'apprendre sa tâche correctement.

V.5.10 Choix des données d'apprentissage et de validation

Le choix des données pour la phase d'apprentissage et pour la phase de validation constitue une étape indispensable pour le développement d'un RNA. Dans la phase d'apprentissage, le RNA s'entraine bien aux données préconisées pour celle-ci, durant laquelle il change ses paramètres (poids et biais) de façon à les ajuster pour que l'erreur de sortie soit minimale c.-à-d. bien approximer la fonction mise en question. Dans la phase de validation, il traite des nouvelles données qui n'ont pas été utilisées dans la phase d'apprentissage et qu'il doit les biens représentés vu la phase d'apprentissage. Il existe alors plusieurs méthodes qui permettent de bien choisir les données pour le traitement. Parmi lesquelles, on note :

a. Validation simple

Diviser les données disponibles en deux ensembles (les ensembles d'apprentissage et de validation), sans qu'une donnée ne soit commune (souvent on garde **2/3** des données dans l'ensemble d'apprentissage et **1/3** pour la validation).

Cette méthode n'est justifiable que lorsque le nombre *N* de données est très important (vis-à-vis de la dimension de l'espace d'entrée et de la complexité de la relation à approximer).

b. Split-samples

Réserve un troisième ensemble de données appelé ensemble de test, pour tester le réseau sur des données qui n'ont jamais été utilisées ni pour l'apprentissage ni pour la validation.

c. Validation croisée

L'ensemble des données de départ est découpé en *k* parties de taille égale. Le réseau est entraîné *k* fois, chaque fois en utilisant *k*-1 parties pour l'apprentissage et la dernière pour la validation.

Si *k* est de la taille de l'ensemble de départ, on parle de "leave-one-out" (car chaque apprentissage n'est validé que sur un seul exemple), sinon on parle de "leave-v-out".

La performance de généralisation du modèle appelée "score de validation croisée", est estimée en réalisant la moyenne quadratique des *k* erreurs obtenues.

VII .Conclusion

Plusieurs étapes doivent être franchies afin d'arriver à une estimation utile :

- Trouver un réseau (c.à.d. une famille paramétrique particulière de fonctions) capable d'approximer de manière satisfaisante la fonction à estimer ;
- Trouver un algorithme d'apprentissage (estimation par modification des paramètres) capable d'arriver à une bonne solution, et ayant une complexité non prohibitive ;
- Mener à bien l'apprentissage et évaluer la qualité de la solution obtenue (validation).

PARTIE II : ETUDE EXPERIMENTALE

CHAPITRE 1: Matériels et méthodes

I. Description de la zone d'étude

I.1) Contextes géographique, climatique et géologique

a. Géographie

La région de Skhirat, appartenant à la Meseta côtière, est limitée par la rivière Oued Ykem au nord-est, la rivière Oued Cherrat au sud et l'océan Atlantique à l'ouest. Il se trouve à environ 25 Km de Rabat, elle se caractérise par une production végétale intense basée sur l'irrigation à partir des eaux souterraines [9].

b. Climat

La zone d'étude fait partie de l'ex-région administrative de Rabat-Salé-Zemmour-Zaer : Il y a cinq stations météorologiques dans et autour d'elle. La station météorologique de l'aéroport de Rabat-Salé est la plus représentative du climat de Skhirat car elle a les plus longs records en particulier les précipitations et la température. La zone littorale de Rabat est caractérisée par une température moyenne annuelle de 17°C due à l'océan Atlantique, une température minimale mensuelle de 10 à 12°C en janvier et une température maximale mensuelle de 20 à 24°C en juillet / août, atteignant très exceptionnellement 32°C.

Les précipitations moyennes annuelles sont d'environ 600 mm; cependant, elles sont très variables et peuvent varier entre 250 et 800 mm.

II. Méthode d'échantillonnage de l'eau

II.1) Choix des sites de prélèvement

Après avoir défini la zone d'étude sur une carte topographique à échelle de 1/50 000, les points d'échantillonnage ont été identifié (Figure 23), en se basant sur deux critères principaux pour que les échantillons prélevés soient les plus représentatifs possible de la zone :

- Caractéristiques spatiales des différentes zones de la région de Skhirat
- Niveaux de salinité de l'eau d'irrigation

II.2) Méthode de prélèvement

L'échantillonnage de l'eau a été réalisé sur 70 puits situés aux alentours de la ville de Skhirat. Les échantillons d'eau ont été prélevés dans des bouteilles en polyéthylène de 1 litre qui ont été pré-nettoyées avec de l'acide chlorhydrique (HCl) concentré et de l'eau distillée et hermétiquement fermées, portant le code du site exploité (P_x) avec P se référant au puits et X le numéro du site.

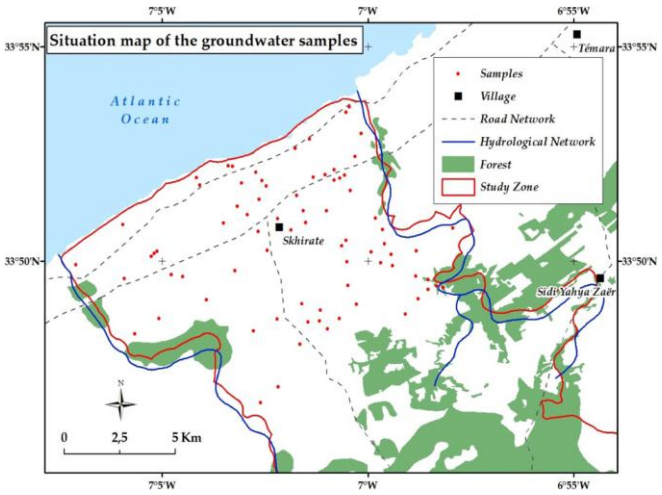


Figure 23: Localisation des sites d'échantillonnage de l'eau.

III. Analyse des paramètres physico-chimiques de l'eau in situ

Afin de caractériser la qualité chimique des eaux souterraines prélevées au niveau de la zone de Skhirat, certains paramètres physico-chimiques ont été mesurés in situ tel que :

III.1 pH de l'eau

Le pH de l'eau a été mesuré immédiatement après l'échantillonnage à l'aide d'un pH Metrohm mètre, modèle 691.

III.2 Conductivité électrique (CE)

La mesure de la CE de l'eau a été réalisée sur champs, en utilisant le conductimètre de laboratoire Orion, modèle 162.

III.3 Niveau piézométrique

Ce paramètre a été mesuré à l'aide d'une sonde piézométrique de 100m. Avoir une idée sur la profondeur de la nappe phréatique peut anticiper la qualité de l'eau. Plus le niveau de la nappe s'élève (s'approche du sol), plus elle est exposée à la pollution due aux activités humaines, notamment celle de l'agriculture qui utilise de nombreux engrais et pesticides que les eaux de pluies vont emmener dans les réservoirs souterrains. L'industrie peut également entrainer une contamination de la nappe par les fuites d'ordures, les métaux lourds et les retombées atmosphériques des fumées. En revanche, si le niveau de la nappe descend trop bas (sous le niveau de la mer), les écoulements d'eau s'inversent (de la mer à la terre, et non pas de la terre à la mer). Ceci entraine à son tour une autre pollution de type saline.

IV. Analyse des paramètres physico-chimiques de l'eau au laboratoire

Afin d'apprécier la qualité de l'eau des puits analysée et particulièrement sa salinité, les analyses suivantes ont été réalisées au laboratoire (Figure 24) :

Figure 24: Appareils des analyses physico-chimiques de l'eau au laboratoire.

IV.1Dosage du sodium (Na^+) et du potassium (K^+)

Le dosage de cet élément s'effectue par l'appareil photomètre à flamme [8].

Les échantillons sont préparés dans des bêchers portant le code (Px) référant aux numéros des sites et aux puits.

On commence à lancer l'analyse par photomètre à flamme, si l'écran affiche « over », on procède à une dilution immédiate de la solution afin de réduire sa concentration et la ramener à la zone de linéarité (zone dont laquelle les résultats affichés sont fiables et exacts). L'appareil photomètre à flamme aspire simultanément l'eau contenu dans le bêcher par une pompe péristaltique vers un système intérieur de gaz vecteur qui l'emporte ensuite dans une flamme de gaz butane ou le processus d'excitation a lieu suite à un chauffage relativement faible des atomes, les atomes de Na^+ et de K^+ possèdent donc une certaine énergie qui leur permettent de passer à un niveau plus haut (état excité), à cet état les atomes se trouvent gênés, instables. Pour retourner à l'état stable (état fondamental), ils cèdent cette énergie sous forme de longueurs d'onde qui vont être détectées par les filtres optiques spécifiques. Les longueurs d'onde émises sont proportionnelles à la quantité d'analyte (Na^+ et K^+) dans l'échantillon.

L'appareil affiche la quantité d'analyte en méq/l, pour l'exprimer en mg/l, on fait la conversion suivante :

$$Na^+ \text{ (mg/l)} = Na^+ \text{ (méq/l)} * 23$$

$$K^+ \text{ (mg/l)} = K^+ \text{ (méq/l)} * 39$$

b. Dosage du calcium (Ca^{2+}) et du magnésium (Mg^{2+})

Le dosage du Ca^{2+} et du Mg^{2+} est volumétrique de type compléxométrie.

Pour le dosage du calcium, on met dans un erlenmeyer de 250ml notre échantillon d'eau (10ml) et on complète avec l'eau distillée jusqu'à 100ml. On ajoute 1ml de solution de triéthanolamine, puis 5gouttes de KCN à 5%, on agite bien la solution et on laisse reposer 5 minutes, après on verse 20ml de solution de NaOH 5N (jusqu'à pH 12) tout en homogénéisant la solution. En fin, on procède à un titrage par E.D.T.A 0,02N en présence d'une pincée d'indicateur calconecarbonique, jusqu'à virage du violet au bleu franc.

Pour le dosage du magnésium, on adopte le même mode opératoire de celui du calcium, sauf ici, au lieu d'ajouter 20ml de NaOH, on ajoute 20ml de solution tampon NH_4 OH-NH_4 CL (jusqu'à pH 10) l'indicateur utilisé pour doser le magnésium est le Noir Eriochrome T.

c. Dosage des carbonates (CO_3^{2-}) et bicarbonates (HCO_3^-)

Pour le dosage des carbonates, on prélève une aliquote de 10ml dans un erlenmeyer de 250ml, on ajoute de 2 à 3gouttes de phénolphtaléine. S'il ne se développe pas une coloration rose, les carbonates sont absents, si une coloration rose apparait, on verse à la burette HCl 0,05N jusqu'à disparition de la coloration.

Pour le dosage des bicarbonates, on ajoute au même prélèvement (qu'il y ait ou non coloration rose) 2 gouttes de méthylorange (ou l'hélianthine) puis on poursuit le titrage avec HCl jusqu'à ce qu'il y ait virage vers le jaune-orangé.

d. Dosage des sulfates (SO_4^{2-})

Pour le dosage des ions sulfates, on prélève 10ml de l'échantillon d'eau dans un erlenmeyer de 250ml, on verse ainsi V ml de HCl nécessaire pour doser les carbonates et les bicarbonates, on porte le mélange à l'ébullition pendant 5 minutes puis on ajoute goutte à goutte 10ml exactement mesurés de $BaCl_2$, après on fait bouillir pendant 5 minutes et on laisse refroidir. Juste avant le titrage, on ajoute 2ml de $MgCl_2$, puis 20ml de tampon (NH_4OH-NH_4Cl), le titrage est réalisé par E.D.T.A en présence de noir Eriochrome T jusqu'à virage du violet au bleu franc.

On opère dans les mêmes conditions avec un témoin réalisé avec de l'eau distillée.

e. Dosage des chlorures (Cl^-)

Pour le dosage des ions chlorures, on met dans un erlenmeyer de 250ml, 10ml d'échantillon d'eau puis on ajoute de 2 à 3 gouttes de chromate de potassium K_2CrO_4 à 10%, le titrage s'effectue au moyen d'une burette avec une solution de nitrate d'argent $AgNO_3$ 0,02N. La fin de la réaction est décelée par l'apparition d'une teinte rougeâtre.

V. Normes et critères utilisées à l'évaluation de la qualité chimique de l'eau d'irrigation

Pour évaluer la qualité chimique de l'eau d'irrigation, on s'intéresse principalement à deux critères ou deux risques qui sont [12] :

V.1 Risque de Salinité :

La salinité de l'eau est exprimée en termes de CE (Conductivité Electrique).

