

**THÈSE DE DOCTORAT DE  
L'UNIVERSITÉ PIERRE ET MARIE CURIE**

Spécialité : **Informatique**

École Doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

**PHAM Nguyen Hoang**

Pour obtenir le grade de

**DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE**

Sujet de la thèse :

**Recherche de réseaux causaux dans des séquences temporelles  
pour l'analyse de données épidémiologique : une application à  
l'analyse de la dengue en Asie du Sud-Est.**

soutenue le ... Septembre 2018 devant le jury composé de :

Directeur de thèse : **Jean-Daniel ZUCKER** Directeur de Recherche de IRD, UMI 209, Directeur de l'unité mixte internationale de Modélisation Math et Informatique des Système complexes, Directeur de ICAN, Assistance Public Hôpitaux de Paris

Encadrants de thèse : **Ho TUONG VINH** Responsable de recherche à Institut Franco-phone International  
**Marc CHOISY** Université de Monpellier, UMR CNRS 5290/IRD 224, Oxford University Clinical Research Unit à Hanoi, Vietnam

Rapporteurs :

Examinateurs :

# Remerciements

Je tiens à remercier en premier lieu mon directeur de thèse, M. Jean-Daniel Zucker, directeur recherche de IRD, directeur de l'équipe UMMISCO, pour la confiance qu'il m'a accordée en acceptant de diriger ce travail doctoral et pour son soutien sans faille.

Je souhaite adresser mes remerciements sincères à M. HO Tuong Vinh, mon co-directeur de thèse au Vietnam, pour m'avoir proposé ce projet et pour m'avoir fait confiance tout au long de ces années.

J'adresse de chaleureux remerciements à mon co-encadrant de thèse, M. Marc Choisy, pour son attention de tout instant sur mes travaux, pour ses conseils avisés et son écoute qui ont été prépondérants pour la bonne réussite de cette thèse. Ses remarques et sa gentillesse sont autant d'élément qui m'ont permis d'atteindre les objectif de ce travail. J'ai pris un grand plaisir à travailler avec lui.

Mes remerciements vont aussi à tous les membres de l'équipe IFI (Hanoi, Vietnam) pour leur amitié, leur aide durant mes séjour au sein de l'équipe. Merci à tous mes collègues à MSI - IFI qui m'avait beaucoup aider à résoudre des problème scientifique et les problème société durant mes séjour à Hanoi.

Merci beaucoup à tous les membres de l'équipe UMMISCO (Bondy, France) et ICAN (Paris, France) de votre accueil et votre aide durant mes séjour en France.

Ces travaux n'auraient pu être réalisés sans les soutiens financiers de l'Agence Universitaire de la Francophonie. J'adresse ainsi mes remerciements aux personnes de l'AUF qui m'ont aidé de faire rapidement les procédures pour renouveler la bourse chaque année de thèse.

Je souhaite remercier vivement Mme. Kathy Baumont, Assistante gestionnaire de l'UMI UMMISCO 209, pour m'avoir toujours aidée à compléter les documents nécessaires à chaque fois que je viens en France.

Enfin, j'adresse mes profonds remerciements à ma famille, particulièrement à mes parents pour leur confiance, leur tendresse. Leur amour m'ont aidé et encouragé tous les jours. Merci pour avoir fait de moi ce que je suis aujourd'hui.

# Résumé

La dengue hémorragique (DH) est une maladie à transmission vectorielle très répandue dans les climats tropicaux et subtropicaux du monde, principalement dans les zones urbaines et semi-urbaines. Selon l'Organisation mondiale de la santé (OMS, 2007), environ 40% de la population mondiale dans 112 pays du monde est exposée au virus. On estime à 390 millions le nombre de nouvelles infections chaque année. Le problème d'identification de la relation entre cette maladie et l'influence des facteurs environnementaux et climatiques existe depuis longtemps. Cependant, les influences réelles exactes des variables climatiques sur la transmission des maladies transmises par les moustiques restent incertaines dans la plupart des cas attendus.

Dans cette thèse, nous analysons des données temporelles incluant le nombre de cas de dengue collectés mensuellement dans 273 provinces de huit pays de l'Asie du Sud-Est : Thaïlande, Cambodge, Laos, Vietnam, Malaisie, Indonésie, Philippines et Taiwan. Ces données couvrent une zone géographique de 3 500 kilomètres d'est en ouest et de 2 500 kilomètres du nord au sud. Au total, elles comptaient 320 millions d'habitants en 2010. Avant d'analyser les données, nous effectuons quelques opérations de pré-traitement. Puis, nous avons construit une vidéo qui décrit le taux d'infection de chaque ville à partir 1994 au 2010. Ensuite, nous appliquons une méthode de clustering k-means pour classifier 273 provinces en fonction de leurs taux d'infection. La distance entre les séries temporelles de leurs provinces est calculée à l'aide de la méthode Dynamic Time Wrapping (DTW). La méthode d'analyse silhouette a été utilisé pour évaluer le nombre des cluster de la méthode k-means.

Nous avons ensuite effectué quelques méthodes pour analyser la relation entre les facteurs climatiques et l'épidémie de dengue dans 64 provinces du pays Vietnam, où les données sont plus complètes. Les résultats des algorithmes montrent une fort effet de la température et de l'humidité absolue sur l'incidence de la dengue. Cependant, il existe toujours un effet indirect des précipitations, de l'humidité relative et des heures d'ensoleillement sur la dynamique de l'épidémie de dengue lorsque leur combinaison donne un meilleur résultat que leur analyse uni-variable.

Notre étude présente un aperçu de l'influence des facteurs environnementaux sur l'évolution de l'épidémie de dengue. À l'avenir, nous devons approfondir l'analyse des données des provinces d'autres pays de la région de l'Asie du Sud-Est afin d'obtenir un résultat plus complet sur ce problème.

---

**Mots clefs : dengue épidémie, modèles VAR, k-means clustering, Granger Causality Analysis, analyse ondelette, modèle EDM, Convergence Cross-Mapping, distance entre les séries temporelles**

---

## Abstract

Dengue hemorrhagic fever (DHF) is a vector-borne disease that is very common in tropical and sub-tropical climates worldwide, mostly in urban and semi-urban areas. According to the World Health Organization (WHO, 2007), about 40% of the world's population in 112 countries of the world is exposed to the virus. An estimated 390 million new infections every year. The problem of identifying the relation between this disease and the influence of environmental and climatic factors has existed for a long time. However, the exact real influences of climatic variables on the transmission of mosquito-borne diseases remain uncertain in most of the expected cases.

In this thesis, we perform some analysis on time series data that included the number of dengue case collected monthly from 273 provinces in eight Southeast Asian countries : Thailand, Cambodia, Laos, Vietnam, Malaysia, Indonesia, Philippines and Taiwan. This data covers a geographical area of 3,500 kilometers from east to west over 2,500 kilometers from north to south and had a combined population of 320 million in 2010. Before analyzing the data, we perform some preprocessing steps on the data. Then we construct a video describing the infection's rate of each city from 1994 to 2010. After that, we perform a k-means clustering method to classified 273 provinces by their infection's rate. The distance between the time series of their provinces are calculated by the dynamic time wrapping (DTW) method. The silhouette analysis method was used to evaluate the number of cluster of the k-means method.

