# INSTITUT FRANCOPHONE INTERNATIONAL

## TRAVAIL PERSONNEL ENCADRE

RÉSEAUX DE NEURONES (ET DEEP LEARNING) POUR L'ANALYSE DE DONNÉES ÉPIDÉMIOLOGIQUE: UNE APPLICATION À L'ANALYSE DE LA DENGUE EN ASIE DU SUD EST

Présenté par : **NSANGU NGIMBI Hervé**

Encadrant Local : **Ho Tuong Vinh**

Encadrant extérieur : **Pham Nguyen Hoang, Thésard MSI**

**Master 1, Promotion 21**

**Janvier 2018**

## Table des matières

LISTE DES FIGURES ..... 4
LISTE DES TABLEAUX ..... 5
INTRODUCTION ..... 6
I. ANALYSE DU SUJET ..... 8
I.1 Présentation du Domaine du Sujet ..... 8
I.2 Les Outils, Techniques et Algorithmes Existants ..... 8
I.2.1 Outils ..... 8
I.2.2 Techniques ..... 8
I.2.3 Algorithmes ..... 9
I.3 Problèmes à Résoudre ..... 9
I.4 Termes Employés ..... 10
I.5 Principales Applications ..... 11
I.6 Résultats Attendus ..... 11
II. RECHERCHE BIBLIOGRAPHIQUE ..... 12
II.1 Concept de base en analyse de données épidémiologique par réseau de neurones (Deep Learning) ..... 12
II.1.1 L'épidémiologie ..... 12
II.1.2 La fièvre de la dengue [15] ..... 12
II.1.3 L'analyse de données [12] ..... 13
II.1.4 Réseau de neurones [3] [4] [5] ..... 14
II.1.5 Introduction à l'apprentissage en profondeur (Deep Learning) dans les réseaux de neurones (NN) ..... 16
II.2. Travaux de recherches sur l'analyse de données épidémiologique de la dengue couplée avec les facteurs climatiques ..... 19
III. SOLUTION PROPOSEE ..... 26
III.1. Contexte ..... 26
III.2. Outils à utilisés ..... 28
IV. IMPLEMENTATION ..... 29
1. Prévision de la dengue ..... 29
2. Transformer l'ensemble de données brutes en quelque chose que nous pouvons utiliser pour la prévision de séries chronologiques. ..... 31

3. Modèle de prévision LSTM multivarié ..... 34
LSTM Préparation des données ..... 34
Faire une prévision ..... 35
V.1. Première Expérimentation : Train 30% et Test 70% ..... 36
V.2. Deuxième Expérimentation : Train 60% et Test 30% ..... 39
V.3. Analyse des résultats ..... 42
CONCLUSION ET PERSPECTIVES ..... 43
REFERENCES SCIENTIFIQUES ..... 44
AUTRES REFERENCES ..... 45

# LISTE DES FIGURES

**Figure 1 :** Réseau de neurones

**Figures 2 :** Deep learning

**Figures 3 :** Aperçu des données brutes sur format csv

**Figures 4 :** Script de transformation de données

**Figures 5 :** Aperçu de données transformées

**Figures 6 :** Script pour créer une intrigue temporelle de données

**Figures 7 :** Intrigue de la série chronologique sur la dengue

**Figures 8 :** Résultat pour Epoch= 10. RMSE est de 173,450

**Figures 9 :** Résultat pour Epoch= 100. RMSE est de 152,094

**Figures 10 :** Résultat pour Epoch= 500. RMSE est de 152,006

**Figures 11 :** Résultat pour Epoch= 10. RMSE est de 148,508

**Figures 12 :** Résultat pour Epoch= 500. RMSE est de 24,621

**Figures 13 :** Résultat pour Epoch= 1000. RMSE est de 24,395

# LISTE DES TABLEAUX

1. **Tableau 1** : Tableau présentant des travaux de recherche sur l'analyse de données épidémiologique de la dengue couplée avec les facteurs climatiques
2. **Tableau 2** : Valeur des paramètres de nos expérimentations

# INTRODUCTION

L'étude des phénomènes épidémiologiques a pour but de comprendre les mécanismes de transmission et de propagation des maladies infectieuses mais aussi de prévoir les politiques et les actions les plus appropriées pour les contenir et les contrôler. C'est un champ de recherche qui connaît actuellement un important regain d'intérêt et qui fait appel à l'analyse de donnée afin de traiter, d'analyser et de comprendre un grand nombre de données pour en dégager les aspects les plus intéressants de la structure de celles-ci. L'analyse de données épidémiologique repose sur une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives. Certaines méthodes, pour la plupart géométriques, aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. D'autres techniques permettent de regrouper les données de façon à faire apparaître clairement ce qui les rend homogènes et ainsi mieux les connaître. [12]

On désigne par réseau de neurones un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement des neurones biologiques. Une attention particulière est portée au deep learning qui est une méthode d'apprentissage en profondeur du réseau de neurones. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, du traitement automatisé du langage.

Dans le souci d'évaluer et de développer notre aptitude à travailler de façon autonome et indépendante, de développer notre sens de curiosité et d'organisation, l'Institut Francophone International (IFI) a intégré, au programme de formation, un module intitulé Travail Personnel Encadré (TPE). Dans ce module chaque étudiant choisit un sujet différent sur lequel il

devra travailler sous l'encadrement des professeurs et/ou chercheurs. Ainsi, ce travail constitue notre rapport final, qui est l'ensemble de la partie théorique et pratique de notre sujet qui s'intitule : **Réseaux de neurones (et Deep Learning) pour l'analyse de données épidémiologique: une application à l'analyse de la dengue en Asie du sud Est**. Où il s'agit dans la partie théorique, de donner dans un premier temps les concepts de base en analyse de données et plus précisément en réseau de neurones puis proposer un état de l'art des méthodes d'analyse de données appliquées dans la dengue couplé avec les facteurs climatiques avant de finir par la présentation de l'approche méthodologique que nous utiliserons pour faire la prévision des séries temporelles multivariées. La seconde partie du rapport décrit l'implémentation de notre model deep learning **de type LSTM** pour faire la **Prévision des séries temporelles multivariées** sur sept (7) facteurs climatiques couplé avec l'indice de la dengue de dans tout l'étendu du Vietnam. Les expérimentations menées et l'analyse des résultats obtenus sont également présentées dans cette partie. Aux deux grandes parties précédemment décrites succèdent la conclusion et les perspectives qui achèvent la présentation de nos travaux.

# I. ANALYSE DU SUJET

Nous présentons dans cette partie de l'analyse du sujet, le domaine de notre sujet, ensuite les outils, techniques et algorithmes existant, les problèmes à résoudre, les termes et outils employés, les références et enfin les résultats attendus de notre travail.

## I.1 Présentation du Domaine du Sujet

Notre sujet se situe dans le domaine de l'Informatique de données, précisément l'analyse de données et dans le sous domaine de l'épidémiologie. Ici, nous sommes en santé publique.

Le sujet que nous allons développer est un sujet de recherche parce qu'ici, nous allons tester si des connaissances, précisément celles de réseaux de neurones (et Deep learning) peuvent être utilisées pour un domaine d'application donnée, qui dans notre cas, correspond au domaine de l'épidémiologie en santé publique.

## I.2 Les Outils, Techniques et Algorithmes Existants

### I.2.1 Outils

Etant donné que nous parlons de réseaux de neurones et précisément du deep learning, nous dirons qu'il y a plusieurs outils qui sont utilisé pour leur implémentation. Nous pouvons citer, des outils d'analyse de données commerciaux comme Matlab : « neural networks » toolbox et Netral : Neuro One et des outils d'analyse de données open-source comme JOONE : bibliothèque JAVA open source (licence LGPL), Scilab : ANN toolbox et Matlab : « netlab » toolbox, le logiciel R, le logiciel Python.

### I.2.2 Techniques

Nous allons appliquer la technique des réseaux de neurones, et en particulier celle du Deep learning qui est une méthode d'apprentissage en profondeur du réseau de neurones.

I.2.3 Algorithmes

Nous pouvons citer quelques algorithmes de réseau de neurones comme, l'algorithme de Hebb, l'algorithme de perceptron, l'algorithme de descente de gradient, l'algorithme de Widrow-Hoff et l'algorithme de rétropropagation de gradient.

Pour le deep learning, nous pouvons citer, Les algorithmes d'apprentissage purement supervisés comme, l'algorithme de Régression logistique, l'algorithme de Multilayer perceptron et l'algorithme de Réseau convolution profond.