La teneur en sel dans l'eau d'irrigation.

Comme ça a été déjà remarqué, l'excès de teneur en sel est l'un des soucis principaux avec l'eau utilisée pour l'irrigation. Pour cela, il est indispensable de mesurer la salinité de l'eau avant de l'utiliser pour les cultures, ainsi d'évaluer le degré de cette salinité. Pour ce faire, certaines normes américaines ont déjà proposé une classification de l'eau d'irrigation selon son degré de salinité (Tableau 2) :

Tableau 2: Répartition de la salinité de l'eau d'irrigation selon la Norme USDA.

Classe de salinité	*Symbole*	*EC (dS/m)*
Non saline	C1	< 0.25
Moyennement saline	C2	0.25–0.75
Très saline	C3	0.75–2.25
Très hautement saline	C4	2.25–5
Extrêmement saline	C5	>5

V.2 Risque d'alcalinité :

L'alcalinité de l'eau d'irrigation est mesurée en utilisant le taux d'adsorption de sodium (SAR). Plus le SAR est élevé, plus le risque de sodicité de l'eau est élevé en raison d'échange qui aura lieu dans l'équilibre entre Na^+ de la solution du sol et Ca^{2+}/ Mg^{2+} du complexe adsorbant. Comme ça a été

déjà indiqué dans la partie qualité d'eau d'irrigation, une teneur en sodium plus élevée dans l'eau réduit la perméabilité, et par conséquent, moins d'eau est disponible pour la plante, etc.

Il est donc primordial de juger le degré de ce risque d'alcalinité en classifiant les eaux selon leur teneur en sodium (Tableau 3).

Tableau 3: Répartition de l'alcalinité de l'eau d'irrigation selon la Norme USDA.

Classe d'alcalinité	*Symbole*	*SAR*
Eau excellente	S1	< 10
Eau bonne	S2	10–18
Eau moyenne	S3	18–26
Eau mauvaise	S4	>26

Le rapport d'absorption du sodium (RAS) sert donc à évaluer la qualité de l'eau d'irrigation pour prévenir ce problème. Le RAS se calcule au moyen de la formule suivante :

$$SAR = \frac{Na}{\sqrt{\frac{Ca + Mg}{2}}}$$

Avec :

Na : teneur en élément Na^{+} exprimé en mg/l

Ca : teneur en élément Ca^{2+} exprimé en mg/l

Mg : teneur en élément Mg^{2+} exprimé en mg/l

V.3 Diagramme de classification américain :

L'utilisation du diagramme de classification américain permet l'attribution à chaque puits d'eau une qualité en termes de risques de salinisation et d'alcalinisation lorsque l'eau est utilisé pour l'irrigation

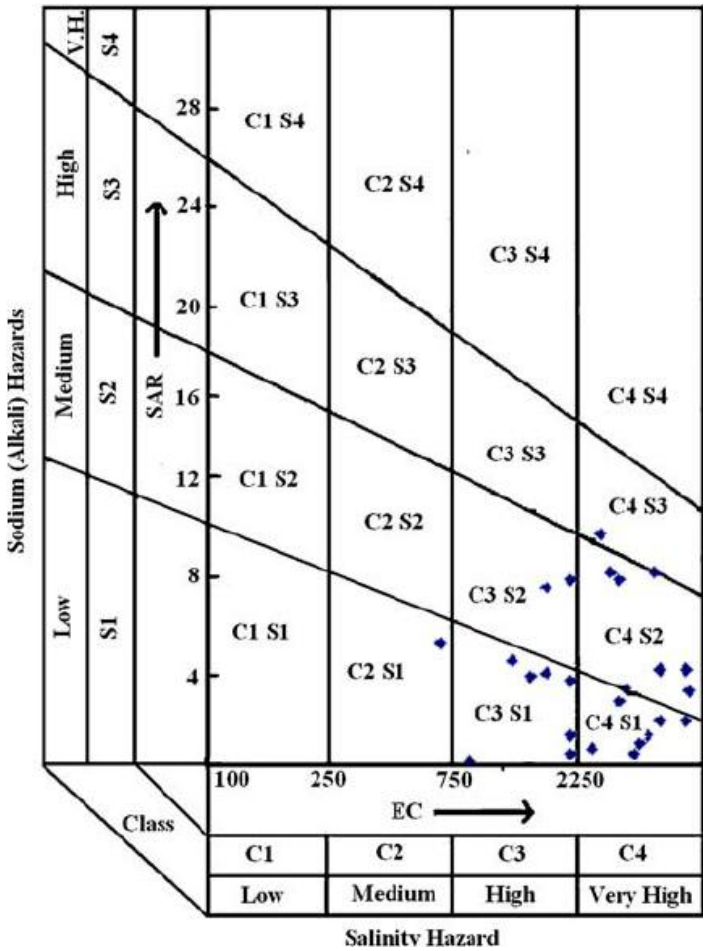


Figure 25 : Diagramme de détermination de la qualité de l'eau.

VI. Logiciel de traitement des données (IBM SPSS Statistics version 20)

Le logiciel **SPSS** (Statistical Package for the Social Sciences) est un logiciel utilisé pour l'analyse statistique [13]. C'est un système complet d'analyse des données, car il offre des informations complètes, cohérentes et précises, ainsi il permet d'exécuter plusieurs fonctions statistiques parmi lesquelles :

- Statistique descriptive : Cross tabulation, Fréquences, Descriptives, Explore, Descriptive Ratio Statistics
- Statistique bivariée : Moyennes, test t, ANOVA, Corrélation (bivariée, partielle, distances), tests non paramétriques
- Prédiction pour numérique outcomes : régression linéaire
- Prédiction pour groupes identifiants : Analyse factorielle, analyse de groupe (deux pas, K-moyennes, hiérarchique), analyse discriminante (en marketing)
- Régression logistique
- Réseaux de neurones artificiels,…

Chapitre 2 : Résultats et discussion

I. Analyse exploratoire des données

II.1 Statistique descriptive des données

La statistique descriptive (Tableau 4) constitue une étape préliminaire et indispensable pour chaque traitement statistique qui a pour objectif de résumer l'information contenue dans un échantillon. Elle permet de décrire la façon dont les résultats des observations se distribuent en utilisant différentes paramètres numériques telle que les paramètres de position (la moyenne, le mode et la médiane), les paramètres de dispersion (l'écart-type et la variance) ainsi que le coefficient de variation. Après avoir effectué dix analyses physico-chimiques sur chaque échantillon prélevé à partir de la zone d'étude, on a récolté les résultats d'analyses sous-forme d'une matrice pour pouvoir appliquer par la suite les deux méthodes statistiques (RNA et RLM).

Tableau 4: La statistique descriptive des données des eaux de puits de Skhirat.

	CE	***pH***	***Ca^{2+}***	***Mg^{2+}***	***Na^{+}***	***K^{+}***	***Cl^{-}***	***SO_4^{2-}***	***HCO_3^{-}***	***CO_3^{2-}***
Moyenne	3,17	7,45	7,41	4,98	26,45	0,15	21,21	8,25	5,09	0,84
Médiane	2,74	7,47	4,70	3,50	22,78	0,08	15,45	6,15	5,11	0,83
Ecart-type	1,72	0,31	6,20	5,30	11,75	0,16	15,00	6,95	1,58	0,52
CV(%)	54,24	4,12	83,64	106,43	44,42	112,43	70,69	84,23	31,11	62,12
Min	0,90	6,20	0,48	0,10	13,68	0,01	2,72	0,00	1,35	0,00
Max	8,08	8,24	26,80	27,80	76,43	1,09	63,40	32,02	9,73	2,05

A partir des résultats de la statistique descriptive **(Tableau 4)**, il est bien évident que les variables pH, HCO_3^- et CO_3^{2-} ont la valeur de la moyenne plus proche de celle de la médiane. Donc ces variables sont probablement symétriques. Les autres variables ont une valeur de la moyenne supérieure à celle de la médiane, donc sont asymétrique.

On remarque ainsi que les variables CE, pH, K^+, HCO_3^- et CO_3^{2-} ont une faible dispersion des données autour de la moyenne (écart-type très faible), contrairement aux variables Ca^{2+}, Mg^{2+} , SO_4^{2-}, Na^+ et Cl^- qui ont une dispersion moyenne des données.

Les valeurs élevées du coefficient de variation pour les trois variables Mg^{2+}, K^+ et NO_3^- montrent la grande dispersion de ces valeurs.

La statistique descriptive permet de montrer aussi la différence entre la valeur maximale et la valeur minimale pour tous les variables.

A partir des résultats qu'on a trouvé on constate que les variables pH, K^+ et CO_3^{2-} ont des amplitudes (E=max-min) petites c.à.d. qu'il n'y a pas une grande différence entre le maximum et le

minimum de ces variables. Les variables Ca^{2+}, Mg^{2+}, Na^{+}, Cl^{-} et SO_4^{2-} ont des amplitudes plus ou moins élevées que les autres variables.

II.2 Analyse de corrélations

L'analyse de corrélation entre les variables est basée sur un test de significativité de la corrélation de Pearson c.à.d. le coefficient de corrélation r est-il significativement différent de zéro ou non ?

Le test s'écrit :

$$\begin{cases} H_0 : r = 0 \\ H_1 : r \neq 0 \end{cases}$$

On peut accepter ou rejeter l'hypothèse nulle H_0 selon la probabilité observée ($\alpha_{observée}$) par rapport à la valeur seuil ou la probabilité critique ($\alpha_{critique} = 5\%\ soit\ 0{,}05$), si l'hypothèse nulle est rejetée, on dit que le coefficient de corrélation est significatif c.à.d. que sa valeur est suffisamment différente de 0 pour pouvoir décrire une relation entre les variables. Si, elle est acceptée, on dit que le coefficient de corrélation n'est pas significatif c.à.d. il n'y a pas de relation entre les variables.

On rejette l'hypothèse H_0 si $\alpha_{observée} < \alpha_{critique}$.

D'après les résultats d'analyse de corrélation **(Tableau 5)**, on a :

- L'existence d'une corrélation forte et positive entre la variable CE et les variables Ca^{2+} (Figure 29), Mg^{2+}, Na^{+} (Figure 30), K^{+}, Cl^{-}(Figure 31) et SO_4^{2-}.
- La variable SAR est :

  Fortement corrélée et négativement corrélée avec la variable Ca^{2+} (Figure 26).

  Fortement corrélée et positivement corrélée avec la variable Na^{+} (Figure 27).

  Moyennement corrélée et négativement corrélée avec la variable Mg^{2+} (Figure 28).

  Moyennement corrélée et positivement corrélée avec la variable CO_3^{2-}.
- Par ailleurs la variable pH est fortement corrélée et négativement corrélée avec les variables CE, Ca^{2+}, Mg^{2+}, Cl^{-} et SO_4^{2-}.
- Les variables CE, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-} et SO_4^{2-} sont fortement et positivement corrélées.

Tableau 5: Corrélations entre toutes les variables.

v	CE	pH	Ca2+	Mg2+	Na+	K+	Cl-	SO4--	HCO3-	CO3--	SAR
CE	1,00	***-0,32***	***0,72***	***0,73***	***0,91***	***0,47***	***0,98***	***0,83***	0,07	0,06	0,09
		0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,57	0,62	0,47
pH	***-0,32***	1,00	***-0,38***	***-0,35***	-0,18	***-0,10***	***-0,32***	***-0,36***	0,21	0,03	0,19
	0,01		0,00	0,00	0,13	0,43	0,01	0,00	0,08	0,80	0,12
Ca2+	***0,72***	***-0,38***	1,00	***0,62***	***0,43***	0,23	***0,77***	***0,64***	-0,13	-0,19	***-0,48***
	0,00	0,00		0,00	0,00	0,06	0,00	0,00	0,27	0,11	0,00
Mg2+	***0,73***	***-0,35***	***0,62***	1,00	***0,64***	***0,36***	***0,74***	***0,83***	0,07	-0,03	**-0,26**
	0,00	0,00	0,00		0,00	0,00	0,00	0,00	0,59	0,83	0,03
Na+	***0,92***	-0,18	***0,43***	***0,64***	1,00	***0,53***	***0,88***	***0,78***	0,16	0,14	***0,36***
	0,00	0,13	0,00	0,00		0,00	0,00	0,00	0,18	0,24	0,00
K+	***0,47***	-0,10	0,23	***0,36***	***0,53***	1,00	***0,49***	***0,36***	0,07	-0,03	0,14
	0,00	0,43	0,06	0,00	0,00		0,00	0,00	0,54	0,84	0,26
Cl-	***0,98***	***-0,32***	***0,77***	***0,74***	***0,88***	***0,49***	1,00	***0,78***	-0,04	0,02	0,03
	0,00	0,01	0,00	0,00	0,00	0,00		0,00	0,75	0,88	0,78
SO4--	***0,83***	***-0,36***	***0,64***	***0,83***	***0,78***	***0,36***	***0,78***	1,00	0,10	-0,07	-0,07
	0,00	0,00	0,00	0,00	0,00	0,00	0,00		0,43	0,57	0,57
HCO3-	0,07	0,21	-0,13	0,07	0,16	0,07	-0,04	0,10	1,00	0,21	0,20
	0,57	0,08	0,27	0,59	0,18	0,54	0,75	0,43		0,08	0,10
CO3--	0,06	0,03	-0,19	-0,03	0,14	-0,03	0,02	-0,07	0,21	1,00	**0,27**
	0,62	0,80	0,11	0,83	0,24	0,84	0,88	0,57	0,08		0,03
SAR	0,09	0,19	***-0,48***	**-0,26**	***0,36***	0,14	0,03	-0,07	0,20	**0,27**	1,00
	0,47	0,12	0,00	0,03	0,00	0,26	0,78	0,57	0,10	0,03	

Gras italique: La corrélation est significative au niveau 0.01 (bilatéral).