Then, we performed some analytical method to analyse the relationship between climatic factors and the dengue epidemic in 64 provinces of Vietnam Country, where data are more complete. The results of their algorithm show the strong effect of the temperature and absolute humidity on the dengue incidence. However, there still exist an indirect effect of rainfall, relative humidity and hours of sunshine on the dynamique of dengue epidemic when their combination gives better result than their uni-variable analysis.

Our study show an overview about the influence of environmental factors ont the evolution of the dengue epidemic. In the future, we need to deepen data analysis of provinces from other countries in the Southeast Asia region to obtain a more complete an detailed result on this problem.

**Keywords:** dengue epidemic, VAR models, k-means clustering, Granger causality analysis, wavelet analysis, EDM model, cross convergence mapping, distance between time series

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Contexte . . . . .	1
1.1.1	Histoire de l'épidémiologie . . . . .	1
1.1.2	Portée, applications et difficultés de l'épidémiologie . . . . .	3
1.1.3	Les réalisations de l'épidémiologie . . . . .	5
1.1.4	Dengue . . . . .	7
1.2	L'état de l'art . . . . .	8
1.3	Notre proposition . . . . .	9
1.4	La structure de la thèse . . . . .	10
<b>2</b>	<b>Données</b>	<b>11</b>
2.1	Les séries temporelles . . . . .	11
2.1.1	Définition . . . . .	11
2.1.2	Analyse des séries temporelles . . . . .	11
2.2	Les données de la dengue . . . . .	12
2.3	Les données environnementaux . . . . .	14
2.4	Les données du Vietnam . . . . .	15
2.4.1	Le Vietnam . . . . .	15
2.4.2	Les étapes pré-traitements des données du Vietnam . . . . .	15
<b>3</b>	<b>Déetectez des zones épidémiologiques.</b>	<b>19</b>
3.1	Distance entre les séries temporelles . . . . .	19
3.1.1	Distance Euclidienne . . . . .	20
3.1.2	Complexe invariance distance - CID . . . . .	20
3.1.3	Dynamic times wrapping - DTW . . . . .	20
3.1.4	Derivative dynamic times wrapping - DDTW . . . . .	21
3.1.5	Adaptive Feature Base Dynamic Time Warping . . . . .	21
3.1.6	Fast time series evaluation - FTSE . . . . .	22
3.1.7	K-means clustering . . . . .	23
3.2	Analysis de Silhouette . . . . .	24
3.3	Clustering les données de la dengue . . . . .	24
<b>4</b>	<b>Identifier la relation entre les facteurs climatiques et l'influence de la dengue</b>	<b>28</b>
4.1	Méthodologie . . . . .	28
4.2	Résultats d'expérimentaux . . . . .	30
<b>5</b>	<b>La méthode Causalité de Granger</b>	<b>36</b>
5.1	Méthodologie . . . . .	36
5.1.1	Le modèle autorégressif - AR . . . . .	36

5.1.2	Le modèle autoregressive vectorielle - VAR . . . . .	37
5.1.3	La méthode causalité de Granger . . . . .	38
5.2	Résultats . . . . .	40
5.2.1	Résultats d'analyse univariable . . . . .	40
5.2.2	Résultats d'analyse multivariable . . . . .	41
5.2.3	Résultats d'analyse sous-séquence . . . . .	41
<b>6</b>	<b>La méthode Convergence Cross Mapping</b>	<b>47</b>
6.1	Méthodologie . . . . .	47
6.2	Résultats . . . . .	47
6.2.1	Résultats d'analyse univariable . . . . .	47
6.2.2	Résultats d'analyse multivariable . . . . .	48
<b>7</b>	<b>Conclusion</b>	<b>50</b>
7.1	Résultats de la thèse . . . . .	50
7.2	Perspectives . . . . .	51
	<b>Bibliographie</b>	<b>52</b>
<b>A</b>	<b>Les résultats complémentaires des méthodes utilisent</b>	<b>55</b>
A.1	Exploration des résultats d'analyse sous-séquence de la méthode Causalité de Granger uni-variable et multivariable . . . . .	55

# Table des figures

1.1	Cause . . . . .	3
1.2	Histoire naturelle . . . . .	4
1.3	Description de l'état de santé des populations . . . . .	4
1.4	Évaluation des interventions . . . . .	5
2.1	Le taux d'infection des provinces des pays du Sud-Est d'Asie à partir Janvier 1994 juste qu'au Décembre 2010 . . . . .	14
2.2	Les étapes de pré-traitements la donnée de la dengue au Vietnam . . . . .	16
2.3	Incidence de la dengue par mois et province au Vietnam en moyenne sur les 13 années de janvier 1998 à septembre 2010. . . . .	17
2.4	Variations des 7 variables climatiques par mois et stations climatiques moyennes sur les 13 années de janvier 1998 à septembre 2010. . . . .	18
3.1	Résultat de la méthode k-means avec $k = 3$ , méthode validation : silhouette	26
3.2	Résultat du test statistique de la méthode k-means avec $k = 3$ . . . . .	27
4.1	Exemple de la transformée de Fourier . . . . .	29
4.2	La processus d'appliquer la méthode analyse ondelette sur les données du Vietnam . . . . .	31
4.3	Cohérence des ondelettes entre le taux d'inflection de Hanoi et l'humidité relative. (a) Les signaux observé : le taux d'inflection est représenté en couleur rouge et l'humidité relative est représenté en couleur bleu. (b) Cohérence des ondelettes entre le taux d'inflection de la dengue et l'humidité relative de Hanoi, calculée à l'aide de la fonction d'ondelette de Morlet. Les couleurs codent les valeurs de puissance du bleu foncé pour une faible cohérence au rouge foncé pour une cohérence élevée. Les lignes noir imbriquées montrent les niveaux de signification $\alpha = 5\%$ calculés sur la base de 1 000 séries amorcées. Le cône d'influence indique la région non influencée par les effets de bord. . . . .	32

4.4 Cohérence des ondelettes entre le taux d'inflection de Da Nang et l'heure du soleil. (a) Les signaux observé : le taux d'inflection est représenté en couleur rouge et l'heure du soleil est représenté en couleur bleu. (b) Cohérence des ondelettes entre le taux d'inflection de la dengue et l'heure du soleil de Da Nang, calculée à l'aide de la fonction d'ondelette de Morlet. Les couleurs codent les valeurs de puissance du bleu foncé pour une faible cohérence au rouge foncé pour une cohérence élevée. Les lignes noir imbriquées montrent les niveaux de signification $\alpha = 5\%$ calculés sur la base de 1 000 séries amorcées. Le cône d'influence indique la région non influencée par les effets de bord. . . . .	33
4.5 Cohérence des ondelettes entre le taux d'inflection de Ho Chi Minh ville et la température moyenne. (a) Les signaux observé : le taux d'inflection est représenté en couleur rouge et la température moyenne est représenté en couleur bleu. (b) Cohérence des ondelettes entre le taux d'inflection de la dengue et la température moyenne de Ho Chi Minh ville, calculée à l'aide de la fonction d'ondelette de Morlet. Les couleurs codent les valeurs de puissance du bleu foncé pour une faible cohérence au rouge foncé pour une cohérence élevée. Les lignes noir imbriquées montrent les niveaux de signification $\alpha = 5\%$ calculés sur la base de 1 000 séries amorcées. Le cône d'influence indique la région non influencée par les effets de bord. . . . .	34
4.6 Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec $s = 20\%$ . . . . .	34
4.7 Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec $s = 30\%$ . . . . .	35
4.8 Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec $s = 40\%$ . . . . .	35
4.9 Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec $s = 50\%$ . . . . .	35
5.1 Résultats du test de causalité Granger pour chaque variable climatique et chaque province. . . . .	40
5.2 Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période $L = 6$ ans . . . . .	42
5.3 Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période $L = 7$ ans . . . . .	43
5.4 Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période $L = 8$ ans . . . . .	43
5.5 Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période $L = 8$ ans . . . . .	44
5.6 Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période $L = 6$ ans . . . . .	44