## I.3 Problèmes à Résoudre

Dans notre travail, les problèmes à résoudre se résument en une question, qui est :

> ***« Est-il possible de prédire la dengue à l'heure actuelle (t) étant donné la mesure de la dengue et les conditions météorologiques des temps précédents (t-1,… t-n)?»***

**Les principales difficultés à prévoir sont** :

- Une telle approche nécessite de données authentiques de l'épidémie de la dengue en Asie de l'Est pour pouvoir effectuer une bonne analyse pour obtenir des résultats fiables, donc, difficulté de compréhension de données ;
- La difficulté d'apprentissages des algorithmes du Deep learning, pour la faisabilité de ce travail ;
- La difficulté d'apprentissage et de la maitrise de l'outil de développement python pour l'implémentation de l'algorithme du deep learning dans ce travail ;
- La difficulté d'applicabilité de l'algorithme du deep learning sur les données de l'épidémie de la dengue en Asie de l'Est.

## I.4 Termes Employés

1. **Réseaux de neurones** : Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit.

2. **Deep Learning** : L'apprentissage profond (en anglais *deep learning*, *deep structured learning*, *hierarchical learning*) est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaires.

3. **Analyse de données** : Est une famille de méthodes statistiques dont les principales caractéristiques est d'être multidimensionnelle et descriptives. Certaines méthodes, pour la plupart géométriques, aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données.

4. **Epidémiologie** : Est l'étude des facteurs influant sur la santé et les maladies de populations. Il s'agit d'une discipline qui se rapporte à la répartition, à la fréquence et à la gravité des états pathologiques.

5. **La dengue** : anciennement appelée « grippe tropicale », « fièvre rouge » ou « petit palu », est une infection virale, endémique dans les pays tropicaux. La dengue est une arbovirose, transmise à l'être humain par l'intermédiaire de la piqûre d'un moustique diurne du genre *Aedes*, lui-même infecté par un virus de la dengue de la famille des flaviviridae.

## I.5 Principales Applications

Nous présentons ci-dessous les différentes applications de réseaux de neurones (et Deep learning) :

- Reconnaissance faciale;
- Reconnaissance vocale;
- Vision par ordinateur ;
- Traitement automatisé du langage.

## I.6 Résultats Attendus

A la fin de ce travail personnel encadré (TPE), nous aurons à implémenter notre model deep learning de type LSTM avec la librairie Kéras de python pour faire la **Prévision des séries temporelles multivariées**, appliquer dans les données climatiques couplées avec l'épidémie de la dengue menée à l'échelle de l'ensemble du pays du Vietnam.

# II. RECHERCHE BIBLIOGRAPHIQUE

Dans cette partie, nous allons effectuer une analyse plus approfondie de notre travail. Avoir la connaissance précise et spécialisée de notre sujet, comprendre les techniques existantes, trouver les solutions possibles à notre sujet et comparer les approches (avantages et désavantages).

## II.1 Concept de base en analyse de données épidémiologique par réseau de neurones (Deep Learning)

### II.1.1 L'épidémiologie

D'après l'OMS, l'épidémiologie est : « l'étude de la distribution et des déterminants des états de santé et des maladies dans les populations humaines ainsi que des influences qui déterminent cette distribution ».

Elle a entre autres pour buts de [1] :

- Déterminer les facteurs de risque des maladies et d'évaluer leur importance ;
- D'évaluer le bien-fondé et l'efficacité des politiques de santé publique ;
- Surveiller l'état de santé des populations pour détecter des épidémies et identifier de nouvelles maladies;
- Définir des stratégies de prévention et de gestion des épidémies.

### II.1.2 La fièvre de la dengue [15]

La dengue (prononcé « dingue »), anciennement appelée « grippe tropicale », « fièvre rouge » ou « petit palu », est une infection virale, endémique dans les pays tropicaux. La dengue est une arbovirose, transmise à l'être humain par l'intermédiaire de la piqûre d'un moustique diurne du genre *Aedes*, lui-même infecté par un virus de la dengue de la famille des flaviviridae.

Cette infection virale entraîne classiquement fièvre, maux de tête, douleurs musculaires et articulaires, fatigue, nausées, vomissements et

éruptions cutanées. Biologiquement, on retrouve habituellement une baisse des plaquettes. La guérison survient généralement en une semaine.

Il existe des formes hémorragiques ou avec syndrome de choc, rares et sévères, pouvant entraîner la mort.

Il n'existe qu'un seul vaccin en cours de mise sur le marché, mais pas de traitement spécifique antiviral, d'autres vaccins sont en cours de développement.

La prise en charge repose sur un traitement symptomatique à base de médicaments contre la fièvre et la douleur. Cependant, la dengue pouvant en de rares cas évoluer vers une forme hémorragique, la prise d'antiagrégants plaquettaires comme l'aspirine est à proscrire. La prévention collective repose sur la lutte contre les moustiques vecteurs (extermination, chasse aux eaux stagnantes…) et sur les mesures de protection préventives individuelles contre les piqûres de moustiques (moustiquaire, répulsif…).

### II.1.3 L'analyse de données [12]

Est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives. Certaines méthodes, pour la plupart géométriques, aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. D'autres techniques permettent de regrouper les données de façon à faire apparaître clairement ce qui les rend homogènes et ainsi mieux les connaître.

L'analyse des données permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Le succès de cette discipline dans les dernières années est dû, dans une large mesure, aux représentations graphiques fournies. Ces graphiques peuvent mettre en évidence des relations difficilement saisies par l'analyse

directe des données ; mais surtout, ces représentations ne sont pas liées à une opinion « a priori » sur les lois des phénomènes analysés contrairement aux méthodes de la statistique classique.

## II.1.4 Réseau de neurones [3] [4] [5]

### II.1.4.1 Présentation

Les méthodes neuronales se sont développées depuis ces trente dernières années simultanément au paradigme de l'apprentissage (machine learning). Selon ce paradigme, les machines ne sont pas programmé à l'avance pour une taches donnée (par exemple, la reconnaissance d'une forme), mais ''apprennent'' à effectuer cette tâche à partir d'exemples.

### II.1.4.2 Définition

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit Ensemble de nœuds connectés entre eux, chaque variable correspondant à un nœud. C'est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement des neurones biologiques.

C'est une transposition simplifiée des neurones du cerveau humain. Les réseaux de neurones formels sont une tentative pour imiter le mécanisme d'apprentissage et qui se produit dans le cerveau.

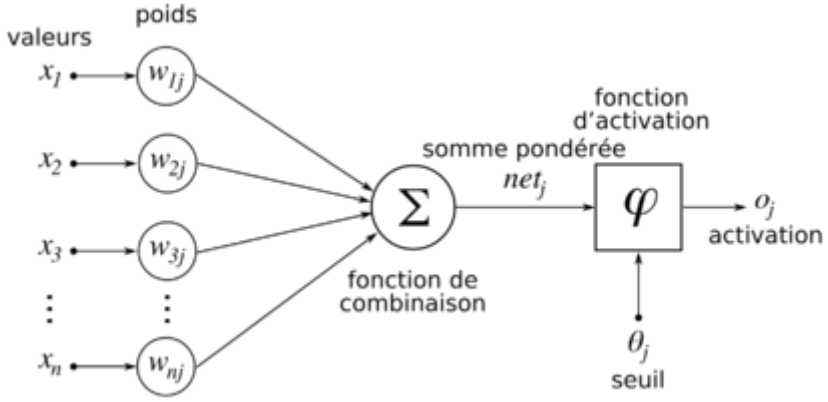



**Figure 1 : Réseau de neurones**

II.1.4.3 Types des réseaux de Neurones

II.1.4.3.1 Le perceptron

Il peut être vu comme le type de réseaux de neurone le plus simple. Il est monocouche et n'a qu'une seule sortie à laquelle toutes les entrées sont connectées.

II.1.4.3.2 Le perceptron multicouche

C'est un réseau de neurone à plusieurs couches. Lors de l'apprentissage du réseau de neurones, pour chaque exemple présenté en entrée, la valeur renvoyée (« rétropropagée ») par le nœud de sortie est comparée à la valeur réelle, et les poids pi sont ajustés.

II.1.4.3.3 Les réseaux à couches cachées

On augmente le pouvoir de prédiction en ajoutant une ou plusieurs couches cachées entre les couches d'entrée et de sortie.

II.1.4.3.4 Les réseaux à plusieurs sorties

La couche de sortie du réseau peut parfois avoir plusieurs nœuds, lorsqu'il y a plusieurs valeurs à prédire.

II.1.4.3.5 Composants de Réseaux de Neurones

Les composants sont :

- Le neurone formel ;
- Une règle d'activation ;
- Une organisation en couches ;

- Une règle d'apprentissage.