Italique : La corrélation est significative au niveau 0.05 (bilatéral).

Première ligne : coefficient de corrélation de Pearson.

Deuxième ligne : probabilité $\alpha_{observés}$ indiquant le degré de significativité du test.

II. Application de la RLM

Comme ça a été déjà indiqué, deux risques sont à prendre en considération pour évaluer la qualité chimique de l'eau à savoir l'alcalinité (SAR) et la salinité (CE). Pour ce faire, on va générer donc deux modèles de régression, le premier pour le SAR et le deuxième pour la CE.

II.1 SAR

Afin d'obtenir un modèle de régression exhaustif qui peut bien expliquer et prévoir le degré d'alcalinité, on fait introduire le SAR comme variable dépendante en fonction des autres quatre variables qui lui sont corrélées (Ca^{2+}, Mg^{2+}, Na^{+} et CO_3^{2-}) (variables indépendantes).

Variables introduites

La méthode de régression utilisée ici, est celle du pas à pas (stepwise), elle consiste à introduire les variables qui présentent une forte corrélation avec la variable dépendante de façon successive (l'une après l'autre) d'où vient le nom du pas à pas.

Après le traitement des données par la RLM à l'aide du logiciel SPSS, le modèle final a retenu trois variables : il a introduit tout d'abord le Ca^{2+}, ensuite le Na^{+} et enfin le Mg^{2+} sur la base du critère de sélection (la probabilité F est significative à $p < 0,05$).

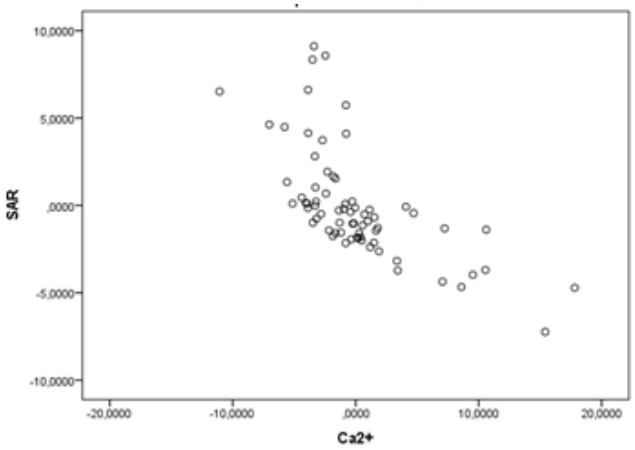


Figure 26 : Corrélation entre le SAR et le Ca^{2+}.

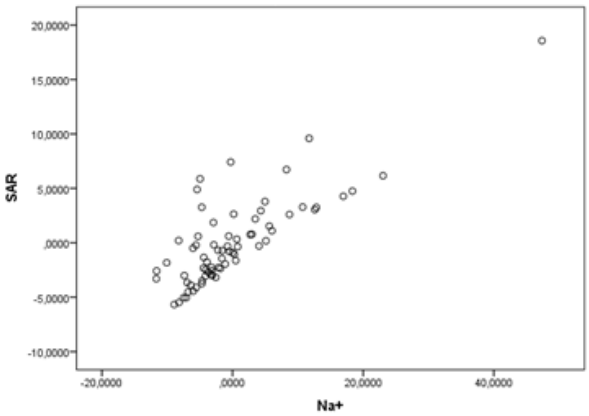


Figure 27 : Corrélation entre le SAR et le Na^{+}.

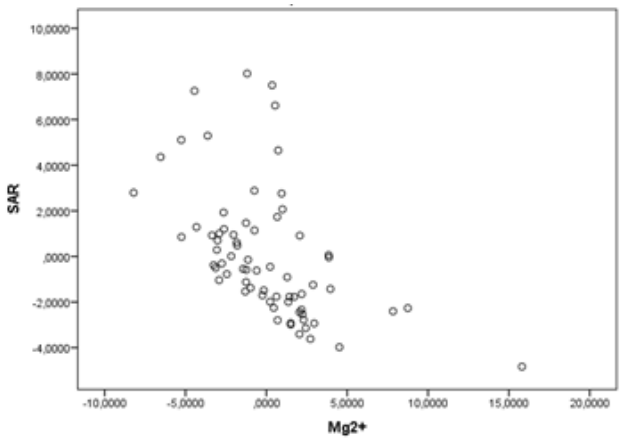


Figure 28 : Corrélation entre le SAR et le Mg^{2+}.

Étape 1 : Évaluation de la qualité du modèle de régression

Tout comme la régression simple, l'interprétation débute en évaluant la qualité du modèle. On vérifie si la première étape du modèle explique significativement plus de variabilité qu'un modèle sans prédicteur. Ensuite, il s'agit de s'assurer que toutes les variables introduites contribuent à améliorer significativement la variabilité expliquée par le modèle final.

***Analyse de variance**

Le tableau d'ANOVA (Tableau 6) nous donne cette information. Il nous permet de déterminer si nous rejetons l'hypothèse nulle (H_0) ou non. Dans notre exemple, nous voulons savoir dans un premier temps si l'élément Ca^{2+} prédit mieux le SAR que ne le fait un modèle sans prédicteur (avec seulement la moyenne), dans un deuxième temps si les deux éléments Ca^{2+} et Na^{+} prédisent mieux le SAR qu'un modèle sans prédicteur et dans un troisième temps si les trois éléments Ca^{2+}, Na^{+} et Mg^{2+} prédisent mieux le SAR qu'un modèle sans prédicteur. L'hypothèse nulle est donc que les trois modèles sont équivalents à la moyenne du salaire.

Tableau 6: ANOVA.

Modèle	*Somme des carrés*	*ddl*	*Moyenne des carrés*	*F*	*Signification*
Régression	1064,46	3	354,82	59,57	0,00[d]
Résidu	393,12	66	5,96		
Total	1457,58	69			

d. Valeurs prédites : (constantes), Ca^{2+}, Na^{+}, Mg^{2+}.

Ici, on a considéré seulement le dernier modèle (troisième bloc) qui tient compte de trois variables car il a la plus grande valeur de F qui est hautement significative comparée au seuil critique de 0,05, ce qui indique que nous avons moins de 0,1 % de chance de se tromper en affirmant que le modèle contribue à mieux prédire le SAR que la simple moyenne.

Étape 2 : Évaluation de l'ajustement du modèle de régression aux données

Maintenant que l'on sait que le modèle est significatif, le tableau récapitulatif des modèles (Tableau 7) permet de déterminer la contribution du troisième bloc de variables. Ce tableau indique le R^2 cumulatif à chaque étape du modèle (colonne R^2).

Tableau 7: Récapitulatif des modèles.

Modèle	*R*	*R^2*	*R^2 ajusté*	*Changement dans les statistiques*				
				Variation de R^2	Variation de F	ddl1	ddl2	Signification Variation de F
3	0,85[c]	0,73	0,72	0,11	25,72	1	66	0

c. Valeurs prédites : (constantes), Ca^{2+}, Na^{+}, Mg^{2+}.

Le tableau contient donc plusieurs informations utiles. Premièrement, la valeur de la corrélation multiple (R) correspond à l'agglomération des points dans la régression simple. Elle représente la force de la relation entre la variable dépendante (SAR) et la combinaison des variables indépendantes (Ca^{2+}, Na^{+}, Mg^{2+}) de chaque modèle. Une valeur de 0,85 suggère que les données sont ajustées de manière très satisfaisante au modèle.

Ensuite, la signification du R^2 est évaluée en fonction de l'apport de chaque étape. La variation de F associée au troisième modèle est significative ($p < .001$). Ce modèle explique donc une proportion significative de la variance de la variable SAR. Le troisième modèle a pour valeur de R^2 0,730 et cette variation qui vaut comme valeur 0,105 apparait comme significative. En effet, la valeur de F est calculée à partir de la variation du R^2 entre les étapes. SPSS détermine donc si la différence (0,105) entre le R^2 du dernier modèle et celui du premier modèle est significative. Cette fois-ci, c'est le cas ($p < 0,001$).

La troisième étape contribue donc significativement à l'amélioration de l'explication de la variabilité de la variable dépendante (SAR).

Étape 3 : Évaluation de la variabilité expliquée par le modèle de régression

Enfin, on se rappelle que la valeur du R^2, lorsqu'elle est multipliée par 100, indique le pourcentage de variabilité de la variable dépendante expliquée par le modèle (les prédicteurs). Les résultats suggèrent que 73% de la variabilité de SAR est expliquée par la combinaison de trois éléments Ca^{2+}, Na^{+} et Mg^{2+}.

Étape 4 : Estimation des paramètres du modèle

Maintenant que nous savons que notre modèle est significatif et que le troisième est celui qui explique le plus de variance, il est possible de construire l'équation de régression pour prédire une valeur de y (SAR). L'équation de base était la suivante :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_p x_{ip} + \varepsilon_i$$

Tableau 8: Coefficients du modèle.

Modèle		*Coefficients non standardisés*		*Coefficients standardisés*	t(Student)	*Signification*
		A	Erreur standard	Bêta		
3	(Constante)	8,26	0,77		10,72	0,00
	Ca2+	-0,43	0,06	-0,58	-7,09	0,00
	Na+	0,36	0,03	0,92	11,03	0,00
	Mg2+	-0,42	0,08	-0,48	-5,07	0,00

Remplaçons maintenant les β par les coefficients fournis dans le tableau ci-dessous :

$$SAR_{prédite} = 8{,}26 - 0{,}43\ Ca^{2+} + 0{,}36\ Na^{+} - 0{,}42\ Mg^{2+}$$

Le signe du coefficient nous indique le sens de la relation. Dans notre cas, plus les éléments Ca^{2+} et Mg^{2+} augmentent plus le SAR diminue. Nous voyons aussi que quand le Na^{+} augmente le SAR également à tendance d'augmenter.

Le coefficient nous informe également sur le degré auquel chaque prédicteur influence la variable dépendante si tous les autres prédicteurs sont constants.

L'erreur standard nous renseigne sur la variabilité du coefficient dans la population. Elle permet également de calculer la valeur de t. Cette dernière nous indique si le coefficient est significatif. Alors que le tableau sur le récapitulatif des modèles confirmait si chaque modèle était significatif, la signification de t nous permet de répondre à la question « est-ce que le b du prédicteur est différent de 0 ? », donc si chaque variable contribue significativement au modèle. Plus la valeur de t est élevée et plus celle de p est petite, plus le prédicteur contribue au modèle. Nous constatons donc que les trois variables sont significatives, mais que la variabilité expliquée par l'élément Na^{+} est plus importante que celle expliquée par les deux éléments Ca^{2+} et Mg^{2+}.