5.7	Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période L = 7 ans . . . . .	45
5.8	Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période L = 8 ans . . . . .	45
5.9	Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période L = 9 ans . . . . .	46
6.1	Detecting cross-map causality beyond shared seasonality of environmental drivers on dengue fever. . . . .	48
6.2	Forecast improvement with multivariate EDM. Causal effect is demonstrated if EDM forecast skill ( $\rho$ ) improves when a driver variables is included in the EDM model . . . . .	49
A.1	The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and average temperature in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	56
A.2	The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and maximal temperature in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	56
A.3	The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and minimal temperature in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	57
A.4	The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and rainfall in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	57
A.5	The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and relative humidity in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	58
A.6	The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and hours of sunshine in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	58

A.7 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and absolute humidity in this province. The Y axis is sorted by the total number of significant for every provinces over the period. . . . .	59
A.8 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and average temperature in this province. Provinces on Y axis are ordered according to their latitude. . . . .	59
A.9 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and maximal temperature in this province. Provinces on Y axis are ordered according to their latitude. . . . .	60
A.10 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and minimal temperature in this province. Provinces on Y axis are ordered according to their latitude. . . . .	60
A.11 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and rainfall in this province. Provinces on Y axis are ordered according to their latitude. . . . .	61
A.12 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and relative humidity in this province. Provinces on Y axis are ordered according to their latitude. . . . .	61
A.13 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and hours of sunshine in this province. Provinces on Y axis are ordered according to their latitude. . . . .	62
A.14 The exploration of result subsequence test with GC univariable method. The length of each period L = 6 years. The red square in each period present the relationship between dengue incidence and absolute humidity in this province. Provinces on Y axis are ordered according to their latitude. . . . .	62

# Liste des tableaux

2.1	Données climatiques des 3 pays : Vietnam, Laos et Thaïlande . . . . .	15
2.2	Changement des noms des provinces au Vietnam durant 1994 - 2010 . .	16
3.1	Silhouette coefficient des différents valeurs de clusters . . . . .	26
5.1	Le nombre des provinces qui ont la relation entre le DHF et chacun des facteurs climatiques. . . . .	40
5.2	Le nombre de provinces a été conclu qui était la relation entre DHF et l'ensemble des paires des facteur climatique . . . . .	41

# Chapitre 1

## Introduction

### Sommaire

---

<b>1.1</b>	<b>Contexte</b>	<b>1</b>
1.1.1	Histoire de l'épidémiologie	1
1.1.2	Portée, applications et difficultés de l'épidémiologie	3
1.1.3	Les réalisations de l'épidémiologie	5
1.1.4	Dengue	7
<b>1.2</b>	<b>L'état de l'art</b>	<b>8</b>
<b>1.3</b>	<b>Notre proposition</b>	<b>9</b>
<b>1.4</b>	<b>La structure de la thèse</b>	<b>10</b>

---

### 1.1 Contexte

#### 1.1.1 Histoire de l'épidémiologie

L'épidémiologie est une domaine de rechercher qui existe depuis très longtemps. On a trouvé les informations sur les maladies infectieuses comme : la variole, la peste, la lèpre,... dans les anciennes documents du Chine, Inde, Égypte, Rome et Grèce. Dans cette période, les gens a connu que la maladie peut être transmise par les patients aux personnes en bonne santé par le contact avec les patients ou leurs objets personnels. La définition sur l'immunité est aussi formé pendant cette période.

L'historien grec Fubidiv (460 - 400 B.C) a découvert que les gens qui sont déjà eu la peste n'auront reçu plus la peste. Alors, ils ont utilisé les personnes qui avaient déjà guéri avec la peste pour soigner celui qui avaient cette malade. En Chine, les gens ont mis des croûte sèche d'un bouton des petite variole dans le nez des enfants pour causer une maladie bénigne, puis créer des anticorps pour lutter contre la variole.

Dans la féodalité au Moyen-Âge, la définition sur les maladies contagieuse sont formés lorsque les gens ont réalisé que le contact avec les object personnel, l'air, l'eau ou le nourriture des malades doit contribuer à la transmission des maladies entre les patients et les personnes en bonne santé.

Jusqu'au *XIX<sup>e</sup>* siècle, les mesures à grande échelle sur la distribution des maladies dans les grandes groupes de population sont premièrement effectué en Europe. Typiquement est la découverte de John Snow sur le risque de choléra dans la ville de

Londres est lié à la consommation de différentes entreprises[1]. John Snow a trouvé une association claire entre la source d'eau et ces décès en identifiant les locations de chacun des décès de choléra à Londres entre 1848 -1849 et 1853 - 1854. Il a comparé les décès des comtés avec différentes sources d'eau et a montré que la mortalité dans les comtés où l'entreprise Southwark est l'approvisionnement en eau, était plus élevés que dans d'autre comtés. Grâce aux leurs résultats méticuleux, Snow a développé une théorie de la transmission des maladies infectieuses et suggère que le choléra se propage à travers l'eau contaminée. Ses recherches ont eu un impact direct et durable sur la politique de santé publique.

En 1861, Ignace Philippe Semmelweis, une des premiers médecins qui utilise les statistiques en médecin pour tester une hypothèse sur une étiologie d'une malade, a été publié son travail dans un livre. Il proposait à ses contemporains de se laver les main dans une solution d'hypochlorite (de l'eau de javel) et stérilisait ses instruments de chirurgie. Malheureusement, l'opposition parmi ses contemporains ne permit pas de faire avancer ses idées. À la fin du XIX<sup>e</sup> siècle et au début du XX<sup>e</sup> siècle, des méthodes mathématiques furent introduit en épidémiologie par Ronald Ross, W.O Kermarck et A.G McKendrick [?], [3], [4]. Cette approche a été initialement appliquée à la lutte contre les maladies infectieuses et s'est avérée efficace pour mettre en évidence une association entre certaines conditions ou agents environnementaux à des maladies déterminée. Dans la deuxième moitié du XX<sup>e</sup> siècle, en particulier dans les pays à revenue élevé ou moyen, cette méthode est applicable aux maladies non transmissible chronique telles que les maladies cardiaques et le cancer.

Avec l'apparition des maladies nosocomiales, d'une antibiorésistance préoccupant, des circulations rapidement des microbes sur notre planète, La veille écoépidémiologique et l'accès rapide à des information transparentes et valides devient majeur problème. Un des publication très célèbre au début du deuxième moitié du XX<sup>e</sup> siècle est la recherche de Richard Doll et Andrew Hill sur la relation entre tabagisme et le cancer du poumon[5]. Ils ont effectué des observation sur 41,000 homme et femme au Royaume-Uni avec des condition médicale qualifié pendant 12 ans. Ils ont montré que la mortalité par le cancer du poumon a fortement augmenté chez les fumeurs pendant des années 1952 - 1961. Le taux de mortalité par cancer du poumon est plus élevé chez les légers et moyen fumeurs qui inhalent que chez ceux qui n'inhalent pas.