II.1.4.3.6 Apprentissage

Peut être considéré comme le problème de la mise à jour des poids des connexions au sein du réseau, afin de réussir la tâche qui lui est demandée. La règle d'apprentissage permet au réseau d'évoluer dans le temps en tenant compte des expériences antérieures. Les poids des connexions sont modifiés en fonction des résultats précédents afin de trouver le meilleur modèle par rapport aux exemples donnés.

Les réseaux de neurones se divisent en deux sortes :
Les réseaux à apprentissage supervisé et les réseaux à apprentissage non supervisé, on parle aussi des réseaux à apprentissage hybride.

Pour les réseaux à apprentissage supervisé (perceptron, Adaline, etc), on présente au réseau des entrées, et au même moment les sorties que l'on désirerait.

Pour les réseaux à apprentissage non supervisé (Hopfield, kohonen, etc), on présente une entrée au réseau et on le laisse évoluer librement jusqu'à ce qu'il se stabilise.

Les réseaux à apprentissage hybride reprennent les deux autres approches en ce sens qu'une partie des poids sera déterminée par apprentissage supervisé et l'autre partie par apprentissage non supervisé.

## II.1.5 Introduction à l'apprentissage en profondeur (Deep Learning) dans les réseaux de neurones (NN)

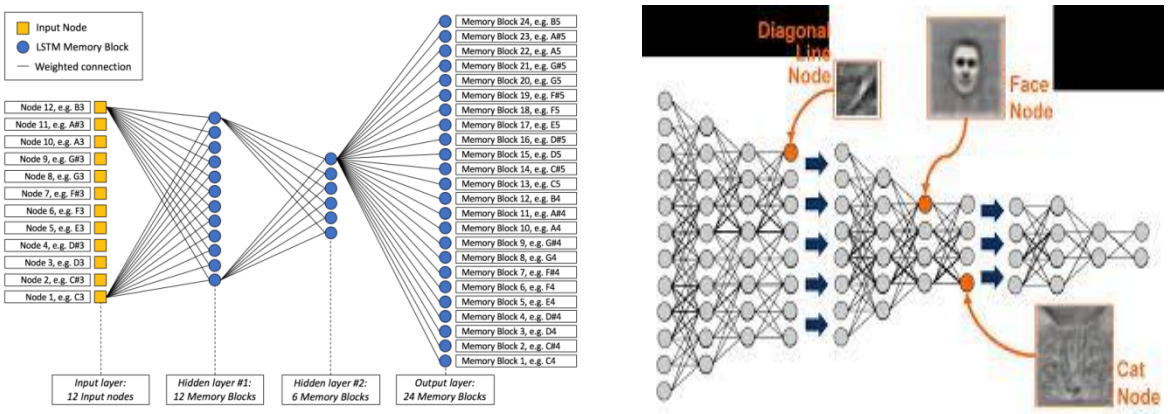
II.1.5.1 Introduction

Au cours des dernières années, les réseaux nerveux artificiels profonds (y compris les récurrents) ont gagné de nombreux concours sur la reconnaissance des formes et l'apprentissage par machine. Cette enquête historique résume de façon compacte le travail pertinent, en grande partie du millénaire précédent. Les apprenants profonds et approfondis se distinguent par la profondeur de leurs chemins d'affectation de crédit, qui sont des chaînes de liens de causalité possédant des liens de causalité entre les actions et les effets. [2]

II.1.5.2 Définition [18]

L'apprentissage en profondeur est une classe d'apprentissage automatique des algorithmes qui :

- Utiliser une cascade de plusieurs couches de traitement non linéaire d'unités pour l'extraction de caractéristiques et de transformation. Chaque couche successive utilise la sortie de la couche précédente en tant qu'entrée. Les algorithmes peuvent être supervisé ou non supervisé et les applications comprennent l'analyse de modèles (non supervisée) et la classification (supervisée) ;
- Sont basés sur la (non supervisée) l'apprentissage de plusieurs niveaux de caractéristiques ou des représentations de données. La hausse des fonctionnalités de niveau sont dérivés de fonctions de niveau inférieur afin de former une représentation hiérarchique ;
- Font partie du champ plus large d'apprentissage automatique des représentations d'apprentissage des données ;
- Apprendre plusieurs niveaux de représentation qui correspondent à différents niveaux d'abstraction; les niveaux forment une hiérarchie de concepts.




**Figures 2 : Deep learning**

Ces définitions ont en commun des couches multiples d'unités de traitement non linéaire et l'apprentissage supervisé ou non supervisé de représentations d'entités dans chaque couche, les couches formant une

hiérarchie de bas niveau à des fonctionnalités de haut niveau. La composition d'une couche d'unités de traitement non linéaires utilisées dans un algorithme d'apprentissage en profondeur dépend du problème à résoudre. Les couches qui ont été utilisés dans l'apprentissage en profondeur comprennent des couches cachées d'un réseau de neurones artificiels et des ensembles de complexes formules propositionnelles. Ils peuvent également inclure des variables latentes organisées couche par couche dans les modèles génératifs profonds tels que les nœuds dans les réseaux profonds de croyance et Machines profondes Boltzmann.

II.1.5.3 Interprétations [18]

Les réseaux de neurones profonds sont généralement interprétées en fonction de: théorème d'approximation universelle ou inférence probabiliste .

- Interprétation de théorème d'approximation universelle : Le théorème d'approximation universelle concerne la capacité des réseaux de neurones, avec une seule couche cachée de taille finie pour se rapprocher des fonctions continues. En 1989, la première preuve a été publiée par George Cybenko pour sigmoïde fonctions d'activation et a été généralisé à précompensation architectures multicouches en 1991 par Kurt Hornik.
- Interprétation probabilistes : La probabiliste interprétation dérive du domaine de l'apprentissage machine. Il dispose d'inférence, ainsi que les optimisations, concepts de formations et des tests liés au montage et généralisation respective. Plus précisément, l'interprétation probabiliste considère la non - linéarité d'activation en fonction de distributions cumulatives .Voir le réseau profond de croyance. L'interprétation probabiliste a conduit à l'introduction de décrochage en tant que régulateur dans les réseaux de neurones.

## II.2. Travaux de recherches sur l'analyse de données épidémiologique de la dengue couplée avec les facteurs climatiques

Le tableau ci-dessous présente cinq (04) articles qui traitent le problème de la prevision des variables multivariées entre les facteurs climatiques et la fièvre de la dengue.

| Articles | Avantages | Désavantages |
|---|---|---|
| 1. Pei-Chih Wu, How-Ran Guo, Shih-Chun Lung, Chuan-Yao Lin and Huey-Jen Su, Weather as an effective predictor for occurrence of dengue fever in Taiwan 2007 [7] | - Prend en compte l'âge et le sexe des personnes pour une bonne catégorisation de personnes infectées.<br><br>- Le modèle avec l'analyse des séries temporelles a bien refléter la tendance de l'incidence de la dengue dans la ville de Kaohsiung.<br><br>- cette étude a tenu compte du réchauffement précoce basé sur les prévisions météorologiques et à prendre des décisions sur le programme de prévention de la santé publique, | - Ne tient pas compte des variations des changements socio-écologiques et des modèles de transmission des maladies afin de mieux proposer le risque croissant d'une épidémie de maladie infectieuse en appliquant l'information convenablement accumulée de la variabilité météorologique.<br><br>- Ne prend pas en compte la densité de moustique et le statut d'immunité qui jouent un grand rôle aussi à la transmission de la dengue |

| | | |
|---|---|---|
| | - cette étude utilise une corrélation croisée pour évaluer les degrés de corrélation entre différentes variables météorologiques vu que c'est très important avant de se lancer à la prédiction. | - ne prend pas en compte l'humidité absolue et les heures du soleil qui peuvent aussi agir sur le comportement de moustiques de la dengue |
| 2. Liang Lu, Hualiang Lin, Linwei Tian, Weizhong Yang, Jimin Sun and Qiyong Liu, Time series analysis of dengue fever and weather in Guangzhou, China 2009 [8] | - Utilise l'approche de régression chronologique pour examiner l'effet de la variabilité météorologique sur l'incidence de la dengue dans la ville subtropicale de Guangzhou pour la période 2001-2006.<br><br>- Effectue des tests de corrélation de rang de Spearman pour examiner la relation entre l'incidence mensuelle de la dengue et les variables météorologiques avec un décalage de zéro à trois mois. | - Ne prend pas en compte les facteurs environnementaux qui influent toujours à la propagation de la dengue<br><br>- Se focalise uniquement sur la température minimale et l'humidité minimale pour conclure sur l'effet de la variabilité météorologique sur l'incidence de la dengue dans la ville subtropicale de Guangzhou pour la période 2001-2006. |