La valeur du Bêta standardisé (β) apporte aussi une information intéressante. Elle indique le changement en écart-type de la variable dépendante pour chaque augmentation d'un écart-type de la variable indépendante quand toutes les autres valeurs sont constantes.

a. Prédiction du $SAR_{binaire}$ et erreur d'estimation

Après avoir postulé et déterminé le modèle final de régression, il nous reste qu'à faire la prédiction désirée. Dans un premier temps on va prédire la valeur du SAR en fonction des trois variables mises en jeu (Ca^{2+}, Na^{+} et Mg^{2+}). Suite à cette valeur prédite, on calcule deux autres valeurs de SAR qui vont nous permettre de juger la performance du modèle pour évaluer la qualité de l'eau d'irrigation tout en les comparant avec celles observées sur le tableau de donnée (on comparant le $SAR_{binaire\ prédit}$ avec le $SAR_{binaire\ observé}$ et le $SAR_{ordinal\ prédit}$ avec le $SAR_{ordinal\ observé}$).

Le $SAR_{binaire\ prédit}$ est calculé à partir de la valeur de $SAR_{prédit}$ (valeur prédite par la droite de régression) sous la condition suivante :

Si, $SAR_{prédit} \leq 18 \longrightarrow SAR_{binaire\ prédit} = 1$ avec : $1 \longrightarrow$ Eau excellente

Sinon $\longrightarrow SAR_{binaire\ prédit} = 0$ $0 \longrightarrow$ Eau mauvaise

Afin de juger l'aptitude du modèle à évaluer correctement la qualité de l'eau, on recourt à la comparaison entre les deux valeurs de SAR, on obtient donc un tableau croisé issu de cette comparaison :

Tableau 9: Tableau croisé sar bin * SARbnPrd.

Effectif		*SARbnPrdt*		*Total*
		0	1	
sar bin	0	1	7	8
	1	1	61	62
Total		2	68	70

D'après le tableau ci-dessus, il est bien clair que le modèle de régression a correctement estimé la qualité de l'eau pour ((1+61)/70)*100 = 88.6% des puits. Cependant, ce modèle a commis deux types d'erreurs. La première erreur concerne le classement de 7 puits (7/70)*100 = 10% comme étant de bonne qualité alors qu'en réalité ils sont de mauvaise qualité alors que la seconde erreur concerne le classement d'un puits (1/70)*100 = 1.4% comme étant de mauvaise qualité alors qu'en réalité il est bonne qualité.

b. Prédiction du $SAR_{ordinal}$ et erreur d'estimation

De la même façon que le $SAR_{binaire\ prédit}$, on va calculer le $SAR_{ordinal\ prédit}$.

Le $SAR_{binaire\ prédit}$ est calculé à partir de la valeur de $SAR_{prédit}$ (valeur prédite par la droite de régression) sous la condition suivante :

Si, $SAR_{prédit} \leq 10$	$\longrightarrow$ $SAR_{ordinal\ prédit} = 1$	Avec : 1 $\longrightarrow$	Eau excellente
$10 < SAR_{prédit} \leq 18$	$\longrightarrow$ $SAR_{ordinal\ prédit} = 2$	2 $\longrightarrow$	Eau bonne
$18 < SAR_{prédit} \leq 26$	$\longrightarrow$ $SAR_{ordinal\ prédit} = 3$	3 $\longrightarrow$	Eau moyenne
$SAR_{prédit} > 26$	$\longrightarrow$ $SAR_{ordinal\ prédit} = 4$	4 $\longrightarrow$	Eau mauvaise

Pour le $SAR_{ordinal\ prédit}$, on obtient :

Tableau 10: Tableau croisé SAR ord * SARordPrdt.

Effectif		*SARordPrdt*				*Total*
		1	2	3	4	
SAR ord	1	13	11	0	0	24
	2	1	36	1	0	38
	3	0	7	0	0	7
	4	0	0	0	1	1
Total		14	54	1	1	70

A propos le $SAR_{\textbf{ordinal prédit}}$, le modèle a correctement estimé la qualité de l'eau pour :

- ((13+36+1)/70)*100 = 71,43% des puits.

Par contre, le modèle généré s'est trompé sur 3 types d'erreur :

- La première concerne le classement de 11 puits (11/70)*100 = 15,71% comme étant de bonne qualité alors qu'en réalité ils ont une excellente qualité.

- La deuxième erreur concerne le classement de 2 puits ((1+1)/70)*100 = 2,86% comme étant d'excellente et de moyenne qualité respectivement cependant ils se caractérisent par une bonne qualité.

- La troisième et la dernière erreur intéresse le classement de 7 puits (7/70)*100 = 10% comme étant de bonne qualité alors qu'ils possèdent une moyenne qualité.

On remarque que le nombre de réponses correctes pour le modèle de SAR généré est beaucoup plus élevé pour le $SAR_{\textbf{binaire prédit}}$ (88,6% de réponses correctes) que pour le $SAR_{\textbf{ordinal prédit}}$ (seulement 71,43% de réponses correctes).

Cependant, le nombre de réponses erronées dans le cas de $SAR_{\textbf{binaire prédit}}$ (11,4% de réponses fausses) est moins élevé que dans le cas de $SAR_{\textbf{ordinal prédit}}$ (28,5% de réponses falsifiées).

II.2 CE

Pour générer ce modèle dédié à l'évaluation du risque de salinité, on procède de la même manière que pour le SAR. On fait introduire la CE comme variable dépendante en fonction des autres sept variables qui lui sont corrélées (Ca^{2+}, Mg^{2+}, Na^{+} ,K^{+}, Cl^{-}, SO_4^{2-} et le pH) (variables indépendantes).

Variables introduites

Le modèle final de CE a retenu trois variables : il a introduit tout d'abord le Cl^{-}, ensuite le Na^{+} et enfin le Ca^{2+} sur la base du critère de sélection (la probabilité F est significative à $p < 0,05$).

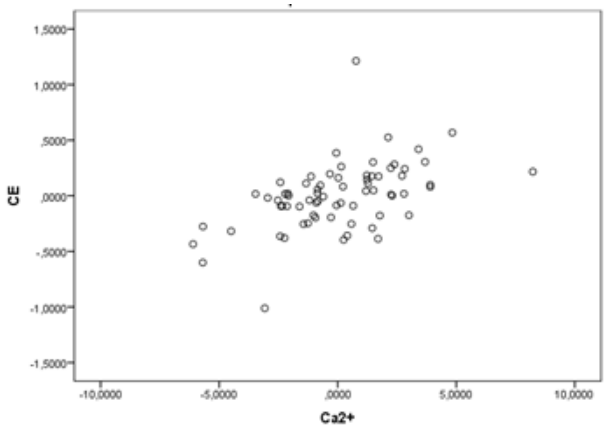


Figure 29 : Corrélation entre la CE et le Ca^{2+}.

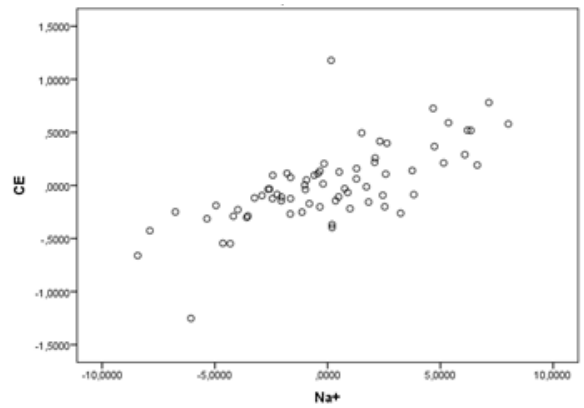


Figure 30 : Corrélation entre la CE et le Na^+.

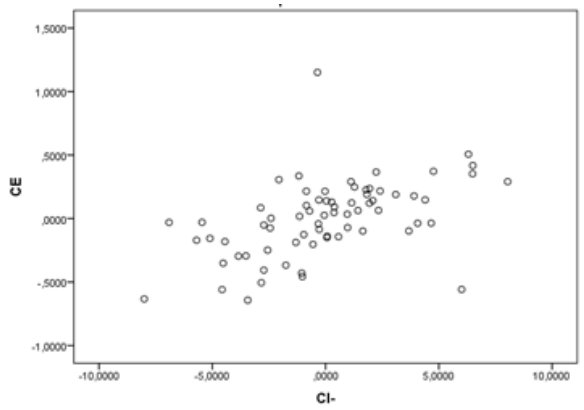


Figure 31 : Corrélation entre la CE et le Cl-.

Étape 1 : Évaluation de la qualité du modèle de régression

***Analyse de variance**

Tableau 11: ANOVA.

Modèle	*Somme des carrés*	*ddl*	*Moyenne des carrés*	*F*	*Signification*
Régression	198,98	3	66,33	950,48	0,00[d]
Résidu	4,61	66	0,07		
Total	203,58	69			

d. Valeurs prédites : (constantes), Cl^-, Na^+, Ca^{2+}.

A partir de ce tableau, il est bien clair que la valeur de F qui est hautement significative $p < 0,001$, ce qui indique que nous avons moins de 0,1 % de chance de se tromper en affirmant que le modèle contribue à mieux prédire la CE que la simple moyenne.

Étape 2 : Évaluation de l'ajustement du modèle de régression aux données

Tableau 12: Récapitulatif des modèles.

Modèle	*R*	*R^2*	*R^2 ajusté*	*Changement dans les statistiques*				
				Variation de R^2	Variation de F	ddl1	ddl2	Signification Variation de F
3	0,98[c]	0,98	0,98	0,01	22,79	1	66	0,00

c. Valeurs prédites : (constantes), Cl^-, Na^+, Ca^{2+}.

La valeur de la corrélation multiple (R) pour le modèle de CE a pour valeur 0,98, ceci indique que les données sont ajustées de manière très satisfaisante au modèle.

Le troisième modèle a pour valeur de R^2 0,98 et sa variation vaut comme valeur 0,01 et elle est significative ($p < 0,001$).

La troisième étape contribue donc significativement à l'amélioration de l'explication de la variabilité de la variable dépendante (CE).

Étape 3 : Évaluation de la variabilité expliquée par le modèle de régression

A partir des résultats du tableau 13, on peut suggérer que 98% de la variabilité de CE est expliquée par la combinaison de trois éléments Cl^-, Na^+, Ca^{2+}.

Étape 4 : Estimation des paramètres du modèle

Pour l'équation de régression concernant la variable dépendante (CE), on obtient :

Tableau 13: Coefficients du modèle.

Modèle		Coefficients non standardisés		Coefficients standardisés	t(Student)	Signification
		A	Erreur standard	Bêta		
3	(Constante)	-0,08	0,12		-,064	0,53
	Cl-	0,05	0,01	0,39	4,70	0,00
	Na+	0,07	0,01	0,48	7,98	0,00
	Ca2+	0,06	0,01	0,21	4,77	0,00

Remplaçons maintenant les β par les coefficients fournis dans le tableau ci-dessous :

$$CE_{prédite} = -0,08 + 0,05\ Cl^- + 0,07\ Na^+ + 0,06\ Ca^{2+}$$

D'après l'équation de la droite de régression, on constate que plus les éléments Cl^-, Na^+ et Ca^{2+} augmentent plus la variable CE augmente.

Nous constatons donc que les trois variables sont significatives, mais que la variabilité expliquée par l'élément Na^+ est plus importante (valeur de t élevée) que celle expliquée par les deux éléments Cl^- et Ca^{2+}.

a. Prédiction du $CE_{binaire}$ et erreur d'estimation

La $CE_{binaire\ prédite}$ est calculé à partir de la valeur de $CE_{prédite}$ (valeur prédite par la droite de régression) sous la condition suivante :

Si, $CE_{prédite} \leq 2,25 \longrightarrow CE_{binaire\ prédit} = 1$ avec : 1 ⟶ Eau non-saline

Sinon $\longrightarrow CE_{binaire\ prédit} = 0$ 0 ⟶ Eau saline

Le tableau (Tableau 14) de comparaison entre les valeurs de $CE_{binaire}$ observées et celles prédites par le modèle de régression généré donne :

Tableau 14: Tableau croisé CE bin * CEbnprédite.

Effectif		*CEbnprédite*		*Total*
		0	1	
CE bin	0	39	4	43
	1	0	27	27
Total		39	31	70

D'après le tableau croisé de CE bin*CEbnprédite (Tableau 14), on constate que :
Le modèle de régression a correctement estimé la qualité de l'eau pour ((39+27)/70)*100 = 94,28% des puits. Par opposition, le modèle a commis une erreur concernant le classement de 4 puits (4/70)*100 = 5,72% comme étant non-salins alors qu'en réalité ils se distinguent par une eau saline.

b. Prédiction du $CE_{ordinal}$ et erreur d'estimation

De la même façon que la $CE_{binaire\ prédite}$, on va calculer la $CE_{ordinale\ prédite}$.