Le tabagisme est un cas particulier dans l'épidémiologie , mais pour la plupart des maladies, il existe plusieurs facteurs qui contribuent à la cause. Le travail de recherche pour identifier la relations entre les maladies et les facteurs causal devient une grande problème dans la domaine l'épidémiologie . De nouvelle méthodes épidémiologiques sont utilisées pour analyser ces relations. L'épidémiologie joue un rôle très important dans la détection et le contrôle des maladies dans des pays à revenue faible ou intermédiaire comme le VIH / SIDA, la tuberculose et le paludisme. Cette branche d'épidémie devient de plus en plus important dans tous les pays du fait de l'émergence de nouvelles maladies transmissibles comme le syndrome respiratoire aigu sévère (SRAS), l'encéphalopathie spongiforme bovine (ESB) et la grippe pandémique. L'épidémiologie a considérablement évolué ces 50 dernières années et son principal défi à l'heure actuelle consiste à explorer et à agir sur les déterminants sociaux de la santé et de la maladie, dont la plupart se trouvent en dehors du secteur sanitaire.

### 1.1.2 Portée, applications et difficultés de l'épidémiologie

#### Portée

L'épidémiologie est une science théorique fondamentale de la médecine et des autres sciences de la santé. Elle est largement utilisée dans la recherche et les services de santé publique quotidiens. L'épidémiologie s'intéresse sur la relation entre la santé humaine et les facteurs environnementaux (géographiques, biologiques, sociaux, etc...). Cette relation peut avoir un effet positif ou négatif sur la santé de la communauté. Le but de l'épidémiologie est de l'analyse et de la comprendre afin de fournir des meilleures interventions au bénéfice de la communauté.

#### Applications

- Cause de la maladie : l'épidémiologie découvre les causes des maladies. La majorité des maladies sont causées par les interactions entre les facteurs génétiques et environnementaux (Figure 1.1).

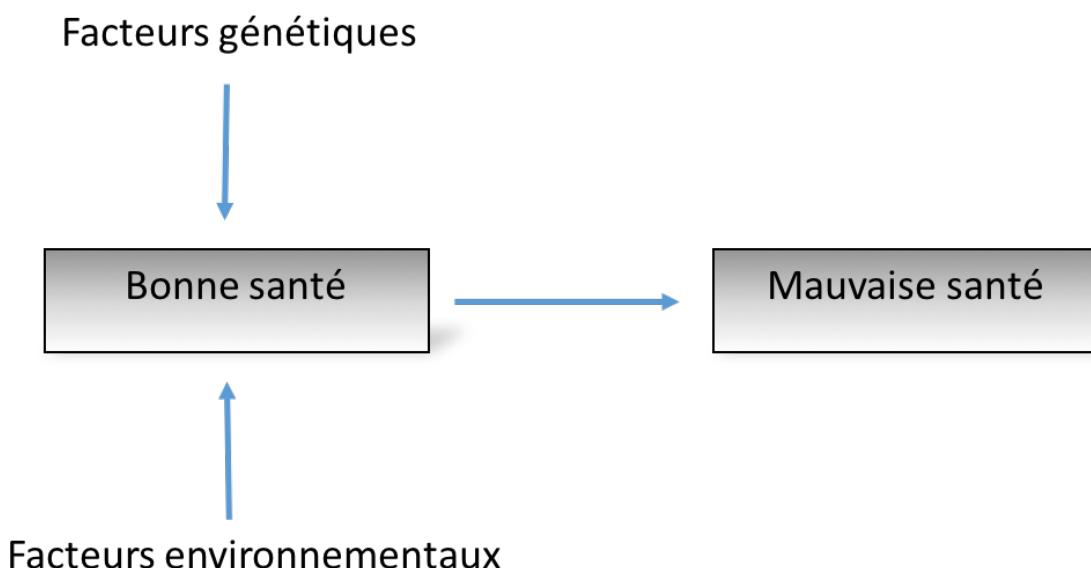


FIGURE 1.1 – Cause

- Histoire naturelle de la maladie : L'épidémiologie aussi s'intéresse sur l'évolution et l'histoire naturelle des maladies chez les individus et les groupes (Figure 1.2)
- Situation sanitaire des populations : L'épidémiologie est souvent utilisée pour décrire état de santé des groupes de population (Figure 1.3).
- Évaluation des interventions : l'épidémiologie est utilisée pour évaluer l'efficacité des services de santé (Figure 1.4). L'application des principes et méthodes épidémiologiques aux problèmes rencontrés dans le domaine médical a conduit à la développement des autres domaines comme l'épidémiologie moléculaire, l'épidémiologie génétique, la pharmaco-épidémiologie ,....