| | | |
|---|---|---|
| | - Modélise l'incidence mensuelle de la dengue selon une approche d'équations d'estimation généralisées (GEE), avec une distribution de Poisson. Ce modèle permet à la fois la spécification d'un terme de sur-dispersion et d'une structure autorégressive de premier ordre qui explique l'autocorrélation du nombre mensuel de cas de dengue<br><br>- Deux modèles de meilleure qualité, avec les plus petites valeurs QICu, sont choisis pour caractériser la relation entre l'incidence mensuelle de la dengue et les variables météorologiques. | |
| 3. Simon Hales, Neil de Wet, John Maindonald and Alistair | - Elle prend en compte des facteurs sociaux tels que | |

| | | |
|---|---|---|
| Woodward, Potential effect of population and climate changes on global distribution of dengue fever: an empirical model 2002 [9] | l'augmentation de la population ou les interactions entre les variables climatiques.<br><br>- Elle prend en compte le changement climatique mondial et une bonne répartition géographique<br><br>- Elle utilise la régression logistique, ajustée par la méthode de maximum de vraisemblance, pour modéliser la présence ou l'absence de la dengue<br><br>- Elle applique le résultat du model trouvé aux situations futures du changement climatique pour | |

| | | |
|---|---|---|
| | générer des projections sur le risque de fièvre dengue dans les années 2050 et 2080 | |
| 4. Jonathan A. Patz, Willem J.M. Martens, Dana A. Focks and and Theo H. Jettend, Dengue Fever Epidemic Potential of Global Climate Change as Projected by General Circulation Models 1998 [10] | - Elle utilise une analyse de simulation pour projeter l'altération de la transmission potentielle de la dengue liée à la température résultant de scénarios climatiques mondiaux<br><br>- Elle se concentre sur l'influence de la température sur la dynamique de la transmission virale pour une population donnée de moustiques infectés; | - Elle peut sous-estimer le changement de potentiel de transmission dans les zones tempérées.<br><br>- Elle n'évalue pas le changement dans les densités de moustiques qui pourraient être anticipées pour passer au réchauffement climatique.<br><br>- Elle ne prend pas en compte les précipitations dans l'analyse. |

| | | |
|---|---|---|
| | - Elle utilise le GCM, qui est les modèles climatiques les plus développés disponibles, pour estimer la contribution potentielle du changement climatique à la capacité vectorielle de la dengue | |

**Tableau 1** : **Tableau présentant des travaux de recherche sur l'analyse de données épidémiologique de la dengue couplée avec les facteurs climatiques**

Les travaux présentés ci-haut résument assez bien l'influence qu'a les variables climatiques face à l'épidémie de la dengue. Tous les travaux ont effectué une prevision multivariée entre les facteurs climatiques et l'épidemie de la dengue. Plusieurs méthodes d'analyse de données comme une analyse de simulation, la régression logistique, l'analyse des séries temporelles ont été utilisé par ces auteurs pour mener à bout leurs recherches. Toutes ses méthodes sont purement de méthodes statistiques, et ont permis de ressortir des fortes **prévisions des séries temporelles multivariées** climatiques étudiées comme la température maximale, moyenne et minimum, la précipitation et l'humidité relative, absolue avec l'épidémie de la dengue. Etant donné, que l'outil informatique est indispensable pour effectuer ses analyses, nous voyons l'utilisation de logiciels comme R et SPSS pour l'implémentation de leurs model d'étude.

Au vu de l'analyse précédente, nous nous proposons dans le cadre de notre travail personnel encadre, d'utiliser une nouvelle méthodologie d'analyse de données autre que le statistique ordinaire basé sur l'apprentissage automatique (Machine Learning). Et, nous allons utiliser précisément le deep learning récurrent de type LSTM pour effectuer une analyse sur **prévision des séries temporelles multivariées** entre les facteurs climatiques et la fièvre de la dengue en Asie du Sud-Est.

# III. SOLUTION PROPOSEE

## III.1. Contexte

Nous y développons en détail notre approche, les outils que nous utiliserons pour nos expérimentations et les raisons de nos choix. Nous présentons aussi les données que nous allons utiliser pour notre application.

Toutes les études réalisées au Vietnam sont réalisées à petite échelle spatiale et ne comprennent donc pas la grande variabilité spatiale des climats qui se trouve au Vietnam. Dans cette étude, nous présentons l'analyse sur la **prévision des séries temporelles multivariées** entre le climat et les syndromes de dengue menée à l'échelle de l'ensemble du pays du Vietnam. En outre, nous explorons, pour la première fois à notre connaissance, le potentiel de la méthodologie de réseau de neurones et plus précisément le Deep Learning pour faire la **prévision des séries temporelles multivariées** entre les variables climatiques et l'épidémie de la dengue. L'analyse est basée sur des séries chronologiques mensuelles d'incidence de syndromes de dengue dans 64 provinces et 7 variables climatiques de 67 stations climatiques.

Les sept facteurs climatiques qui seront utilisés dans notre travail sont :

- Température maximale ;
- Température moyenne ;
- Température minimale ;
- Humidité absolue ;
- Humidité relative ;
- Précipitation ;
- Heure soleil.

Nos données sont sous format csv. Nous appliquerons le deep learning pour faire la **prévision des séries temporelles multivariées**

entre l'incidence de la dengue et les  sept variables climatiques dans chacune des 64 provinces du Vietnam.

Donc, nous allons analyser le taux d'erreur quadratique moyenne (RMSE) pour l'ensemble des facteurs climatiques avec le cas de la dengue. C'est grâce à ce taux d'erreur quadratique moyenne (RMSE) que nous pourrons dire si oui ou non, il y a une bonne prévision de la dengue en utilisant l'ensemble des facteurs climatiques énumérés ci-haut en entrée sur tout l'ensemble du Vietnam.

Nous disons que le problème de prévision des séries temporelles multivariées par deep learning, est une régression linéaire où nous utilisons un model temporel basé sur les données des années précédentes (t-n…t-1) pour prédire la dengue à l'année actuel au temps (t).

Les problèmes de prédiction des séries chronologiques sont un problème difficile de modélisation prédictive. Contrairement à la modélisation prédictive de régression, les séries chronologiques ajoutent également la complexité d'une dépendance de séquence parmi les variables d'entrée. C'est pourquoi, Un type puissant de réseau neuronal conçu pour gérer la dépendance des séquences s'appelle les réseaux neuronaux récurrents. Nous utiliserons le réseau LSTM récurrent (The Long Short-Term Memory) utilisé dans l'apprentissage en profondeur parce que des architectures très importantes peuvent être formées avec succès.

Le réseau LSTM, En tant que tel, il peut être utilisé pour créer de grands réseaux récurrents qui, à leur tour, peuvent être utilisés pour résoudre des problèmes de séquence difficiles dans l'apprentissage par machine et obtenir des résultats à la fine pointe de la technologie. Au lieu des neurones, les réseaux LSTM possèdent des blocs de mémoire connectés via des couches. Un bloc comporte des composants qui le rendent plus intelligent qu'un neurone classique et une mémoire pour les séquences récentes. Un bloc contient des portes qui gèrent son état et sa  sortie. Un bloc fonctionne sur une séquence d'entrée et chaque grille dans un bloc utilise les unités d'activation sigmoïde

pour contrôler si elles sont déclenchées ou non, rendant le changement d'état et l'ajout d'informations à travers le bloc conditionnel.

## III.2. Outils à utilisés

Nous utiliserons les outils ci-après pour notre implémentation :

- Logiciels
  - La librairie Keras pour sa puissante architecture en apprentissage en profondeur.
  - Le langage de programmation Python

- Algorithmes
  - Algorithme de Backpropagation Through Time (BPTT)

Nous avons fait le choix du langage Python parce qu'il est un langage de programmation très puissant, flexible, pas très compliquer à implémenter et surtout qu'il comporte des librairies très puissantes comme TensorFlow, Theano, Keras pour l'apprentissage en profondeur. Nous avons choisi la librairie Keras parce qu'il est construit sur une architecture deep learning très puissante qui utilise même Tensorflow et Theano comme base. C'est un peu comme une mise en ensemble de beaucoup de librairies deep learning. En plus, il a la particularité de traiter le cas de réseau LSTM récurent dont est implémenté l'algorithme Backpropagation Through Time (BPTT) pour le problème de séries chronologiques. Nous retrouvons aussi beaucoup de documentations sur la librairie Keras.

# IV. IMPLEMENTATION

Pour cette partie, nous allons implémenter notre model deep learning récurrent de type LSTM pour faire la prévision des séries temporelles.