La $CE_{ordinale\ prédit}$ est calculée à partir de la valeur de $CE_{prédite}$ (valeur prédite par la droite de régression) sous la condition suivante :

Si, $CE_{prédit} \leq 0{,}25 \longrightarrow CE_{ordinal\ prédit} = 1$ Avec : 1 ⟶ Eau non-saline

$0{,}25 < CE_{prédit} \leq 0{,}75 \longrightarrow CE_{ordinal\ prédit} = 2$ 2 ⟶ Eau moyennement saline

$0{,}75 < CE_{prédit} \leq 2{,}25 \longrightarrow CE_{ordinal\ prédit} = 3$ 3 ⟶ Eau très saline

$2{,}25 < CE_{prédit} \leq 5 \longrightarrow CE_{ordinal\ prédit} = 4$ 4 ⟶ Eau très hautement saline

$CE_{prédit} > 5 \longrightarrow CE_{ordinal\ prédit} = 5$ 5 ⟶ Eau extrêmement saline

Pour la $CE_{ordinale\ prédite}$, on obtient :

Tableau 15: Tableau croisé CEord * CEordnPrédite

Effectif		*CEordnPrédite*			*Total*
		1	2	3	
CE ord	1	38	2	0	40
	2	1	17	1	19
	3	0	0	11	11
	Total	39	19	12	70

Concernant la $CE_{ordinal\ prédit}$, le modèle a correctement estimé la qualité de l'eau pour :

- ((38+17+11)/70)*100 = 94,28% des puits.

Par contre, le modèle généré s'est trompé sur 2 types d'erreur :

- La première concerne le classement de 2 puits (2/70)*100 = 2,86% comme étant moyennement salins alors qu'en réalité ils sont non salins.

- La deuxième erreur concerne le classement de 2 puits (1+1/70)*100 = 2,86% comme étant non salins et très salins, cependant ils sont en réalité moyennement salins.

A l'opposition du modèle de SAR, pour le modèle de CE généré on constate que le nombre de réponses correctes de $CE_{binaire\ prédite}$ est équivalent à celui de $CE_{ordinale\ prédite}$ (94,28% de réponses correctes), ainsi que pour le nombre de réponses fausses il est semblable pour les deux variétés de CE (5,72% de réponses fausses).

III. Application des réseaux de neurones artificiels

Pour les réseaux de neurones artificiels, on va générer deux modèles neuronaux comme le traitement par la RLM, le premier modèle neuronal va se générer pour le SAR, alors que le deuxième pour la CE.

Pour le traitement des données par les RNA, on a choisi le perceptron multicouche comme base de traitement.

III.1 SAR

En ce qui concerne la variable SAR, on a introduit comme covariables (les variables présentant une corrélation significative), la variable Ca^{2+}, Mg^{2+}, Na^{+} et la variable CO_3^{2-}, on a obtenu comme résultat :

a. Pour le $SAR_{binaire}$

Récapitulatif de traitement des observations

Tableau16 : Récapitulatif de traitement des observations

		N	Pourcentage
Echantillon	**Apprentissage**	**49**	**70,0%**
	Test	**21**	**30,0%**
Valide		**70**	**100,0%**
Exclus		**0**	
Total		**70**	

Le récapitulatif du traitement des observations montre que 49 observations ont été attribuées à l'échantillon d'apprentissage et 21 à l'échantillon traité. Aucune observation n'a été exclue de l'analyse.

Informations sur le réseau

Le tableau d'informations (Tableau 17) sur le réseau affiche des informations sur le réseau neuronal et permet de vérifier que les spécifications sont correctes. En l'occurrence, notez les points suivants :

*Le nombre de neurones (d'unités) dans la couche d'entrée correspond au nombre de covariables utilisés; un neurone spécifique est créé pour chaque covariable présent. Pour la variable $SAR_{bianire}$, on a introduit quatre covariables et par conséquent le réseau a créé 4 neurones.

* De même, un neurone de résultat spécifique est créé pour chaque modalité de $SAR_{binaire}$, pour un total de 2 neurones dans la couche de sortie.

*La sélection automatique de l'architecture a choisi 1 seul neurone dans la couche cachée.

*Toutes les autres informations sur le réseau correspondent aux valeurs par défaut pour la procédure.

Tableau 17 : Informations réseau.

Strate d'entrée	Covariables	1 Ca2+ 2 Mg2+ 3 Na+ 4 CO3--
	Nombre d'unités[a]	4
	Méthode de rééchelonnage pour les covariables	Standardisé
Strate(s) masquée(s)	Nombre de strates masquées	1
	Nombre d'unités dans la strate masquée 1[a]	1
	Fonction d'activation	Tangente hyperbolique
Strate de sortie	Variables dépendantes	1 sar bin
	Nombre d'unités	2
	Fonction d'activation	MaxMou
	Fonction d'erreur	Entropie croisée

a. Exclusion de l'unité biaisée.

Récapitulatif des modèles

Tableau 18 : Récapitulatif des modèles.

Apprentissage	Erreur d'entropie croisée	0,11
	Prévisions de pourcentage incorrectes	0,0%
	Arrêt de la règle utilisée	1 étape(s) consécutive(s) sans diminution dans l'erreur[a]
	Durée de formation	0:00:00,02
Test	Erreur d'entropie croisée	0,00
	Prévisions de pourcentage incorrectes	0,0%

Variable dépendante : sar bin.

a. Les calculs d'erreurs sont basés sur l'échantillon de test.

Le récapitulatif du modèle (Tableau 18) affiche des informations sur les résultats de l'apprentissage du réseau final et de son application à l'échantillon traité.

* Une erreur d'entropie croisée apparaît, car la couche de sortie utilise la fonction d'activation softmax. Il s'agit de la fonction d'erreur que le réseau essaie de minimiser pendant l'apprentissage.

* Le pourcentage de prévisions incorrectes provient du tableau de classement et sera abordé plus loin dans cette section.

* L'algorithme d'estimation s'est arrêté, car l'erreur n'a pas diminué après un pas dans l'algorithme.

Classification

Tableau 19 : Classification.

Echantillon		Prévisions		
		0	1	Pourcentage correct
Apprentissage	0	4	0	100,0%
	1	0	45	100,0%
	Pourcentage global	8,2%	91,8%	100,0%
Test	0	4	0	100,0%
	1	0	17	100,0%
	Pourcentage global	19,0%	81,0%	100,0%

Le tableau de classement affiche les résultats pratiques de l'utilisation du réseau pour chaque échantillon :

*Les cellules situées sur la diagonale de la classification croisée des observations sont des prévisions correctes.

*Les cellules hors de la diagonale de la classification croisée des observations sont des prévisions incorrectes.

*4 puits (4/49)*100 = 8,2% de mauvaise qualité ont été bien estimés par le modèle neuronal et 45 puits (45/49)*100 = 91,8% d'excellente qualité ont été bien classés également. Au total, 100% des observations d'apprentissage ont été classées correctement, ce qui correspond à la proportion de 0 % de prévisions incorrectes indiquée dans le tableau récapitulatif des modèles (Tableau 18).

Les classements basés sur les observations utilisées pour créer le modèle tendent à être trop « optimistes » dans le sens où leur taux de classification est augmenté. L'échantillon traité (base de test) permet de valider le modèle ; en l'occurrence, le modèle a correctement classé 4 puits (4/21)*100 = 19,0% comme étant de mauvaise qualité ainsi que 17 puits (17/21)*100 = 81,0% comme étant d'excellente qualité. Au total, 100% des observations de test ont été bien classées. Ceci suggère qu'en général notre modèle est en fait correct environ trois fois sur quatre.

b. Pour le $SAR_{ordinal}$

Récapitulatif de traitement des observations

Pour le $SAR_{ordinal}$, on a obtenu :

Tableau 20 : Récapitulatif de traitement des observations.

		N	Pourcentage
Echantillon	**Apprentissage**	**54**	**77,1%**
	Test	**16**	**22,9%**
Valide		**70**	**100,0%**
Exclus		**0**	
Total		**70**	

Le récapitulatif du traitement des observations montre que 54 observations ont été attribuées à l'échantillon d'apprentissage et 16 à l'échantillon traité. Aucune observation n'a été exclue de l'analyse.

Informations sur le réseau

* Comme pour le $SAR_{binaire}$, le réseau a créé 4 neurones dans la couche d'entrée pour le $SAR_{ordinal}$.

* De même, un neurone de résultat spécifique est créé pour chaque modalité de $SAR_{ordinal}$, pour un total de 4 neurones dans la couche de sortie (car on se dispose pour le $SAR_{ordinal}$ de 4 modalités).

*La sélection automatique de l'architecture a choisi 2 neurones dans la couche cachée.

*Toutes les autres informations sur le réseau correspondent aux valeurs par défaut pour la procédure.

Tableau 21 : Récapitulatif de traitement des observations.

Strate d'entrée	Covariables	1	Ca2+
		2	Mg2+
		3	Na+
		4	CO3--
	Nombre d'unités[a]		4
	Méthode de rééchelonnage pour les covariables		Standardisé
Strate(s) masquée(s)	Nombre de strates masquées		1
	Nombre d'unités dans la strate masquée 1[a]		2
	Fonction d'activation		Tangente hyperbolique
Strate de sortie	Variables dépendantes	1	SAR ord
	Nombre d'unités		4
	Fonction d'activation		MaxMou
	Fonction d'erreur		Entropie croisée

Récapitulatif des modèles

Tableau 22 : Récapitulatif des modèles.

Apprentissage	Erreur d'entropie croisée	11,53
	Prévisions de pourcentage incorrectes	7,4%
	Arrêt de la règle utilisée	1 étape(s) consécutive(s) sans diminution dans l'erreur[a]
	Durée de formation	0:00:00,08
Test	Erreur d'entropie croisée	1,90
	Prévisions de pourcentage incorrectes	0,0%

Pour le tableau récapitulatif des modèles de $SAR_{ordinal}$, on a :

* Une erreur d'entropie croisée plus élevée que le $SAR_{binaire.}$

* le pourcentage de prévisions incorrectes pour la base d'apprentissage est de l'ordre de 7,4%, par opposition il est nul pour la base de test.

* L'algorithme d'estimation s'est arrêté, car l'erreur n'a pas diminué après un pas dans l'algorithme.

Classification

Tableau 23 : Classification.

Echantillon		Prévisions				
		1	2	3	4	Pourcentage correct
Apprentissage	1	15	2	0	0	88,2%
	2	0	29	1	0	96,7%
	3	0	0	6	0	100,0%
	4	0	0	1	0	0,0%
	Pourcentage global	27,8%	53,7%	11,1%	0,0%	92,6%
Test	1	7	0	0	0	100,0%
	2	0	8	0	0	100,0%
	3	0	0	1	0	100,0%
	4	0	0	0	0	0,0%
	Pourcentage global	43,8%	50,0%	6,2%	0,0%	100,0%

Les résultats de classification des modalités pour le modèle de $SAR_{ordinal}$ montrent :

Que 15 puits (15/54)*100 = 27,8% comme étant d'excellente qualité ont été bien classés, ainsi que 29 puits (29/54)*100 = 53,7% se caractérisent par une bonne qualité ont été classés correctement et 6 puits (6/54)*100 = 11,1% ont été attribué correctement comme étant de moyenne qualité. Au total, 92,6% des observations d'apprentissage ont été classées correctement, cependant le modèle s'est trompé pour 3 types d'erreur, la première concerne le classement de 2 puits (2/54)*100 = 3,7% comme étant de bonne qualité, alors qu'en réalité ils sont d'excellente qualité, la deuxième erreur concerne le classement d'un puits (1/54)*100 = 1,85% comme étant de moyenne qualité tandis qu'il est de bonne qualité, la dernière se rapporte au classement d'un puits (1/54)*100 = 1,85% comme étant de moyenne qualité cependant il est en réalité de mauvaise qualité, ceci correspond à la proportion de 7,4% de prévisions incorrectes indiquées dans le tableau récapitulatif des modèles (Tableau 22).

Pour l'échantillon traité (base de test), le modèle a correctement classé 7 puits (7/16)*100 = 43,8% comme étant d'excellente qualité, il a également bien classé 8 puits (8/16)*100 = 50% comme étant de bonne qualité et en fin il a bien attribué à un puits (1/16)*100 = 6,2% sa classe correspondante (puits de moyenne qualité). Au total, environ 100% des observations de test ont été bien classées. Ceci suggère qu'en général notre modèle est en fait correct environ trois fois sur quatre.