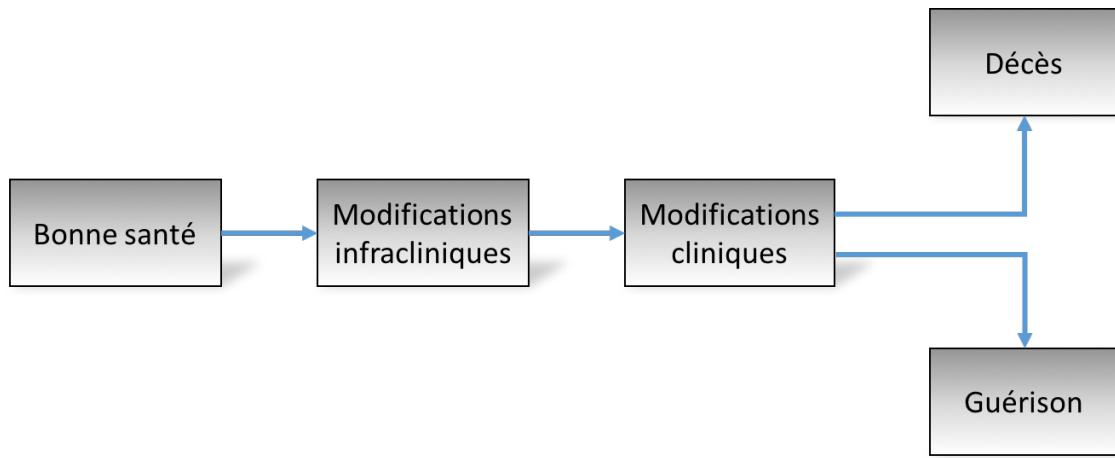


FIGURE 1.2 – Histoire naturelle

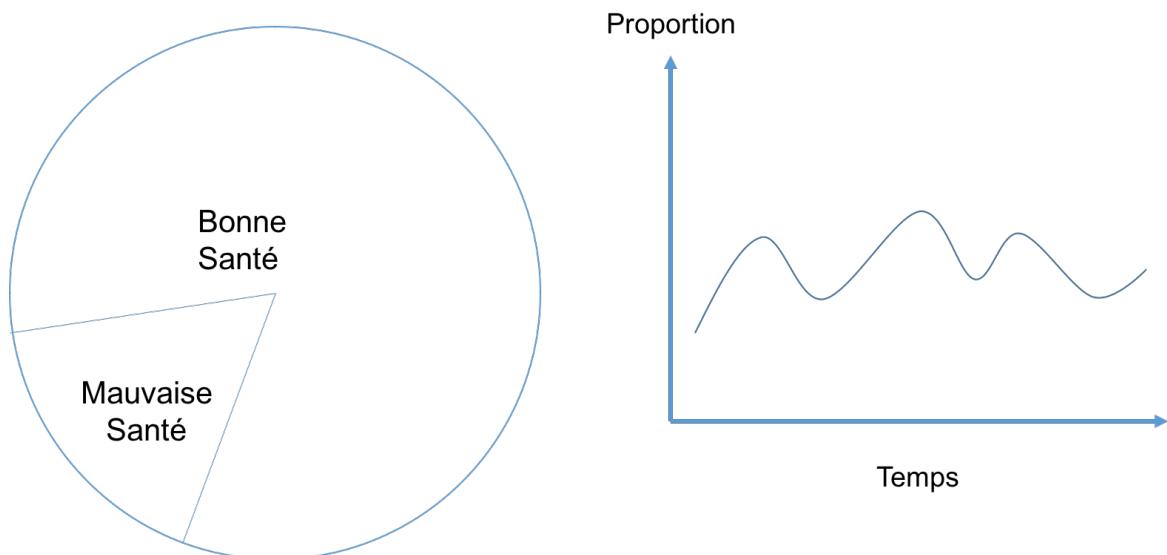


FIGURE 1.3 – Description de l'état de santé des populations

### Difficultées

Sous l'impact du changement climatique et de la pollution de l'environnement, de nombreuses nouvelles maladies infectieuses apparaissent et causent de grands dommages à la santé publique comme le SARS, l'Ebola, le Zika, le grippe pandémique (H5N1, H7N9, H1N1, H5N6),... L'épidémiologie doit continuellement traiter les problèmes découlant de ces nouvelles infectieuses maladies et fournir des interventions opportunes pour limiter leur propagation au sein des populations.

Une autre limite de l'épidémiologie est la contradiction entre les résultats des recherches de l'épidémiologie. Par exemple, en Janvier 1994, une étude en Suisse a révélé une association statistiquement significative entre l'exposition au radon et le cancer du poumon, contrairement à l'étude menée au Canada. Dans le même temps, en Janvier 1994, une étude sur les électriques américains a révélé qu'il peut y avoir un lien entre des champs électromagnétiques dans les fils électriques et le cancer au cerveau. Cependant, une étude précédente en 1993 en France et au Canada a montré qu'il n'y avait pas d'association entre le champ électromagnétique et la leucémie. Les

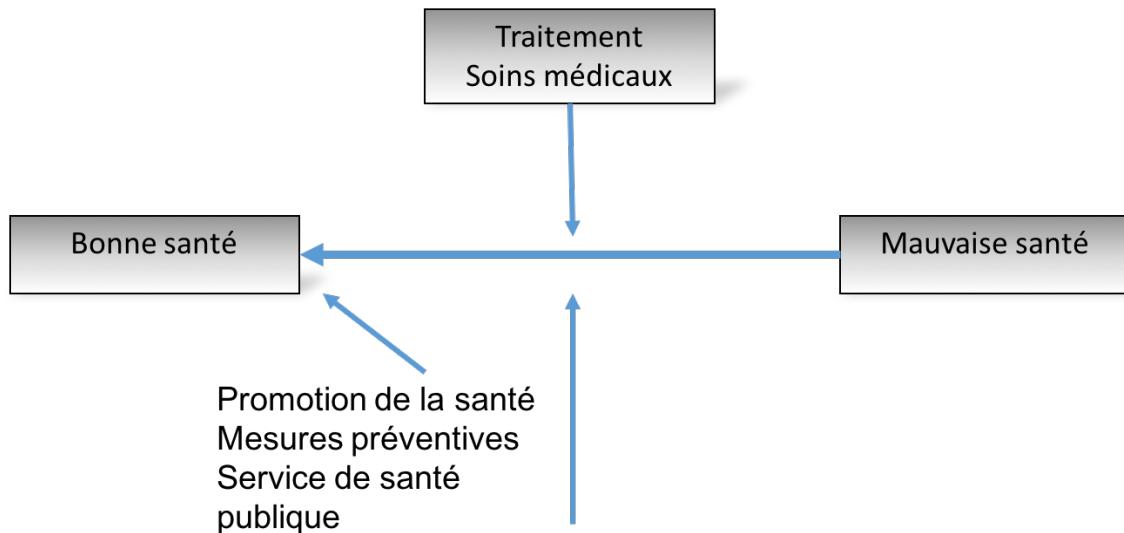


FIGURE 1.4 – Évaluation des interventions

contradictions dans les résultats de recherche ont causé la confusion sur la population et cela a affecté négativement sur le succès des campagnes prophylactique de l'épidémiologie .

### 1.1.3 Les réalisations de l'épidémiologie

#### Tuberculose

Plus de 1 milliard de personnes sont mortes de la tuberculose depuis 200 ans et le nombre de cas de tuberculose est devenu la maladie infectieuse la plus meurtrière de l'histoire de l'humanité. Avec la prévalence de la tuberculose, cette maladie est présente dans tous les pays. Cependant, la tuberculose affecte beaucoup les pays en cours de développement, où la maladie n'est pas bien contrôlée et le taux d'inflection est élevé.

En fait, l'incidence de la tuberculose et leurs décès a diminué régulièrement depuis les années 1900. Avant que l'homme a trouvé le spécifique contre la tuberculose, l'amélioration de l'hygiène, du logement et de la vaccination ont entraîné une réduction des taux d'infection dans certains pays. Mais seulement lorsque les antibiotiques sont nés, la tuberculose devient une maladie qui peut être contrôlée et guérie. Sans antibiotiques, jusqu'au 70% des patients tuberculeux seront morts. Le vaccin antituberculeux a été existé depuis un siècle. Mais il ne se révèle efficace que dans la prévention de la tuberculose chez les enfants. Les vaccins antituberculeux ne sont pas efficaces lorsqu'ils sont utilisés avec les adultes.

En général, la tuberculose est toujours sous contrôle avec les progrès du traitement. Cependant, la tuberculose est encore assez complexe dans certains pays en développement. Des ressources financières et politiques seront nécessaires pour soutenir le contrôle de la tuberculose.

## VIH/AIDS

Au début des années 1980, aux États-Unis, certains hommes gais ont eu une infection anormale. Dans les années suivantes, la maladie est devenue de plus en plus commune. Les chercheurs ont finalement découvert un virus qui causait l'immuno-déficience chez eux. C'est le VIH - un virus originaire d'Afrique qui infecte uniquement par des primates. Dans 100 dernières années, le VIH a commencé à infecter les humaines et s'est rapidement propagé dans le monde entier.

Le VIH se propage dans le sang, principalement par des activités sexuelles non sécuritaires, de transfusions sanguines provenant de sources infectées, de partage d'aiguilles, de la mère à l'enfant pendant la grossesse, l'accouchement ou l'allaitement.

Une personne infectée par le VIH non traitée doit devenir le SIDA, également connu sous le nom de syndrome d'immuno-déficience acquise. La maladie détruit le système immunitaire humain, offrant des possibilités d'infections opportunistes et causant la mort.

Il n'y a pas de vaccins disponibles pour prévenir l'infection par le VIH, et aucun traitement ne peut éliminer complètement le virus du VIH de l'organisme. Cependant, les personnes vivant avec le SIDA peuvent maintenant étendre et améliorer leur qualité de vie grâce à la thérapie anti-rétrovirale (ART). ART est la thérapie qui utilise des médicaments anti-rétroviraux, qui ralentissent la réPLICATION du VIH dans le corps. Par conséquent, le traitement anti-rétroviral augmentera l'immunité et réduira les risques d'infections opportunistes chez les patients infectés par le VIH.