Les réseaux neuronaux tels que les réseaux neuronaux récurrents à longue durée de vie (LSTM) sont capables de modéliser de façon quasi transparente des problèmes avec plusieurs variables d'entrée.

C'est un grand avantage dans **la prévision des séries temporelles**, où les méthodes linéaires classiques peuvent être difficiles à adapter aux problèmes de prévision multivariés ou à entrées multiples.

Les étapes de notre implémentation se présentent comme suit :

- Transformer l'ensemble de données brutes en quelque chose que nous pouvons utiliser pour la prévision de séries chronologiques.

- Préparer les données et adapter un LSTM pour un problème de prévision de séries chronologiques multivariées.

- Faire une prévision

- Evaluer notre model

## 1. Prévision de la dengue

C'est un ensemble de données qui rend compte de la météo et du cas de la dengue pour l'ensemble du Vietnam.

Les données comprennent l'année (year), le mois (month), le jour (day), la dengue appelée cases et les renseignements météorologiques, notamment la température maximale, la température moyenne, la température minimale, l'humidité absolue, l'humidité relative, la précipitation et les heures

de soleil. La liste complète des fonctionnalités dans les données brutes est la suivante:

1. **No** : La numérotation
2. **Year** : L'année
3. **Month** : Le mois
4. **Day** : Le jour
5. **Cases** : Le cas de l'épidémie de la dengue
6. **Ta** : Température Moyenne
7. **Tx** : Température Maximale
8. **Tm** : Température Minimale
9. **Rf** : Précipitation
10. **rH** : Humidité Rélative
11. **aH** : Humidité Absolue
12. **Sh** : Heure de soleil

Nous allons utiliser ces données et faire une prévision où, compte tenu des conditions météorologiques et de la dengue des jours précédents, nous prévoyons la dengue à l'heure suivante.

| No | year | month | day | cases | Ta | Tx | Tm | Rf | rH | Sh | aH |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1977 | 1 | 1 | 9 | 25 | 28.1 | 22.7 | 1020.1 | 89 | 138.6 | 28.2 |
| 2 | 1977 | 1 | 2 | 14 | 25.8 | 30.4 | 22.9 | 6.2 | 85 | 163 | 28 |
| 3 | 1977 | 1 | 3 | 31 | 25.7 | 29.5 | 23 | 32.5 | 77 | 181.5 | 25.3 |
| 4 | 1977 | 1 | 4 | 131 | 25.7 | 26.6 | 22.7 | 1005.7 | 89 | 140.5 | 29.2 |
| 5 | 1977 | 1 | 5 | 256 | 25.6 | 30 | 23.1 | 61.2 | 82 | 173.7 | 26.7 |
| 6 | 1977 | 1 | 6 | 340 | 23.4 | 28.3 | 20.2 | 244.1 | 83 | 149.3 | 23.6 |
| 7 | 1977 | 1 | 7 | 390 | 24.9 | 31.3 | 21.4 | 66.8 | 80 | 189.5 | 24.9 |
| 8 | 1977 | 1 | 8 | 418 | 26.7 | 30.5 | 24.1 | 46.8 | 72 | 176.9 | 25.1 |
| 9 | 1977 | 1 | 9 | 514 | 23.2 | 28.5 | 14.7 | 5.5 | 80 | 178.4 | 22.4 |
| 10 | 1977 | 1 | 10 | 184 | 25.4 | 30.3 | 22.4 | 83.5 | 84 | 145.2 | 26.8 |
| 11 | 1977 | 1 | 11 | 89 | 28 | 31 | 25.5 | 734.5 | 83 | 202.1 | 31.3 |
| 12 | 1977 | 1 | 12 | 34 | 26.8 | 31 | 24.5 | 199.6 | 88 | 193.6 | 30.6 |
| 13 | 1977 | 1 | 13 | 58 | 25.7 | 29.8 | 23 | 102.8 | 82 | 165.9 | 26.9 |
| 14 | 1977 | 1 | 14 | 39 | 26.7 | 30.2 | 24.1 | 338.1 | 86 | 186 | 29.8 |
| 15 | 1977 | 1 | 15 | 54 | 25.9 | 29.4 | 23.7 | 97.1 | 81 | 151.9 | 27 |
| 16 | 1977 | 1 | 16 | 76 | 27.2 | 31 | 24.6 | 97.4 | 84 | 193.9 | 30.1 |
| 17 | 1977 | 1 | 17 | 175 | 25.8 | 30.3 | 22.9 | 30.5 | 79 | 174.3 | 26 |
| 18 | 1977 | 1 | 18 | 206 | 25.6 | 30.3 | 22.8 | 18.6 | 79 | 210.4 | 25.7 |
| 19 | 1977 | 1 | 19 | 241 | 25.7 | 31 | 22.6 | 385.5 | 85 | 186 | 27.7 |
| 20 | 1977 | 1 | 20 | 312 | 27.5 | 29.9 | 25.7 | 212.4 | 86 | 207.2 | 31.4 |
| 21 | 1977 | 1 | 21 | 402 | 21.9 | 26.6 | 18.9 | 179.8 | 89 | 156.7 | 23.1 |
| 22 | 1977 | 1 | 22 | 390 | 25.9 | 29.4 | 23.7 | 1186.1 | 90 | 135.4 | 29.9 |
| 23 | 1977 | 1 | 23 | 303 | 27.4 | 31 | 25.1 | 375.4 | 84 | 167.3 | 30.4 |
| 24 | 1977 | 1 | 24 | 199 | 27.9 | 30.4 | 25.4 | 207.7 | 84 | 181 | 31.1 |
| 25 | 1977 | 1 | 25 | 71 | 16.6 | 20.5 | 14.4 | 55.8 | 93 | 96.7 | 17.6 |

**Figures 3 : Aperçu des données brutes sur format csv**

Nous tenons à préciser que pour une bonne application de notre deep learning, nous avons eu à modifier et réarranger la plage temps. Nous avons créé une plage chronologique allant de 1977 à 2009. Toute cette modification est faite pour pouvoir avoir un très grand nombre d'enregistrement (ligne) pour permettre à notre model deep learning de pouvoir être applicable. Donc, après ce réarrangement, nous avons maintenant **12045** enregistrements.

## 2. Transformer l'ensemble de données brutes en quelque chose que nous pouvons utiliser pour la prévision de séries chronologiques.

Les données ne sont pas prêtes à être utilisées. Nous devons le préparer en premier.

- La première étape consiste à consolider les informations de date (year, month et day) en une seule « date » afin que nous puissions l'utiliser comme un index dans Pandas.
- Il y a quelques valeurs manquantes "NA" dispersées dans l'ensemble de données; nous pouvons les marquer avec 0 valeurs pour le moment.

Le script ci-dessous charge l'ensemble de données brut et analyse les informations de date sous la forme de l'index Pandas DataFrame. La colonne "Nom" est supprimée et des noms plus clairs sont spécifiés pour chaque colonne. Enfin, les valeurs NA sont remplacées par des valeurs "0".

Donc, nous avons modifié les noms de toutes les variables comme suit :

- **No** est supprimé ;
- **Year**, **Month** et **Day** sont indexé en une seule colonne appelée « date » ;
- **Cases** est appelé « dengue » ;
- **Ta** est appelé « Temp_moy » ;

- **Tx** est appelé « Temp_max » ;
- **Tm** : est appelé « Temp_Min » ;
- **Rf** est appelé « Precip » ;
- **rH** est appelé « hum_rela » ;
- **aH** est appelé « hum_abs » ;
- **Sh** est appelé « heur_sol ».

```
# -*- coding: utf-8 -*-
"""
Created on Fri Dec 15 09:57:47 2017

@author: HERVE
"""

from pandas import read_csv
from datetime import datetime

# load data
def parse(x):
      return datetime.strptime(x, '%Y %m')
dataset = read_csv('dengue_hanoi.csv',  parse_dates = [['year', 'month']], index_col=0, date_parser=parse, sep=";")
#dataset.drop('No', axis=1, inplace=True)
# manually specify column names
dataset.columns = ['dengue', 'temp_moy', 'temp_max', 'temp_min', 'hum_rela', 'hum_abs', 'precip', 'heur_sol']
dataset.index.name = 'date'
# mark all NA values with 0
#dataset['pollution'].fillna(0, inplace=True)
# drop the first 24 hours
#dataset = dataset[24:]
# summarize first 5 rows
print(dataset.head(12))
# save to file
dataset.to_csv('DENGUE.csv')
```