Si on compare alors les résultats fournis pour les deux catégories de SAR, on aperçut que le nombre de réponses correctes pour le $SAR_{bianire}$ est supérieure à celui pour le $SAR_{ordinal}$ et le nombre de réponses falsifiées pour le $SAR_{bianire}$ est nul donc nettement inférieure à celui pour le $SAR_{ordinal}$ dans la base d'apprentissage. Dans la base de test, les résultats fournis pour les deux catégories de SAR sont semblables, le modèle neuronal a bien classé tous les puits.

Donc pour le $SAR_{bianire}$ dans la base d'apprentissage, le modèle a correctement classé tous les puits (4+45/49)*100 = 100% en revanche pour le $SAR_{ordinal}$ le modèle a correctement classé 50 puits (15+29+6/54)*100 = 92,6% et il s'est trompé dans le classement de 4 puits (2+1+1/54)*100 = 7,4%. Dans la base de test concernant le $SAR_{binaire}$ le modèle a correctement classé tous les puits (4+17/21)*100 = 100% ainsi que pour le $SAR_{ordinal}$ (7+8+1/16)*100 = 100%.

III.2 CE

Pour la variable CE, on a introduit comme covariables : Le Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-} et le pH. Comme pour le SAR, on va fractionner la variable dépendante CE en deux modalités à savoir la $CE_{binaire}$ et $CE_{ordinale}$.

a.$CE_{binaire}$

Récapitulatif de traitement des observations

Pour la $CE_{binaire}$, on a obtenu :

Tableau 24 : Récapitulatif de traitement des observations.

		N	Pourcentage
Echantillon	Apprentissage	51	72,9%
	Test	19	27,1%
Valide		70	100,0%
Exclus		0	
Total		70	

Le récapitulatif du traitement des observations montre que 51 observations ont été attribuées à l'échantillon d'apprentissage et 19 à l'échantillon traité. Aucune observation n'a été exclue de l'analyse.

Informations sur le réseau

Le tableau d'informations (Tableau 25) sur le réseau montre :

*Le nombre de neurones (d'unités) dans la couche d'entrée correspond au nombre de covariables utilisés; un neurone spécifique est créé pour chaque covariable présent. Pour la variable $CE_{bianire}$, on a introduit sept covariables et par conséquent le réseau a créé 7 neurones.

* De même, un neurone de résultat spécifique est créé pour chaque modalité de $CE_{binaire}$, pour un total de 2 neurones dans la couche de sortie.

*La sélection automatique de l'architecture a choisi 5 neurones dans la couche cachée.

*Toutes les autres informations sur le réseau correspondent aux valeurs par défaut pour la procédure.

Tableau 25 : Informations réseau.

Strate d'entrée	Covariables	1	Ca2+
		2	Mg2+
		3	Na+
		4	K+
		5	Cl-
		6	SO4--
		7	pH
	Nombre d'unités[a]		7
	Méthode de rééchelonnage pour les covariables		Standardisé
Strate(s) masquée(s)	Nombre de strates masquées		1
	Nombre d'unités dans la strate masquée 1[a]		5
	Fonction d'activation		Tangente hyperbolique
Strate de sortie	Variables dépendantes	1	CE bin

Nombre d'unités	2
Fonction d'activation	MaxMou
Fonction d'erreur	Entropie croisée

Récapitulatif des modèles

Pour le tableau récapitulatif des modèles de $CE_{binaire}$ (Tableau 26), on a :

* Une erreur d'entropie croisée dans la base d'apprentissage qui vaut 4,05 mais elle a diminué jusqu'à une valeur de 0,33% dans la base de test.

* le pourcentage de prévisions incorrectes pour la base d'apprentissage est de l'ordre de 2%, par opposition il est nul pour la base de test.

* L'algorithme d'estimation s'est arrêté, car l'erreur n'a pas diminué après un pas dans l'algorithme.

Tableau 26 : Récapitulatif des modèles.

Apprentissage	Erreur d'entropie croisée	4,05
	Prévisions de pourcentage incorrectes	2,0%
	Arrêt de la règle utilisée	1 étape(s) consécutive(s) sans diminution dans l'erreur
	Durée de formation	0:00:00,08
Test	Erreur d'entropie croisée	0,33
	Prévisions de pourcentage incorrectes	0,0%

Classification

Tableau 27 : Classification.

Echantillon		Prévisions		
		0	1	Pourcentage correct
Apprentissage	0	31	1	96,9%
	1	0	19	100,0%
	Pourcentage global	60,8%	37,2%	98,0%
Test	0	11	0	100,0%
	1	0	8	100,0%
	Pourcentage global	57,9%	42,1%	100,0%

D'après le tableau de classification (Tableau 27) pour la variable $CE_{binaire}$, on se rend compte que : 31 puits (31/51)*100 = 60,8% ont été classées correctement comme étant salins, ainsi que 19 puits (19/51)*100 = 37,2% ont été bien classés comme étant non-salins. Au total, 98% des observations d'apprentissage ont été classées correctement, en revanche le modèle neuronal généré a commis un type d'erreur concerne l'attribution d'un puits (1/51)*100 = 1,96% ≈ 2% à une modalité non-saline alors qu'en réalité, il se caractérise par une eau saline, ce qui correspond à la proportion de 2 % de prévisions incorrectes indiquée dans le tableau récapitulatif des modèles (Tableau 26).

En ce qui concerne l'échantillon traité (base de test), on voit que le modèle a correctement classé 11 puits (11/19)*100 = 57,9% comme étant salins ainsi que 8 puits (8/19)*100 = 42,1% comme étant non-salins. Au total, 100% des observations de test ont été bien classées. Ceci suggère qu'en général notre modèle est en fait correct environ trois fois sur quatre.

b.$CE_{ordinale}$

Récapitulatif de traitement des observations

Pour la $CE_{ordinale}$, on a obtenu :

Tableau 28 : Récapitulatif de traitement des observations.

		N	Pourcentage
Echantillon	Apprentissage	53	75,7%
	Test	17	24,3%
Valide		70	100,0%
Exclus		0	
Total		70	

Le récapitulatif du traitement des observations montre que 53 observations ont été attribuées à l'échantillon d'apprentissage et 17 à l'échantillon traité. Aucune observation n'a été exclue de l'analyse.

Informations sur le réseau

Tableau 29 : Informations réseau.

Strate d'entrée	Covariables	1	Ca2+
		2	Mg2+
		3	Na+
		4	K+
		5	Cl-
		6	SO4--
		7	pH
	Nombre d'unités[a]		7
	Méthode de rééchelonnage pour les covariables		Standardisé

Strate(s) masquée(s)	Nombre de strates masquées	1
	Nombre d'unités dans la strate masquée 1[a]	2
	Fonction d'activation	Tangente hyperbolique
Strate de sortie	Variables dépendantes 1	CE ord
	Nombre d'unités	3
	Fonction d'activation	MaxMou
	Fonction d'erreur	Entropie croisée

Le tableau d'informations (Tableau 29) sur le réseau montre :

*Le nombre de neurones (d'unités) dans la couche d'entrée pour la variable $CE_{ordinale}$, correspond à 7 neurones (équivalent aux 7 covariables présentes).

* De même, un neurone de résultat spécifique est créé pour chaque modalité de $CE_{ordinale}$, pour un total de 3 neurones dans la couche de sortie.

*La sélection automatique de l'architecture a choisi 2 neurones dans la couche cachée.

*Toutes les autres informations sur le réseau correspondent aux valeurs par défaut pour la procédure.

Récapitulatif des modèles

Tableau 30 : Récapitulatif des modèles.

Apprentissage	Erreur d'entropie croisée	3,81
	Prévisions de pourcentage incorrectes	1,9%
	Arrêt de la règle utilisée	1 étape(s) consécutive(s) sans diminution dans l'erreur[a]
	Durée de formation	0:00:00,03
Test	Erreur d'entropie croisée	2,86
	Prévisions de pourcentage incorrectes	11,8%

Pour le tableau récapitulatif des modèles de $CE_{ordinale}$, on a :

* Une erreur d'entropie croisée dans la base d'apprentissage qui vaut 3,81, elle a diminuée jusqu'à une valeur de 2,86 dans la base de test.

* le pourcentage de prévisions incorrectes pour la base d'apprentissage est de l'ordre de 1,9%, par opposition il a augmenté jusqu'à 11,8% pour la base de test.

* L'algorithme d'estimation s'est arrêté, car l'erreur n'a pas diminué après un pas dans l'algorithme

Classification

Tableau 31 : Classification.

Echantillon		Prévisions			
		1	2	3	Pourcentage correct
Apprentissage	1	30	1	0	96,8%
	2	0	12	0	100,0%
	3	0	0	10	100,0%
	Pourcentage global	56,6%	22,6%	18,9%	98,1%
Test	1	9	0	0	100,0%
	2	1	5	1	71,4%
	3	0	0	1	100,0%
	Pourcentage global	52,9%	29,4%	5,9%	88,2%

D'après le tableau de classification (Tableau 31) pour la variable $CE_{ordinale}$, on se rend compte que : 30 puits (30/53)*100 = 56,6% ont été classées correctement comme étant non-salins, ainsi que 12 puits (12/53)*100 = 22,6% ont été bien classés comme étant moyennement salins et 10 puits (10/53)*100 = 18,9% ont été regroupés correctement comme étant très salins. Au total, 98,1% des observations d'apprentissage ont été classées correctement, en revanche le modèle neuronal généré a commis un seul type d'erreur concerne le classement d'un puits (1/53)*100 = 1,9% comme étant moyennement salin alors qu'en réalité il est non-salin, ce qui correspond à la proportion de 1,9 % de prévisions incorrectes indiquée dans le tableau récapitulatif des modèles (Tableau 30).

En ce qui concerne l'échantillon traité (base de test), on voit que le modèle a correctement classé les 3 modalités, premièrement 9 puits (9/17)*100 = 52,9% comme étant non-salins, deuxiément 5 puits (5/17)*100 = 29,4% comme étant moyennement salins et troisièment 1 puits (1/17)*100 = 5,9% comme étant très salin. Au total 88,2% des observations de test ont été bien classées. Inversement, le modèle s'est trompé pour deux types d'erreur, la première concerne l'attribution d'un puits (1/17)*100 = 5,9% à une modalité non saline alors il appartient à une modalité moyennement saline, la deuxième erreur concerne le classement d'un puits également (1/17)*100 = 5,9% comme étant très salin par contre il est moyennement salin, ce qui correspond à la proportion de 11,8 % de prévisions incorrectes indiquée dans le tableau récapitulatif des modèles (Tableau 30) dans la base de test.

En ce qui concerne les résultats fournis pour les deux catégories de CE, on constate qu'ils sont semblables dans la base d'apprentissage, pour la $CE_{binaire}$ le modèle a correctement classé 50 puits (31+19/51)*100 = 98%.

Ainsi que pour la $CE_{ordinale}$ le modèle a bien classé 52 puits (30+12+10/53)*100 = 98,1% et c'est la même chose pour le nombre de réponses fausses, pour la $CE_{binaire}$ le modèle neuronal a commis une erreur sur un puits (1/51)*100 = 1,96% de même pour la

$CE_{ordinale}$ il s'est trompé dans le classement d'un puits (1/53)*100 = 1,9%. En revanche, dans la base de test, le nombre de réponses correctes pour la $CE_{binaire}$ est supérieure à celui pour la $CE_{ordinale}$, cependant le nombre de réponses fausses pour la $CE_{binaire}$ est inférieur à celui pour la $CE_{ordinale}$. Alors le modèle a bien classé tous les puits pour la $CE_{binaire}$ (11+8/19)*100 = 100% ainsi il a bien classé 15 puits (9+5+1/17)*100 = 88,2%, en outre il s'est trompé dans deux puits (1+1/17)*100 = 11,8% pour la $CE_{ordinale}$.

IV. Comparaison entre les résultats de la RLM et des RNA

L'objectif principal de cette étude est de comparer entre les résultats fournis par les réseaux de neurones artificiels et ceux fournis par la régression linéaire multiple afin de déterminer la méthode la plus performante à l'évaluation de la qualité chimique de l'eau utilisée pour l'irrigation. Pour ce faire, on procède à une comparaison des résultats donnés pour les deux catégories de SAR ainsi que pour les catégories de CE approvisionnés par les deux méthodes.

a. Comparaison de $SAR_{binaire}$

Pour pouvoir comparer les résultats de $SAR_{binaires}$ donnés par chaque approche on a les tableaux suivants:

Tableau 32 : Tableau croisé entre le $SAR_{binaire}$ observé et celui estimé par la droite de régression.