Actuellement, les scientifiques travaillent toujours pour trouver un remède et arrêter le VIH / SIDA. Ils espèrent d'arrêter cette pandémie au 2030.

## Variole

La variole ou petite vérole est une maladie infectieuse d'origine virale, très contagieuse et épidémique, due à un poxvirus. Il est apparu chez les humains depuis des millénaires et devient un facteur majeur dans l'histoire de l'humanité. À la fin des années 1960, la variole était encore complexe en Asie et en Afrique, avec environ 2 millions de décès chaque année.

La variole se propage facilement entre les gens, par les éternuements ou les contacts occasionnels. Il provoque des symptômes simples avec des pustules sur la peau, et peut entraîner des complications causées par les nombreuses cicatrices comme l'arthrite, la cécité, la mort... Au XVIII<sup>e</sup> siècle, un médecin britannique, Edward Jenner, a découvert que les vaches pouvaient également être infectées par le virus de la variole. Les personnes qui sont infectées par le virus de la variole chez les vaches peuvent facilement être guérie et ils sont également immunisées contre la variole humaine. Jenner devina que c'était une type de virus variola plus légère. Il a extrait les croûtes d'un bouton d'une personne qui a été infecté par la variole chez les vaches et les a injectées dans les bras d'un garçon de huit ans. Étonnamment, ce garçon était immunisé contre la maladie. Jenner est devenu l'invention de la vaccination et aussi le créateur du vaccin contre la variole.

Grâce à la méthode vaccination de Jenner, l'homme a trouvé et développé des méthodes de prévention de la variole à l'échelle mondiale. Les campagnes de vaccination sont poursuivies du *XIX<sup>e</sup>* siècle au *XX<sup>e</sup>* siècle. Jusqu'au 8 Mai 1980, l'éradication de la variole a été proclamée par l'Organisation Mondiale de la Santé (OMS).

### **Grippe**

Au cours des derniers siècles, la grippe a également causé plus de décès que le VIH/AIDS. Les pandémies saisonnières de grippe se produisent chaque année et touchent environ quatre millions de personnes, avec environ 250 000 décès dans le monde entier. Au cours du dernier siècle, il y avait des nouveaux types de virus de la grippe qui étaient propagés à l'homme à partir de la volaille ou du bétail comme les grippes H1N1, H5N1, H5N6, H7N9,... Ces types de virus sont complètement nouvelles et ils créent des pandémies dévastatrices lorsqu'ils commencent à se propager entre les hommes.

Un exemple typique est la pandémie de la grippe A/H1N1 en 1918, également connue sous le nom de grippe espagnole, qui a causé la mort de plus 50 - 100 millions de personnes, soit 2 - 5 fois plus grande que le nombre des victimes de la première guerre mondiale.

De nos jours, l'épidémie de grippe a lieu partout dans le monde. Selon l'OMS, cette maladie survient chaque année, entraînant la situation entre 3 millions et 5 millions de personnes atteintes de la grippe sévère et environ 250 000 à 500 000 décès chaque année dans le monde. Il n'existe pas encore de vaccins qui peuvent lutter contre la variation des nouveaux types de virus de la grippe. C'est pourquoi l'OMS a donné des notifications en haut niveau qui appellent les pays à surveiller la situation de la grippe afin de réduire la possibilité d'explosion d'une épidémie de grippe pandémique à l'échelle mondiale.

Les quatre maladies ci-dessus sont des maladies transmissibles plus typiques dans les 100 dernières années. En fait, il existe d'autres maladies qui peuvent également figurer sur cette liste, notamment la poliomyélite, le paludisme, le choléra et la syphilis.

#### **1.1.4 Dengue**

La dengue est une infection virale qui est endémique dans les pays tropicaux. Elle est transmise entre les êtres humains par l'intermédiaire des moustiques. Cette infection virale entraîne classiquement fièvre, maux de tête, douleurs musculaires et articulaires, fatigue, nausées, vomissements et éruptions cutanées. La guérison survient généralement en une semaine. À côté des symptômes cités là, il existe des formes hémorragiques ou avec syndrome de choc, rares et sévères, pouvant entraîner la mort.

Dans le passé, les africains ont utilisé le mot Ka-dinga pepo pour appeler la maladie dengue qui est transmise par le moustique Aedes Aegypti. En Swahili, cet mot signifie "malade du diable" parce qu'il peut créer le bouleversement et la débilité de l'intérieur du corps des victimes qui mènent à l'hémorragie des organes internes. La maladie

Ka-dinga pepo n'a été apparu que dans les régions tropicales et subtropicales. Au cours des siècles, cette maladie a été propagée partout dans le monde entier à cause de la développement de la circulation et de la commerce. Jusqu'au XVIII<sup>e</sup> siècle, en 1780, elle a causé une pandémie sur la province Philadelphie d'Amérique.

Au dernier du XX<sup>e</sup> et début du XXI<sup>e</sup> siècle, la dengue est devenu une menace très dangereux avec la vitesse de propagation très horrible dans le monde entier, même aux États-Unis avec le système sanitaire le plus énergique. Le taux de propagation de la dengue est devenu très sérieux ces dernières années. Avant 1970, il y avait que 9 pays qui sont touché par cette maladie mais cette nombre a été quadruplé en 1995. Selon l'Organisation mondiale de la santé (OMS, 2007), environ 40% de la population mondiale dans 112 pays du monde est exposée au virus. On estime à 390 millions le nombre de nouvelles infections chaque année[6].

Afin de prévenir la maladie, les premiers vaccins développés par des scientifiques japonais et américains après l'identification du virus. La problème principal est que la dengue a non seulement un mais 4 type de virus : Dengue 1, Dengue 2, Dengue 3, Dengue 4. Une zone peut être infecter par un, deux ou plusieurs types de virus en même temps. De plus, ces types de virus se mutent souvent très rapidement, rendant le système immunitaire humain difficile à résister. Avec la plupart des types de virus, les personnes qui sont déjà infecté par une malade ont souvent des anticorps qui aident le corps à la résistance lors d'infections ultérieures. Les vaccins sont construits sur le même mécanisme. Cependant, en raison de la complexité du virus, les personnes atteintes d'anticorps Dengue 1 peuvent toucher aux trois autres types de virus Dengue.

Avec le développement de la technologie moderne, les scientifiques ont obtenu quelques succès dans l'étude des vaccins contre la dengue. Dans la deuxième guerre mondiale, les scientifiques avaient utilisé Dichloro-Diphényl-Trichloréthane (DDT) pour exterminer les moustiques. Cela a contribué à arrêter la fièvre dengue et malaria au centre et au sud d'Amérique. Pourtant, depuis la fin des années 1960, le DDT a été abandonné en raison de problèmes sanitaires et environnementaux. Ceci est entraîné le retour des moustiques et leurs maladies infectieuses. Récemment, le Dengvaxia dengue vaccine a été distribué au Mexique, au Brésil, au Salvador et aux Philippines. Cependant, leurs efficacité ne sont pas élevée car ils existe des inconvénients et le coût sont très élevée. Jusqu'à maintenant, le monde n'a pas encore trouvé un bon et efficace vaccin pour cette maladie. C'est la raison pour laquelle les domaines d'épidémiologie ont activées les recherches et les analyses des impacts et des interactions environnementaux sur le vecteur le plus dangereux dans l'histoire humanité.