**Figures 4 : Script de transformation de données**

L'exécution de l'exemple imprime les 12 premières lignes de l'ensemble de données transformé et enregistre l'ensemble de données dans " *DENGUE.csv* ".

```
            dengue  temp_moy  temp_max  temp_min  hum_rela  hum_abs  precip  \
date
1977-01-01       9      25.0      28.1      22.7    1020.1     89.0   138.6
1977-01-02      14      25.8      30.4      22.9       6.2     85.0   163.0
1977-01-03      31      25.7      29.5      23.0      32.5     77.0   181.5
1977-01-04     131      25.7      26.6      22.7    1005.7     89.0   140.5
1977-01-05     256      25.6      30.0      23.1      61.2     82.0   173.7
1977-01-06     340      23.4      28.3      20.2     244.1     83.0   149.3
1977-01-07     390      24.9      31.3      21.4      66.8     80.0   189.5
1977-01-08     418      26.7      30.5      24.1      46.8     72.0   176.9
1977-01-09     514      23.2      28.5      14.7       5.5     80.0   178.4
1977-01-10     184      25.4      30.3      22.4      83.5     84.0   145.2

            heur_sol
date
1977-01-01      28.2
1977-01-02      28.0
1977-01-03      25.3
1977-01-04      29.2
1977-01-05      26.7
1977-01-06      23.6
1977-01-07      24.9
1977-01-08      25.1
1977-01-09      22.4
1977-01-10      26.8
```

**Figures 5 : Aperçu de données transformées**

- Maintenant que nous avons les données sous une forme facile à utiliser, nous pouvons créer une intrigue rapide de chaque série et voir ce que nous avons.

Le code ci-dessous charge le nouveau fichier " *DENGUE.csv* " et trace chaque série sous forme de sous-parcelle séparée.

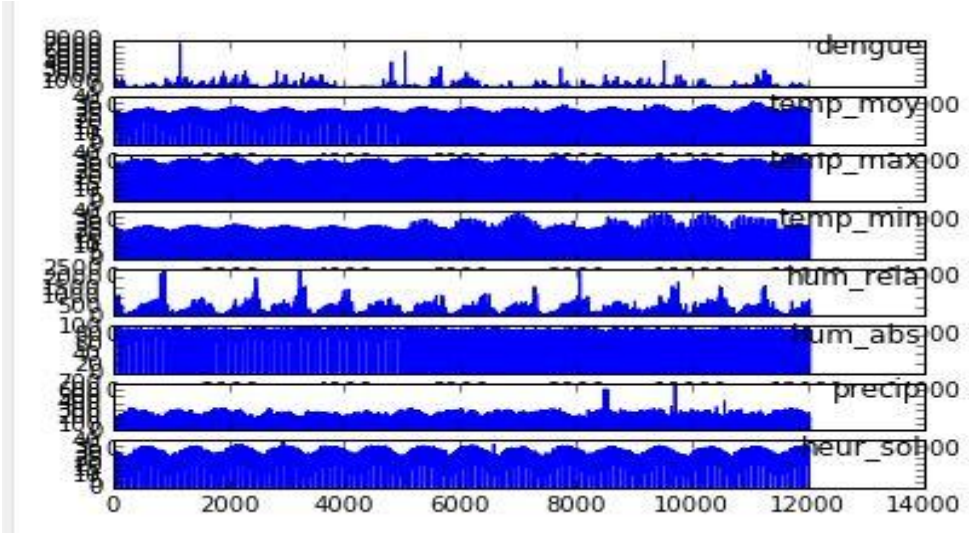
```
# -*- coding: utf-8 -*-
"""
Created on Sat Dec 16 03:56:18 2017

@author: HERVE
"""

from pandas import read_csv
from matplotlib import pyplot
# load dataset
dataset = read_csv('DENGUE.csv', header=0, index_col=0)
values = dataset.values
# specify columns to plot
groups = [0, 1, 2, 3, 4, 5, 6, 7]
i = 1
# plot each column
pyplot.figure()
for group in groups:
	pyplot.subplot(len(groups), 1, i)
	pyplot.plot(values[:, group])
	pyplot.title(dataset.columns[group], y=0.5, loc='right')
	i += 1
pyplot.show()
```

**Figures 6 : Script pour créer une intrigue temporelle de données**

L'exécution de ce code crée un graphique avec 8 sous plans montrant les 15 années de données pour chaque variable.




**Figures 7 : Intrigue de la série chronologique sur la dengue**

## 3. Modèle de prévision LSTM multivarié

Dans cette section, nous allons adapter un LSTM au problème.

### LSTM Préparation des données

La première étape consiste à préparer l'ensemble de données sur la pollution pour le LSTM.

Cela implique de cadrer l'ensemble de données comme un problème d'apprentissage supervisé et de normaliser les variables d'entrée.

Nous allons encadrer le problème d'apprentissage supervisé comme prédisant la dengue à l'heure actuelle (t) étant donné la mesure de la dengue et les conditions météorologiques à l'étape de temps précédente.

Nous pouvons transformer l'ensemble de données en utilisant la fonction *series_to_supervised ()*

- Tout d'abord, les données " *dengue_okkk.csv* " sont chargées ;
- Ensuite, toutes les entités sont normalisées, puis l'ensemble de données est transformé en un problème d'apprentissage supervisé ;
- Diviser l'ensemble de données préparé en trains et ensembles de tests et en variables d'entrée et de sortie. Enfin, les entrées (X) sont remodelées dans le format 3D attendu par les LSTM, à savoir [**samples, timesteps, features**] ;
- Nous définissons **le LSTM avec 100 neurones** dans la première couche cachée et **1 neurone** dans la couche de sortie pour prédire **la dengue ;**

- Nous utilisons **la fonction de perte d'erreur absolue moyenne** (**MAE**) et **la version efficace d'Adam d'une descente de gradient stochastique ;**
- Le modèle sera adapté pour **différents époques (epoch) (10, 100 et 500)** de formation avec une **taille de lot de 100 (batch_size).**

**Faire une prévision**

- Nous combinons la prévision avec l'ensemble de données de test et inversons la mise à l'échelle ;
- Nous calculons un **score d'erreur** pour le modèle. Dans ce cas, nous calculons **l'erreur quadratique moyenne (RMSE)** qui donne l'erreur dans les mêmes unités que la variable elle-même ;
- Nous traçons la courbe de perte du train et du test à la fin.

# V. EXPERIMENTATION

Pour nos expérimentations, nous avons allons jouer avec les taux de division de nos données en « Train » et en « Test ». Et nous allons l'effectuer sur différents « Epoch ».

Le tableau ci-dessous résume un peu les paramètres et les mesures utilisées.

| | | |
|---|---|---|
| **Première Expérimentation** | **Train : 30%**<br>**Test : 70%** | **Epoch : 10** |
| | | **Epoch : 100** |
| | | **Epoch : 500** |
| **Deuxième Expérimentation** | **Train : 60%**<br>**Test : 40%** | **Epoch : 10** |
| | | **Epoch : 500** |
| | | **Epoch : 1000** |

**Tableau 2 : Valeur des paramètres de nos expérimentations**

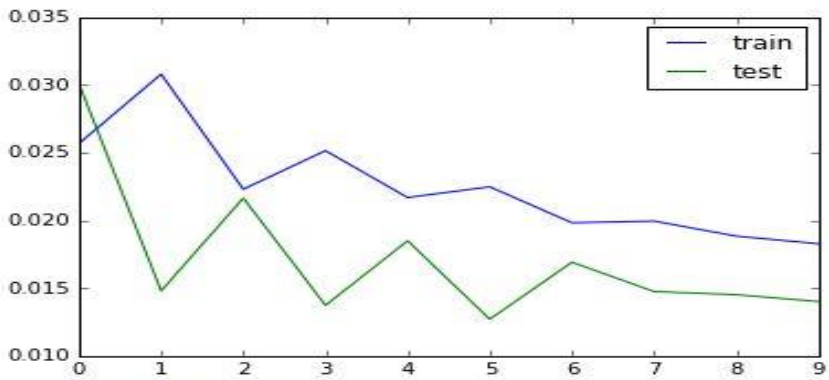
Il faut retenir qu'une Epoch signifie une passe de l'ensemble d'entraînement complet. Habituellement, il peut contenir quelques itérations. Chaque Epoch passe par tout l'ensemble d'entraînement.

## V.1. Première Expérimentation : Train 30% et Test 70%

Cette division c'est fait d'une manière intuitive. Nous savons qu'un réseau de neurones doit apprendre avec un grand nombre de données de Train pour pouvoir être plus efficace. Alors, nous commençons d'abord par lui donner peu de données de Train et beaucoup de données de Test, pour voir comment il agira.