Effectif		SARbnprédite		Total
		0	1	
sar bin	0	1	7	8
	1	1	61	62
Total		2	68	70

Tableau 33 : Tableau croisé entre le $SAR_{binaire}$ observé et celui estimé par le modèle neuronal.

Effectif		Prévision pour sarbin		Total
		0	1	
sar bin	0	8	0	8
	1	1	61	62
Total		9	61	70

Alors, si on compare les deux résultats de $SAR_{binaire}$ fournis par les deux approches, on voit bien que l'approche neuronale a bien classé les puits en termes de risque d'alcalinité (98,6% de réponses correctes) et il s'est trompé juste pour 1,4% des puits. Cependant, le modèle de régression linéaire multiple a classé seulement 88,6% des puits, en revanche il s'est trompé dans 11,4% des puits.

b. Comparaison de $SAR_{ordinal}$

En ce qui concerne les résultats de $SAR_{ordinal}$, on a :

Tableau 34 : Tableau croisé entre le $SAR_{ordinal}$ observé et celui estimé par la droite de régression.

Effectif		SARordprédit				Total
		1	2	3	4	
SAR ord	1	13	11	0	0	24
	2	1	36	1	0	38
	3	0	7	0	0	7
	4	0	0	0	1	1
Total		14	54	1	1	70

Tableau 35 : Tableau croisé entre le $SAR_{ordinal}$ observé et celui estimé par le modèle neuronal.

Effectif		Prévision pour SARord				Total
		1	2	3	4	
SAR ord	1	24	0	0	0	24
	2	1	37	0	0	38
	3	0	1	6	0	7
	4	0	0	0	1	1
Total		25	38	6	1	70

Ainsi que pour le $SAR_{ordinal}$, le pourcentage de réponses correctes fourni par le modèle neuronal vaut 97,1%, il a commis une petite erreur de 2,9% pour certains puits. Le modèle de régression a convenablement classé 71,4% des puits, par contre il s'est trompé pour un pourcentage de 28,6% des puits.

c. Comparaison de $CE_{bianire}$

Les résultats de $CE_{bianire}$ donnés par chaque méthode, sont les suivants :

Tableau 36 : Tableau croisé entre la $CE_{bianire}$ observée et celle estimée par la droite de régression.

Effectif		CEbnPrédite		Total
		0	1	
CE bin	0	39	4	43
	1	0	27	27
Total		39	31	70

Tableau 37 : Tableau croisé entre la $CE_{bianire}$ observée et celle estimée par le modèle neuronal.

Effectif		Prévision pour CEbin		Total
		0	1	
CE bin	0	40	3	43
	1	1	26	27
Total		41	29	70

Pour la $CE_{bianire}$, on voit bien que les deux approches ont fourni les mêmes résultats concernant la classification des puits en termes de risque de salinité, alors pour l'approche neuronale il est bien clair que le modèle a bien classé 94,3% des puits en revanche, il a commis des erreurs pour 5,7% des puits et c'est les même prédictions ont été élaborées pour l'approche régression.

d. Comparaison de $CE_{ordinale}$

Pour la $CE_{ordinale}$, on a obtenus comme résultats de comparaison :

Tableau 38 : Tableau croisé entre la $CE_{ordinale}$ observée et celle estimée par la droite de régression.

Effectif		CEordsnpérdite			Total
		1	2	3	
CE ord	1	38	2	0	40
	2	1	17	1	19
	3	0	0	11	11
Total		39	19	12	70

Tableau 39 : Tableau croisé entre la $CE_{ordinale}$ observée et celle estimée par le modèle neuronal.

Effectif		Prévision pour CEord			Total
		1	2	3	
CE ord	1	39	1	0	40
	2	1	17	1	19
	3	0	0	11	11
Total		40	18	12	70

Concernant la $CE_{ordinale}$, on aperçut que le modèle neuronal a bien estimé le risque de salinité pour 95,7% des puits, par contre il s'est trompé pour 4,3% des puits. En ce qui concerne le modèle de régression il a classé correctement 94,3% des puits et il s'est trompé pour 5,7% des puits.

Récapitulation

A partir de la comparaison des résultats fournis pour les quatre catégories de SAR et de CE, on constate bien que le modèle neuronal estime plus d'observations comme étant correctes que le modèle de régression linéaire ainsi que l'erreur d'estimation pour l'approche neuronale est beaucoup plus inférieur à celle fournie par l'approche de régression pour les trois cas de $SAR_{binaire}$, $SAR_{ordinal}$ et $CE_{ordinale.}$ A l'exception de $CE_{binaire}$, on constate que les deux modèles ont donnés les mêmes résultats concernant la bonne classification des puits ainsi que pour la classification falsifiée. Mais dans le cas général, les réseaux de neurones artificiels sont plus puissants dans la classification et la prédiction que la régression linéaire multiple.

V. Calcul de sensibilité et de spécificité pour les modèles de SAR et de CE

Le calcul de ces deux critères est fait seulement pour les variables qui sont binaires, il est très important car ils nous indiquent sur le type d'erreur de prédiction fournie par les modèles concernant les deux risques à savoir la salinité et l'alcalinité des eaux. En général on a deux types d'erreur, dont la première il se peut par exemple que les puits présentent une bonne qualité de l'eau en ce qui concerne les deux risques alors que les modèles générés par le RLM et les RNA nous donnent qu'il sont de mauvaise qualité, dans la deuxième type d'erreur il se peut que les puits sont de très mauvaise qualité en ce qui concerne les deux risques alors que les deux modèles fournissent de bonnes résultats c.-à-d. que les puits sont de bonne qualité et dans ce cas là que réside le danger de la deuxième type d'erreur. On peut par conséquent utiliser ces eaux pour l'irrigation des cultures bien qu'elles présentent une qualité médiocre, de ce fait il est fort probable que le développement agricole peut subir une Détérioration.

a. Calcul de la sensibilité et de la spécificité pour le modèle de $SAR_{binaire}$ généré par la RLM

La sensibilité désigne la probabilité d'obtenir un test positif sur un échantillon négatif (le faux négatif). Au contraire, la spécificité désigne la probabilité d'obtenir un test négatif sur un échantillon positif (le faux positif). Le concept de sensibilité et de spécificité est utilisé pour les tests dichotomiques (oui/non, positif/négatif, etc.)

Sensibilité : S_e = (61/ (7+61) *100) = 89.7%.

Une sensibilité de 0.897 signifie que, lorsque l'échantillon est positif, il y a 89.7% de chance que la prédiction du modèle de $SAR_{binaire}$ soit positive. Autrement dit il y a 89.7% de chance que le modèle de SARbinaire généré par l'approche régression prédit correctement une bonne qualité de l'eau en terme d'alcalinité.

Spécificité : S_p = 1/ (1+2) *100 = 33,3%.

Une spécificité de 0,33 signifie que, lorsque l'échantillon est négatif, il y a 33, 3% de chance que la prédiction du modèle de SAR binaire soit négative. Autrement dit il y a 33,3% de chance que le modèle de $SAR_{binaire}$ généré par l'approche régression prédit correctement une mauvaise qualité de l'eau en termes d'alcalinité.

b. Calcul de la sensibilité et de la spécificité pour le modèle de $SAR_{binaire}$ généré par les RNA

Pour le modèle de $SAR_{binaire}$ généré par les RNA, on a obtenu Concernant la base d'apprentissage :

Sensibilité : S_e = (45/ (0+45) *100) = 100%.

Une sensibilité de 1 signifie que, lorsque l'échantillon est positif, il y a 100% de chance que la prédiction du modèle de $SAR_{binaire}$ soit positive. Autrement dit il y a 100% de chance que le modèle de SARbinaire généré par l'approche régression prédit correctement une bonne qualité de l'eau en terme d'alcalinité.

Spécificité : S_p = 4/ (4+0) *100 = 100%

Ainsi une spécificité de 1 signifie que, lorsque l'échantillon est négatif, il y a 100% de chance que la prédiction du modèle de SAR binaire soit négative. Autrement dit il y a 100% de chance que le modèle de $SAR_{binaire}$ généré par l'approche régression prédit correctement une mauvaise qualité de l'eau en termes d'alcalinité.

Concernant la base de test, on a :

Sensibilité : S_e = (17/ (0+17) *100) = 100%.

Spécificité : S_p = 4/ (4+0) *100 = 100%.

Les mêmes résultats sont fournis pour la base de test.

c. Calcul de la sensibilité et de la spécificité pour le modèle de $CE_{binaire}$ généré par la RLM

Sensibilité : S_e = (27/ (4+27) *100) = 87,1%.

Une sensibilité de 0.871 signifie que, lorsque l'échantillon est positif, il y a 87.1% de chance que la prédiction du modèle de $SAR_{binaire}$ soit positive. Autrement dit il y a 87.1% de chance que le modèle de SARbinaire généré par l'approche régression prédit correctement une bonne qualité de l'eau en terme de salinité.

Spécificité : S_p = 39/ (39+0) *100 = 100%

Une spécificité de 1 signifie que, lorsque l'échantillon est négatif, il y a 100% de chance que la prédiction du modèle de SAR binaire soit négative. Autrement dit il y a 100% de chance que le modèle de $SAR_{binaire}$ généré par l'approche régression prédit correctement une mauvaise qualité de l'eau en termes de salinité.

d. Calcul de la sensibilité et de la spécificité pour le modèle de $CE_{binaire}$ généré par les RNA

Pour le modèle de $CE_{binaire}$ généré par les RNA, on a obtenu Concernant la base d'apprentissage :

Sensibilité : S_e = (19/ (1+19) *100) = 95%.

Une sensibilité de 0.95 signifie que, lorsque l'échantillon est positif, il y a 95% de chance que la prédiction du modèle de $SAR_{binaire}$ soit positive. Autrement dit il y a 95% de chance que le modèle de SARbinaire généré par l'approche régression prédit correctement une bonne qualité de l'eau en terme de salinité.

Spécificité : S_p = 31/ (31+0) *100 = 100%

Une spécificité de 1 signifie que, lorsque l'échantillon est négatif, il y a 100% de chance que la prédiction du modèle de SAR binaire soit négative. Autrement dit il y a 100% de chance que le modèle de $SAR_{binaire}$ généré par l'approche régression prédit correctement une mauvaise qualité de l'eau en termes de salinité.

Concernant la base de test, on a :

Sensibilité : S_e = (8/ (0+8) *100) = 100%.

Spécificité : S_p = 11/ (11+0) *100 = 100%

Les résultats de CEbianire fournis pour le modèle neuronal dans la base de test est semblables à ceux fournis pour le SARbinaire.

Récapitulation

D'après les calculs de sensibilité et de spécificité effectués pour les deux approches, il est bien clair que les deux pourcentages (S_e et S_p) dans le modèle neuronal sont beaucoup plus élevés que dans la régression linéaire multiple que ça soit pour le SAR ou pour la CE. Alors, plus les deux pourcentages augmentent plus le modèle à tendance de bien prédire les résultats positifs et négatifs. Ceci nous permet d'affirmer que le modèle neuronal est beaucoup plus performant que le modèle de régression pour l'évaluation de la qualité chimique des eaux.

Conclusion

La qualité de l'eau d'irrigation a une forte influence sur la qualité du sol et des cultures et par conséquent sur le rendement agricole. De ce fait, l'objectif de ce travail était la comparaison de la régression linéaire multiple et des réseaux de neurones artificiels pour une bonne évaluation de la qualité chimique des eaux d'irrigation dans la région de Skhirat. Cette évaluation est faite par les deux méthodes à savoir la RLM et les RNA. La surveillance des paramètres physico-chimiques de l'eau a été réalisée sur 70 puits. Des paramètres tels que le pH et la conductivité électrique ont été mesurés in situ tandis que d'autres comme l'équilibre ionique ont été mesurés en laboratoire alors que d'autres paramètres ont été calculés comme le SAR.

Le traitement des données par la RLM et les RNA montre que les eaux souterraines dans la région présentent un risque moyen de salinité avec un faible risque d'alcalinité. Donc notre échantillon d'eau se situe dans la classe C2-S1 dans le diagramme de classification américain.

Après l'analyse des données par nos modèles de RLM et de RNA, on conclut donc que le modèle neuronal est plus performant que le modèle de régression vu que son erreur d'estimation des deux risques est moins élevée, donc il a un pouvoir prédictif et explicatif plus grand que celui de la RLM. Par conséquent, on peut l'exploiter par la suite pour de nouveaux échantillons afin de mieux évaluer la qualité de l'eau désignée à l'irrigation.