## 1.2 L'état de l'art

Un certain nombre d'études ont cherché à caractériser les influences du climat, de la démographie humaine et du comportement sur l'épidémiologie des maladies infectieuses, en utilisant différentes méthodologies. Zhang et al. [7], Cumming et al. [8], Barreto et al. [9] ont utilisé des modèles mathématiques pour étudier l'effet du changement climatique, de la transition démographique et de la structure urbaine sur la transmission de la dengue. Un résultat global est que les relations quantitatives

entre le climat et les maladies à transmission vectorielle sont incohérentes entre les différentes études. Certaines données établissent un lien étroit entre l'incidence de la dengue et El Niño dans les îles du Pacifique [10, 11] et en Thaïlande [12], tandis que d'autres constatent que l'augmentation de la densité de population et l'insuffisance des sources d'eau sont les principaux facteurs de l'épidémiologie de la dengue [13, 14]. En analysant les données mensuelles de la dengue hémorragique (DHF) en Thaïlande de 1983 à 1997, Cumming et al. [15] a montré que les zones rurales à faible densité de population peuvent également connaître de graves épidémies.

L'Asie du Sud-Est est une région de circulation dengue particulièrement intense [16], et c'est aussi le cas du Vietnam. Le Vietnam a cette particularité supplémentaire d'avoir une diversité importante de climats sur une zone relativement petite ( $330\,000\text{ km}^2$ ). Cette diversité climatique est générée par une grande latitude orthogonale à une large gamme d'élévations depuis le niveau de la mer sur la côte Est jusqu'à plus de 3 000 m sur la frontière Ouest. Schmidt et al. [17] a réalisé une étude de cohorte et une analyse spatiale au Vietnam et a montré que le risque de dengue était plus élevé dans les zones rurales avec une mauvaise alimentation en eau que dans les zones urbaines à forte densité de population. Do et collaborateurs [18, 19] ont étudié la dengue dans la ville de Hanoi entre 2004 et 2009 et ont mis en évidence une forte saisonnalité de l'épidémiologie de la dengue avec des incidences élevées entre juin et novembre, saison des pluies qui correspond également à des températures élevées. Le et al. [20] a analysé l'influence du tourisme sur l'incidence de la dengue sur l'île de Cat Ba (nord du Vietnam) de septembre à novembre 2013. Ils ont montré que la transmission de la dengue est peu probable sur l'île de Cat Ba. l'introduction de virus de la partie continentale, très probablement Hanoi. Cazelles et ses collaborateurs [21] ont analysé la dynamique de l'incidence de la dengue dans la province de Binh Thuan, au sud du Vietnam. Ils ont utilisé la décomposition en ondelettes pour détecter et quantifier la périodicité de la dengue entre janvier 1994 et juin 2009 et pour décrire le profil de la synchronie dans le temps et l'espace. Ils ont également utilisé cette méthode pour explorer la relation entre l'incidence de la dengue et l'oscillation australe El Niño (ENSO) et ont trouvé une forte association non stationnaire entre les indices ENSO et les variables climatiques pour la période de 2-3 ans.

Ces études ont montré l'influence des facteurs environnementaux sur la transmission de la dengue épidémie. Cependant, ce sont des analyses locales à petites échelles. Il n'existe pas encore une étude à grande échelle permettant d'analyser spécifiquement la relation entre les facteurs environnementaux et la propagation de cette épidémie.

### 1.3 Notre proposition

Notre premier objectif était d'effectuer un pré-traitement sur les données, puis le visualiser pour avoir un aperçu sur la situation d'épidémie du région Sud-Est d'Asie durant la période 17 ans. Ensuite, nous appliquons les méthodes classification pour grouper la situation d'épidémie des provinces des pays dans la région Sud-Est d'Asie.

Notre prochain objectif est déterminer la relation entre les facteurs environnementaux et la développement de l'épidémie durant la période 1994 au 2010. Plus précisément, les facteurs d'environnementaux sont les facteurs climatiques comme les températures (average, maximal, minimal), l'absolute humidité, la relative humidité, la pluviosité,

l'heure de soleils. Parmi les pays dans la région Asie du Sud-Est, nous avons choisi le pays Vietnam où leurs données sont les plus complètes. Nous avons utilisée les méthodes d'analyse ondelette, causalité de Granger et convergence cross mapping pour analyser les relations entre l'incidence de la dengue et les facteurs climatiques sur 64 provinces et 67 station climatique au Vietnam.

## 1.4 La structure de la thèse

La présentation des travaux réalisés au cours de cette thèse est organisée de la manière suivante :

Dans le deuxième chapitre, nous rappelons les connaissances de base sur les séries temporelles car notre données sont présente sous formes des séries temporelles. La structures des données d'épidémie de la dengue et des facteurs climatiques sont présente dans la partie suivant de cette chapitre. Ensuite, la processus de pré-traitement sur les données sont présente à la fin de cette chapitre.

Les algorithmes de clustering sont présentées dans le chapitre 3. Nous parlons tout d'abord le manière de calculation la distance entre les séries temporelles. Puis, le méthode classification k-means a été appliqué pour grouper les provinces des pays dans la région Asie du Sud-Est basé sur leurs taux d'infection de l'épidémie de la dengue.

Dans la chapitre 4, nous avons analysé les relations entres les facteurs environnementaux et l'incidence de la dengue en utilisant la méthode d'analyse ondelette. Cette relation est analysées plus détaillé aux chapitre 5 en appliquant la méthode d'analyse Granger causalité univariable et multivariable.

En outre, les relations potentielles entre les facteurs environnementaux sont clarifiées au chapitre 6 par l'utilisation la méthode Convergence Cross Mapping. En fin, la conclusion et notre perspective sera présenté dans la chapitre conclusion.

# **Chapitre 2**

## **Données**

### **2.1 Les séries temporelles**

#### **2.1.1 Définition**

Une série temporelle est une série de points de données indexées (ou répertoriés ou représentés graphiquement) dans l'ordre chronologique. Plus précisément, une série temporelle est une séquence prise à des points successifs espacés régulièrement dans le temps ou on peut rappelé une séquence de données en temps discret. Les séries temporelles sont utilisées dans les statistiques, le traitement du signal, la reconnaissance des formes, l'économétrie, la finance mathématique, la prévision météorologique, l'électroencéphalographie, l'ingénierie de contrôle, l'astronomie, l'ingénierie des communications et dans tous les domaines de la science appliquée.

Une série chronologique normale comprend 4 parties :

- Niveau : La valeur de référence pour la série s'il agissait d'une ligne droite.
- Tendance : Le comportement croissant ou décroissant optionnel et souvent linéaire de la série dans le temps.
- Saisonnalité : Les motifs ou les cycles répétitifs facultatifs de comportement dans le temps.
- Bruit : La variabilité optionnelle des observations qui ne peut être expliquée par le modèle.

Toutes les séries temporelles ont un niveau, la plupart ont du bruit, la tendance et saisonnalité sont optionnelles.

#### **2.1.2 Analyse des séries temporelles**

L'analyse de séries temporelles comprend des méthodes d'analyse de données de séries temporelles afin d'extraire des statistiques significatives et d'autres caractéristiques des données. Il existe de nombreuse des applications de l'analyse des séries temporelles et ils sont largement utilisés dans les différences domaines de recherche.