- Pour **Epoch = 10**, nous pouvons voir nos courbes de Train et Test ainsi que le RMSE ci-dessous.

```
Epoch 9/10
1s - loss: 0.0188 - acc: 0.2170 - val_loss: 0.0145 - val_acc: 0.4508
Epoch 10/10
1s - loss: 0.0183 - acc: 0.2170 - val_loss: 0.0140 - val_acc: 0.4508
```



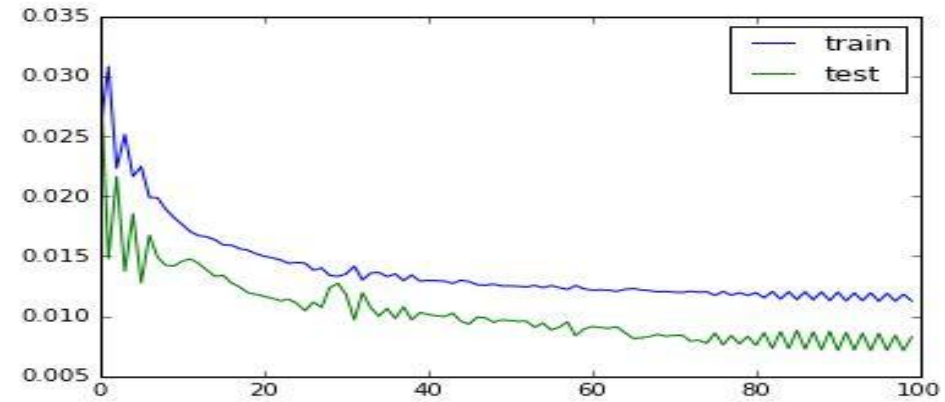

```
Test RMSE: 173.450
```

**Figures 8 : Résultat pour Epoch= 10. RMSE est de 173,450**

Nous pouvons voir que notre RMSE est de 173,450, il est supérieur à 100. Nous sommes dans un cas de surapprentissage. Nous pouvons voir aussi la tendance de la courbe de Train en bleu et de Test en vert qui ne se confondent même pas.

- Pour **Epoch = 100**, nous pouvons voir nos courbes de Train et Test ainsi que le RMSE ci-dessous

```
Epoch 99/100
1s - loss: 0.0118 - acc: 0.2173 - val_loss: 0.0071 - val_acc: 0.4508
Epoch 100/100
1s - loss: 0.0113 - acc: 0.2173 - val_loss: 0.0083 - val_acc: 0.4508
```



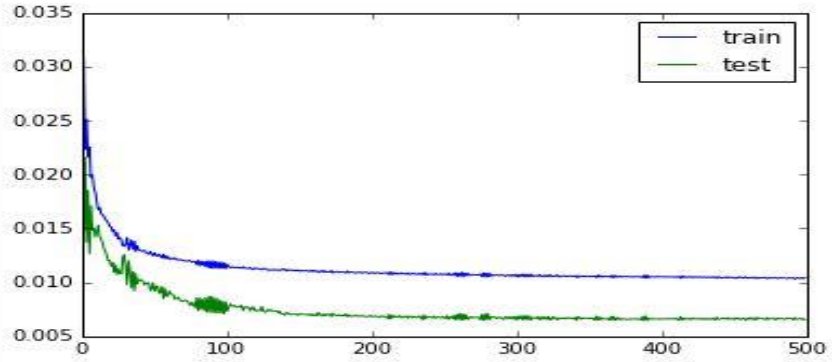

```
Test RMSE: 152.094
```

**Figures 9 : Résultat pour Epoch= 100. RMSE est de 152,094**

Nous pouvons voir que notre RMSE est de 152,094, il est supérieur à 100. Nous sommes dans un cas de surapprentissage. Nous pouvons voir aussi la tendance de la courbe de Train en bleu et de Test en vert qui ne se confondent même pas.

- Pour **Epoch = 500**, nous pouvons voir nos courbes de Train et Test ainsi que le RMSE ci-dessous

```
Epoch 499/500
1s - loss: 0.0103 - acc: 0.2170 - val_loss: 0.0066 - val_acc: 0.4508
Epoch 500/500
1s - loss: 0.0104 - acc: 0.2170 - val_loss: 0.0065 - val_acc: 0.4508
```




```
Test RMSE: 152.006
```

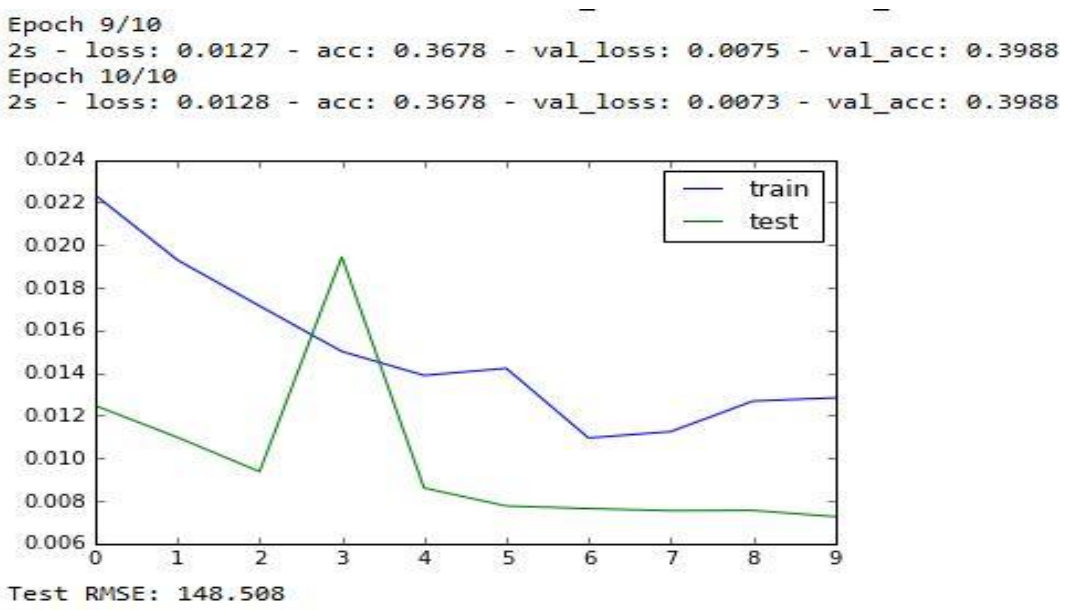
**Figures 10 : Résultat pour Epoch= 500. RMSE est de 152,006**

Nous pouvons voir que notre RMSE est de 152,006, il est supérieur à 100. Nous sommes dans un cas de surapprentissage. Nous pouvons voir aussi la tendance de la courbe de Train en bleu et de Test en vert qui ne se confondent même pas.

## V.2. Deuxième Expérimentation : Train 60% et Test 30%

Cette division c'est fait d'une manière intuitive. Nous savons qu'un réseau de neurones doit apprendre avec un grand nombre de données de Train pour pouvoir être plus efficace. Alors, nous prenons maintenant un bon taux de données Train et Test pour voir comment notre réseau de neurones réagira.

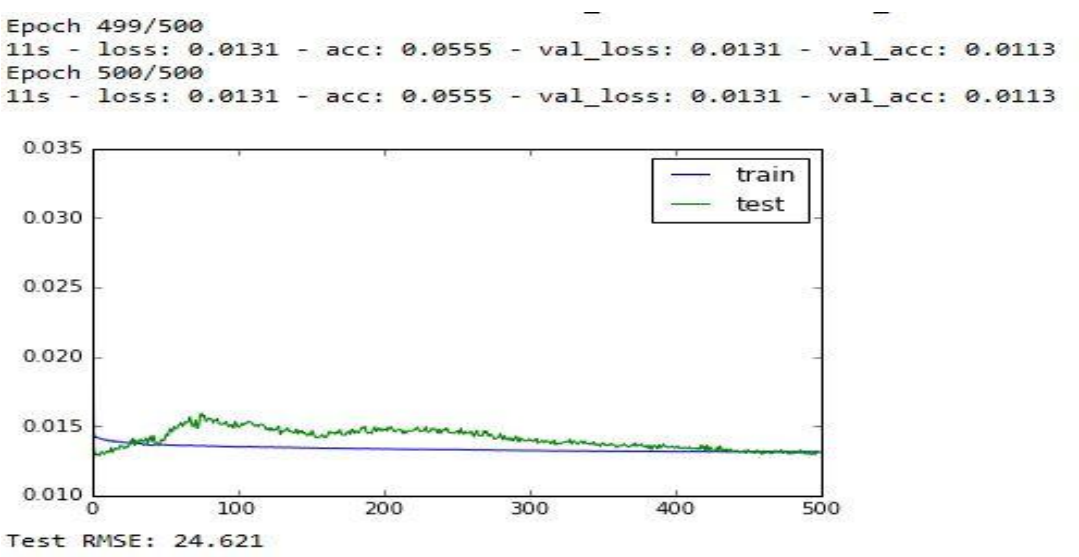
- Pour **Epoch = 10**, nous pouvons voir nos courbes de Train et Test ainsi que le RMSE ci-dessous




**Figures 11 : Résultat pour Epoch= 10. RMSE est de 148,508**

Nous pouvons voir que notre RMSE est de 148,508, il est supérieur à 100. Nous sommes dans un cas de surapprentissage. Nous pouvons voir aussi la tendance de la courbe de Train en bleu et de Test en vert qui ne se confondent même pas.