En résumé, bien que l'eau souterraine dans la région de Skhirat présente un moyen risque de salinité, elle est de bonne qualité pour l'irrigation. Les pratiques agricoles devraient être bien gérées pour sécuriser l'utilisation de la ressource en eau pour un développement durable de l'agriculture dans la région.

Références bibliographiques et webographie

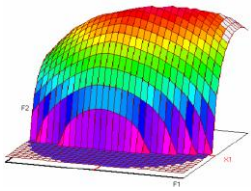
[1] www.inra.org.ma/rabat
[2] Rapport d'activité 2011 de CRRA de Rabat
[3] www.lenntech.fr - L'eau d'irrigation –Lenntech
[4] AgriMaroc: http://www.agrimaroc.ma/les-differentes-techniques-dirrigation/
[5] https://arrosage.ooreka.fr/astuce/voir/744035/irrigation-agricole
[6] https://www.memoireonline.com/01/14/8616/m_Impact-de-la-salinite-due-au-traitement-de-sel-sur-l-environnement-Cas-d-ENASEL-El-Outaya-wilaya17.html
[7] http://agro-reporter.blogspot.com/2012/01/magnetique-magnesium-partie-12.html
[8]https://www.lachimie.fr/analytique/photometrieflamme/
[9] Dakak H, Zouahri A, Iaaich H, Moussadek R, El Khadir M, Douaik A, Soudi B et Benmohammadi A. (2014). Apports des systèmes d'information géographique au diagnostic de la poulltion nitrique des eaux souterraines : cas de la zone de Skhirate, Maroc. Revue des Régions Arides, 33 : 75-79
[10] cours de régression linéaire multiple / https://bu.univ-ouargla.dz/master/pdf/BERKANI-Fatiha.pdf?idmemoire=3548
[11] cours de réseaux de neurones artificiels/ http://www.univ-usto.dz/theses_en_ligne/doc_num.php?explnum_id=1702
[12] Zouahri A, Dakak H, Douaik A, El Khadir M et Moussadek R. (2015). Evaluation of the groundwater suitability for irrigation in the Skhirate region, Northwest of Morocco. Environmental Monitoring and Assessment, 187: 4184.
[13] https://developer.ibm.com/predictiveanalytics/2017/08/08/spssstatistics-25-now-available-purchase.

Annexes

Les résultats des analyses des paramètres physico-chimiques de 70 échantillons d'eau prélevés à partir des puits dans la zone de Skhirat.

X_COORD	Y_COORD	CE	pH	Ca2+	Mg2+	Na+	K+	Cl-	SO4--	HCO3-	CO3--
-6,97	33,82	2,04	7,90	4,85	0,65	20,43	0,06	11,80	5,80	6,70	1,25
-6,98	33,82	3,48	7,05	19,60	0,40	22,67	0,05	24,00	10,20	6,53	0,00
-6,98	33,83	5,43	7,25	26,80	7,20	28,81	0,11	45,60	10,62	3,70	0,75
-6,99	33,83	3,56	6,96	14,90	3,60	21,45	0,09	24,70	8,79	5,55	0,00
-6,99	33,83	1,87	7,55	2,67	4,12	20,56	0,08	9,76	7,00	4,50	0,45
-6,98	33,84	2,59	7,66	3,70	7,30	22,26	0,17	21,40	6,59	3,75	0,70
-6,97	33,85	5,31	6,99	23,00	12,20	29,62	0,26	44,20	14,80	4,03	0,00
-7,01	33,83	2,93	7,60	8,50	5,00	23,49	0,08	24,00	5,50	4,28	1,38
-6,98	33,83	4,73	7,06	18,60	10,90	25,54	0,11	39,80	9,22	3,00	1,13
-7,00	33,84	2,20	7,14	4,84	4,00	18,83	0,10	11,44	4,50	5,33	1,00
-6,99	33,84	4,20	7,45	16,80	11,98	25,20	0,07	32,10	16,70	2,95	0,30
-6,99	33,84	2,14	7,20	3,80	0,12	18,01	0,06	10,96	3,20	5,33	0,95
-7,01	33,84	4,05	7,33	12,52	15,14	28,80	0,07	28,40	20,33	5,55	0,25
-6,99	33,85	4,03	7,07	13,90	12,60	32,89	0,20	40,20	12,42	3,90	1,08
-7,00	33,85	4,64	7,04	10,00	6,60	26,76	0,07	24,10	12,33	3,98	1,03
-7,01	33,84	0,97	7,70	2,69	3,72	15,28	0,07	5,20	3,00	2,80	0,00
-7,03	33,85	2,07	7,12	3,00	4,50	20,63	0,14	11,10	2,89	6,25	2,03
-7,03	33,85	1,12	7,49	3,42	2,84	15,64	0,08	6,32	2,80	1,88	0,75
-7,02	33,85	1,33	7,18	2,84	3,40	20,14	0,22	7,12	5,90	2,45	0,88
-7,01	33,86	2,76	7,18	3,50	0,25	27,99	0,05	18,10	5,10	5,78	1,40
-7,01	33,87	7,39	7,14	13,60	12,90	60,69	0,17	53,20	24,80	5,28	1,85
-7,01	33,87	6,18	7,34	12,60	27,80	51,61	0,26	45,20	32,02	8,25	0,80
-7,01	33,86	2,11	8,24	4,60	2,96	22,32	0,09	10,60	4,47	8,05	0,85
-7,00	33,88	3,95	7,19	3,40	6,90	34,53	0,38	30,00	5,01	6,15	2,05
-7,01	33,87	2,90	6,20	1,00	3,90	29,21	0,08	20,60	11,00	1,35	0,00
-7,01	33,87	7,89	6,86	17,70	23,60	56,61	0,27	56,80	28,50	8,03	1,90
-7,02	33,87	3,73	7,48	6,00	3,10	34,53	0,19	21,40	13,96	4,63	1,88
-7,04	33,86	1,21	7,34	4,52	3,96	16,54	0,05	4,84	1,60	6,88	1,50
-7,05	33,86	2,72	7,40	4,95	0,45	22,88	0,04	15,30	5,50	4,53	1,65
-7,05	33,87	2,24	7,25	3,84	1,40	20,79	0,22	9,40	8,50	5,40	1,63
-7,06	33,87	2,31	7,37	1,90	1,10	20,83	0,15	12,30	4,20	4,78	1,70
-7,05	33,87	3,19	7,32	6,86	8,22	28,80	0,13	16,90	20,00	5,25	0,00
-7,04	33,86	3,28	7,56	8,00	5,10	27,17	0,05	15,60	15,60	6,75	0,73
-7,03	33,88	5,71	7,35	10,30	7,90	47,61	0,71	40,20	16,82	7,50	0,00
-7,02	33,88	2,65	7,70	4,30	1,70	24,72	0,08	15,00	10,32	3,33	1,15
-7,01	33,89	1,93	7,46	4,81	1,37	18,16	0,09	8,40	10,48	4,55	0,00
-7,07	33,87	4,89	7,80	8,30	5,20	39,44	0,21	32,50	12,50	5,70	0,83
-7,07	33,86	4,94	7,33	14,20	12,30	36,16	1,09	42,00	14,23	5,08	0,45
-7,10	33,85	5,60	7,54	15,40	5,60	41,07	0,29	42,60	13,51	3,23	1,03
-7,12	33,83	2,27	7,80	2,55	4,50	22,06	0,05	13,80	3,90	3,95	0,90

-7,05	33,85	1,70	7,60	1,95	3,36	17,26	0,09	7,10	1,90	7,40	0,78
-7,05	33,85	3,70	7,80	8,00	5,00	31,67	0,11	28,00	8,25	5,45	1,08
-7,10	33,83	3,11	7,70	3,50	0,10	31,26	0,32	21,40	6,40	4,70	0,90
-7,09	33,84	1,26	7,87	4,49	2,41	15,62	0,28	5,86	1,10	4,25	1,03
-7,07	33,83	1,83	8,08	1,40	0,10	18,99	0,05	10,00	3,10	5,38	1,08
-7,08	33,83	2,89	7,67	3,00	1,90	25,33	0,06	14,00	5,60	7,53	1,60
-7,04	33,84	3,13	7,55	6,75	1,15	30,03	0,06	19,00	9,30	7,00	0,90
-7,06	33,85	2,95	7,66	5,80	1,10	26,76	0,07	20,60	4,89	5,33	1,13
-7,04	33,86	2,37	7,45	4,80	1,70	22,06	0,11	9,00	8,00	9,73	0,88
-7,04	33,85	2,14	7,73	4,00	4,96	20,47	0,04	9,76	2,40	6,38	0,28
-7,03	33,85	0,90	7,62	1,32	1,46	13,68	0,06	2,72	0,60	3,98	0,78
-7,04	33,84	1,47	7,62	0,70	1,32	15,52	0,05	6,08	0,50	5,20	0,50
-7,05	33,83	6,21	7,74	12,50	8,50	50,06	0,36	47,00	15,50	6,10	0,45
-7,08	33,81	0,92	7,80	1,78	2,02	14,91	0,03	3,10	0,49	4,65	0,50
-7,09	33,81	1,27	7,65	0,48	1,00	15,64	0,04	4,36	0,00	6,83	0,73
-7,05	33,81	2,42	7,56	3,95	0,55	20,83	0,05	13,70	3,20	6,33	0,63
-7,04	33,81	1,35	7,57	0,55	0,45	14,66	0,04	5,02	0,00	6,35	0,88
-7,03	33,82	3,86	7,58	9,20	2,60	27,58	0,04	29,00	2,07	4,58	0,78
-7,01	33,82	2,38	7,57	7,30	0,20	17,56	0,07	15,30	3,50	4,00	0,83
-7,04	33,78	1,99	7,48	4,40	2,70	18,38	0,16	11,80	1,56	5,10	0,75
-7,04	33,78	1,70	7,40	2,47	0,56	17,15	0,05	7,96	0,50	5,70	0,63
-7,03	33,80	1,30	7,50	0,50	0,55	14,33	0,01	5,30	0,00	4,55	0,88
-7,02	33,81	5,30	7,26	17,90	14,10	33,71	0,08	43,40	16,24	3,75	0,40
-7,02	33,81	8,08	7,66	3,70	7,30	76,43	0,44	63,40	15,80	5,25	0,88
-7,02	33,81	1,68	7,45	2,35	1,53	17,36	0,16	9,04	1,50	4,85	0,65
-7,02	33,81	1,54	7,56	3,61	1,69	20,02	0,23	8,50	1,90	3,78	0,58
-7,01	33,81	4,34	7,33	11,90	9,10	27,99	0,14	34,00	8,31	4,08	0,75
-7,00	33,82	1,94	7,46	3,95	0,55	16,95	0,03	12,20	2,40	5,13	0,75
-6,98	33,82	5,23	7,34	20,10	7,90	31,26	0,19	40,20	12,70	3,58	0,98
-6,98	33,81	4,16	7,31	17,84	6,28	27,20	0,05	31,30	15,38	2,70	0,00



Master ST CAC Agiq

Mémoire de fin d'études pour l'obtention du Diplôme de Master Sciences et Techniques

Nom et prénom: Fekiyer youssra

Année Universitaire : 2017/2018

Titre : Comparaison de la régression linéaire multiple et des réseaux de neurones artificiels pour l'évaluation de la qualité chimique des eaux d'irrigation dans la région de Skhirat

Résumé

Notre travail a pour but d'évaluer la qualité de l'eau d'irrigation tout en comparant entre les deux méthodes d'analyse multivariée dans la région de Skhirat qui est caractérisée par une activité maraîchère intense. Cette évaluation est faite à partir des méthodes d'analyse statistique; la RLM et les RNA.

Des échantillons d'eau sont prélevés à partir de 70 puits d'une façon représentative de la région. Ensuite, d'analyses des paramètres physico-chimiques de ces échantillons sont réalisées. Les résultats obtenus sont organisées dans un tableau appelé matrice des données.

Le traitement des données par la RLM et les RNA montre que les eaux souterraines dans la région présentent un risque moyen de salinité avec un faible risque d'alcalinité.

L'analyse des données par nos modèles montre que les modèles de prédiction de RNA sont plus prédictifs que ceux fournis de la RLM.

Mots clés: Paramètres physico-chimique d'eau, Skhirat, Régression linéaire multiple, réseaux de neurones artificiels.