- Pour **Epoch = 500**, nous pouvons voir nos courbes de Train et Test ainsi que le RMSE ci-dessous

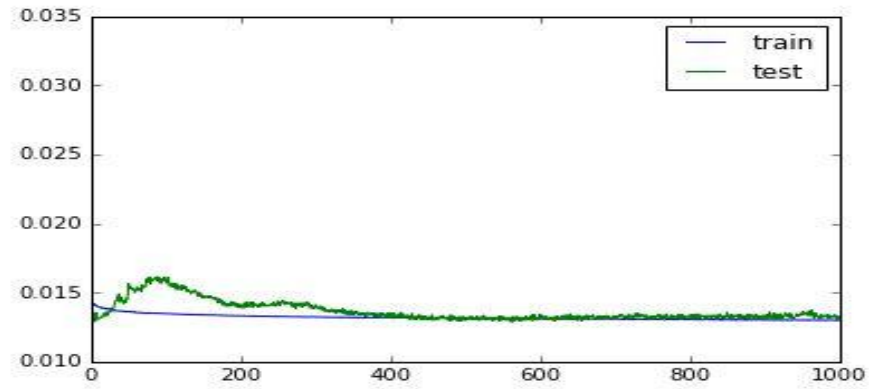



**Figures 12 : Résultat pour Epoch= 500. RMSE est de 24,621**

Nous pouvons voir que notre RMSE est de 24,621, il est compris entre 0 et 100. Nous sommes dans un cas normal, assez réaliste. Nous pouvons voir aussi la tendance de la courbe de Train en bleu et de Test en vert qui se confondent quand même.

- Pour **Epoch = 1000**, nous pouvons voir nos courbes de Train et Test ainsi que le RMSE ci-dessous

```
Epoch 999/1000
8s - loss: 0.0130 - acc: 0.0555 - val_loss: 0.0133 - val_acc: 0.0113
Epoch 1000/1000
8s - loss: 0.0130 - acc: 0.0555 - val_loss: 0.0131 - val_acc: 0.0113
```




```
Test RMSE: 24.395
```

**Figures 13 : Résultat pour Epoch= 1000. RMSE est de 24,395**

Nous pouvons voir que notre RMSE est de 24,395, il est compris entre 0 et 100. Nous sommes dans un cas normal, assez réaliste aussi. Nous pouvons voir aussi la tendance de la courbe de Train en bleu et de Test en vert qui se confondent quand même.

## V.3. Analyse des résultats

- Dans la première expérimentation avec le 30% de Train et 70% de Test, nous constatons que le modèle est en sur-apprentissage pour les trois Epoch déterminé (10, 100 et 500). Cela signifie que le model ne donne pas des résultats réaliste.
- Quand l'Erreur Quadratique Moyenne (RMSE) > 100 : Sur-apprentissage.
- Donc, il faudra un réajustement soit du model, soit retravailler les données.
- Dans la seconde expérimentation avec le 60% de Train et 40% de Test, nous constatons qu'on nous utilisons l'Epoch à 10, le model est encore en sur-apprentissage. Mais quand nous augmentons l'Epoch à 500 et à 1000, le model donne une prédiction normal avec de taux d'erreurs quadratiques appréciables (24,621% et 24,395%).
- Donc, la prévision est possible vu que nous pouvons ajuster notre modèle pour avoir des bons résultats qui ne nous conduirons pas à un sous-apprentissage ou sur-apprentissage.
- Il faut signaler que l'ajustement de l'Epoch (son augmentation), donne plus des meilleurs résultats avec un bon taux d'Erreur Quadratique Moyenne.

# CONCLUSION ET PERSPECTIVES

En guise de conclusion, ce travail personnel encadré nous aura permis de prendre connaissance des notions générales sur le réseau de neurones et plus approfondi sur le Deep Learning. Nous avons pris connaissance du sujet et des travaux connexes qui tournent autour des techniques d'analyse sur la dengue et les facteurs climatiques. Nous avons développé un model Deep Learning Récurrent de type LSTM pour la prévision des séries temporelles multivariées implémenté sous Pythons avec la librairie Kéras. L'évaluation de notre model c'est effectuée avec le calcul de l'erreur quadratique moyenne (RMSE) et du graphique de la courbe de perte de Train et la courbe de perte de Test. Nous avons effectué l'expérimentation avec différentes divisions de données de train et de test, aussi avec différentes mesures de l'Epoch. Les résultats est assez satisfiable si nous ajustant bien les paramètres notre model.

Comme perspective, il faudra effectuer une étude assez appropriée pour une bonne correspondance de données dans la résolution de problème de prévisions de séries temporelles multivariées. Et essayer d'explorer d'autres types de réseaux de neurones pour trouver le bon model qui conviendrait le mieux au problème de prévision de séries temporelles multivariées.

## REFERENCES SCIENTIFIQUES

**[1]** M. Souques : Notions de base sur l'épidémiologie -SPS n° 286, juillet-septembre 2009.

**[2]** Jürgen Schmidhuber: Deep learning in neural networks: An overview, 13 October 2014

**[3]** MAISON David, *DATAWAREHOUSE et DATAMINING,* Conservatoire Régional des Arts et Métiers Centre de Versailles, le 11 décembre 2006.

**[4]** *Data Warehouse et data mining*, Conservatoire National des Arts et Métiers de Lille, Départe-ment Informatique Version 1.1, 15 Juin 1998.

**[5]** PREUX .Ph, *Fouille de données*, Notes de Université de Lille 3, 31 août 2009.

**[6]** Pham Nguyen Hoang, Jean Daniel Zucker, Marc Choisy and and HoTuong Vinh, Causality Analysis Between Climatic Factors And Dengue Fever Using The Granger Causality 2016

**[7]** Pei-Chih Wu, How-Ran Guo, Shih-Chun Lung, Chuan-Yao Linand Huey-Jen Su, Weather as an effective predictor for occurrence of dengue fever in Taiwan 2007

**[8]** Liang Lu, Hualiang Lin, Linwei Tian, Weizhong Yang, Jimin Sun and Qiyong Liu, Time series analysis of dengue fever and weather in Guangzhou, China 2009

**[9]** Simon Hales,Neil de Wet, John Maindonald and Alistair Woodward, Potential effect of population and climate changes on global distribution of dengue fever: an empirical model 2002

**[10]** Jonathan A. Patz, Willem J.M. Martens, Dana A. Focks and and Theo H. Jettend, DengueFever Epidemic Potential of Global Climate Change as Projected by General Circulation Models 1998

**[11]** M. van Lieshout, R.S. Kovats, M.T.J. Livermore and P. Martens, Climate change and malaria: analysis ofthe SRES climate and socio-economic scenarios 2004

**[12]** S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh et al., "The global distribution and burden of dengue," Nature, vol. 496, no. 7446, pp. 504–507, 2013.

**[13]** W. G. van Panhuis, M. Choisy, X. Xiong, N. S. Chok, P. Akarasewi, S. Iamsirithaworn, S. K. Lam, C. K. Chong, F. C. Lam, B. Phommasak et al., "Region-wide synchrony and traveling waves of dengue across eight countries in southeast asia," Proceedings of the National Academy of Sciences, vol. 112, no. 42, pp. 13 069–13 074, 2015.

## AUTRES REFERENCES

[14] https://fr.wikipedia.org/wiki/Analyse_des_donn%C3%A9es

[15] https://www.tensorflow.org/install/install_windows

[16] https://fr.wikipedia.org/wiki/Apprentissage_automatique

[17] https://fr.wikipedia.org/wiki/Dengue

[18] http://deeplearning.net/tutorial/

[19] http://deeplearning.stanford.edu/tutorial/

[20] https://en.wikipedia.org/wiki/Deep_learning

[21] https://machinelearningmastery.com/time-series-forecasting-long-short term memory-network-python/