L'application qui est utilisé plus souvent est la prédition de série temporelle. Il utilise un modèle pour prédire des valeurs futures basées sur des valeurs

précédemment observées. Dans le traitement statistique classique sur des données de séries temporelles, la prédiction sur le futur s'appelle l'extrapolation. Une distinction importante dans la prédiction est que l'avenir est complètement indisponible et doit seulement être estimé à partir de ce qui est déjà arrivé. La compétence d'un modèle de prévision de série temporelle est déterminée par sa performance à prédire l'avenir. Cela se fait souvent au détriment de pouvoir expliquer pourquoi une prédiction spécifique a été faite, des intervalles de confiance et encore mieux comprendre les causes sous-jacentes du problème.

Une autre application est l'analyse de régression qui est souvent utilisée de manière à tester les théories selon lesquelles les valeurs actuelles d'une ou plusieurs séries temporelles indépendantes affectent la valeur courante d'une autre série temporelle, ce type d'analyse de séries temporelles se concentre sur la comparaison des valeurs d'une série temporelle unique ou de séries temporelles dépendantes multiples à différents moments.

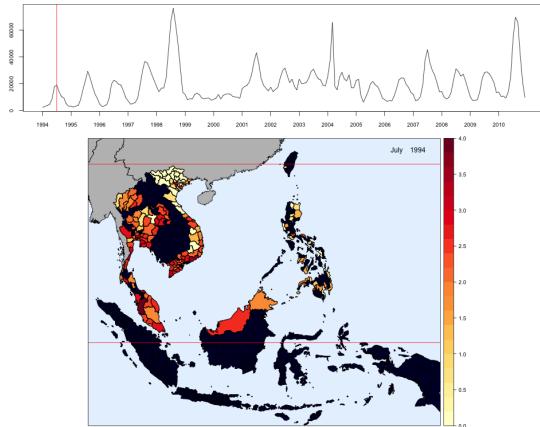
## 2.2 Les données de la dengue

Nous avons utilisé des données mensuelles de surveillance de la dengue provenant des provinces de huit pays couvrant une zone géographique de 3 500 kilomètres d'est en ouest sur 2 500 kilomètres du nord au sud et qui comptaient une population combinée de 320 millions en 2010. Des données mensuelles sur la surveillance de la dengue et les données démographiques et climatiques correspondantes au niveau provincial étaient disponibles pour 273 provinces en Thaïlande, au Cambodge, au Laos, au Vietnam, en Malaisie, en Indonésie, aux Philippines et à Taïwan. Les nombres des cas de dengue mensuels ont été calculés de 1993 à 2010 pour les provinces de Thaïlande, de Malaisie et de Singapour, de 1994 pour les Philippines et le Vietnam et de 1998 pour les autres pays[16]. Nous avons calculé les taux d'infection en divisant le nombre des cas de dengue de chaque mois par le nombre de population typique pour chaque provinces. Les données des zones administratives globales (GADM, [www.gadm.org](http://www.gadm.org)) des pays en Sud-Est d'Asie ont été utilisées pour les visualisation géographiquement. La visualisation des taux d'infection sont présenté sous forme une vidéo.

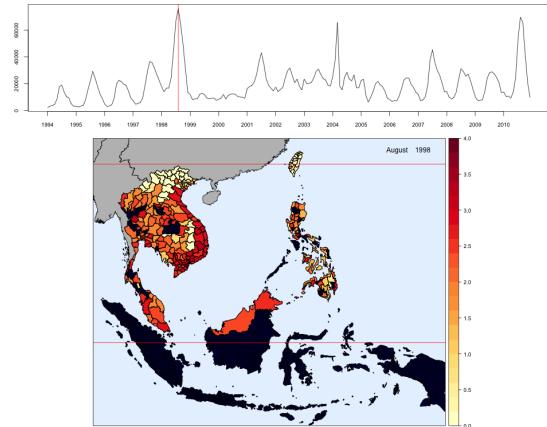
La figure 2.1 représente quelques images export de la vidéo qui exprime des importants situations de la taux d'infection durant la période 1994 - 2010. Le taux d'infection des provinces est représentées de moins au plus d'infection en degré croissant de couleur rouge. C'est à dire les provinces qui ont eu moins infection était en couleur jaune, tandis que les provinces avec le taux d'infection élevé sont eu le couleur rouge. Les provinces manquant les données sont représenté en couleur noir. Les deux ligne rouge horizontal représente le zone tropical. Le premier sommet de l'épidémie de la dengue en Juillet 1994 est présenté en figure 2(a). On peut voir que dans le durée à partir Janvier 1994 jusqu'au Décembre 1997, les taux d'infections viennent des pays comme Thaïlande, Vietnam, Malaisie, Philippines et quelques province de Taïwan. Pendant cette période, le sommet du taux d'infection de chaque années tombait souvent au mois Juillet. La figure 2(b) montre le moment au Août 1998 - la sommet le plus élevé durant toutes la période après l'ajouté les données du Laos et du Cambodge à partir du Janvier 1998. Au début du Janvier 1999, le donné du dernière pays, l'Indonésie, se joindre au modèle et se représente dans la figure 2(c). Ensuite, les deux sommet typique du taux d'infection sont présenté aux figures 2(d) -

Juillet 2001 et 2(e) - Mars 2004. Les données de l'Indonésie sont terminé en Avril 2005 et se représente dans la figure 2(f). Les deux dernières sommet les plus élevés apparaissent en Juillet 2007 (figure 2(g)) et Août 2010 (figure 2(h)). On peut voir que les sommets d'infection des années sont toujours tombés sur les mois Juillet et Août. Les taux d'infection dans des pays tels que le Vietnam, la Thaïlande, la Malaisie et les Philippines sont souvent plus élevés que dans d'autres pays. Cela peut être affecté par les conditions météorologiques de la mousson de la région.

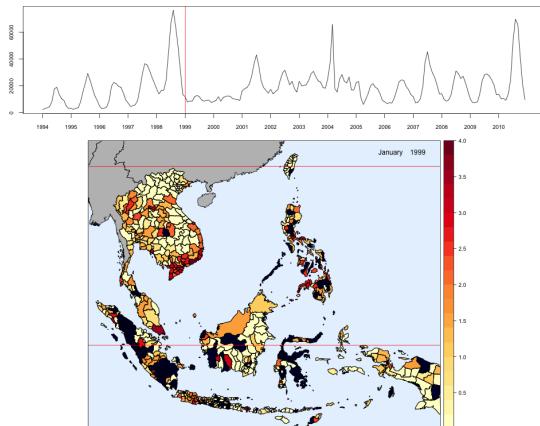
(a) Taux d'infection en Juillet 1994



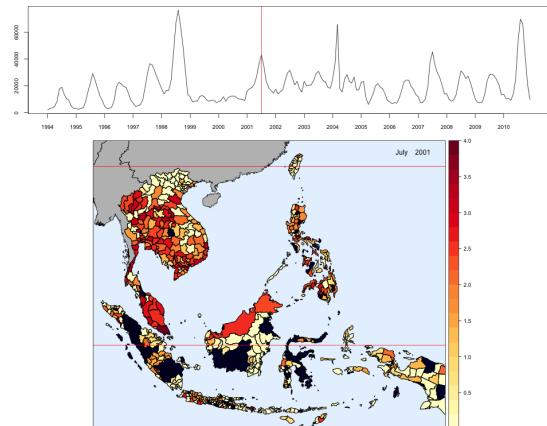
(b) Taux d'infection en Août 1998



(c) Taux d'infection en Janvier 1999



(d) Taux d'infection en Juillet 2001



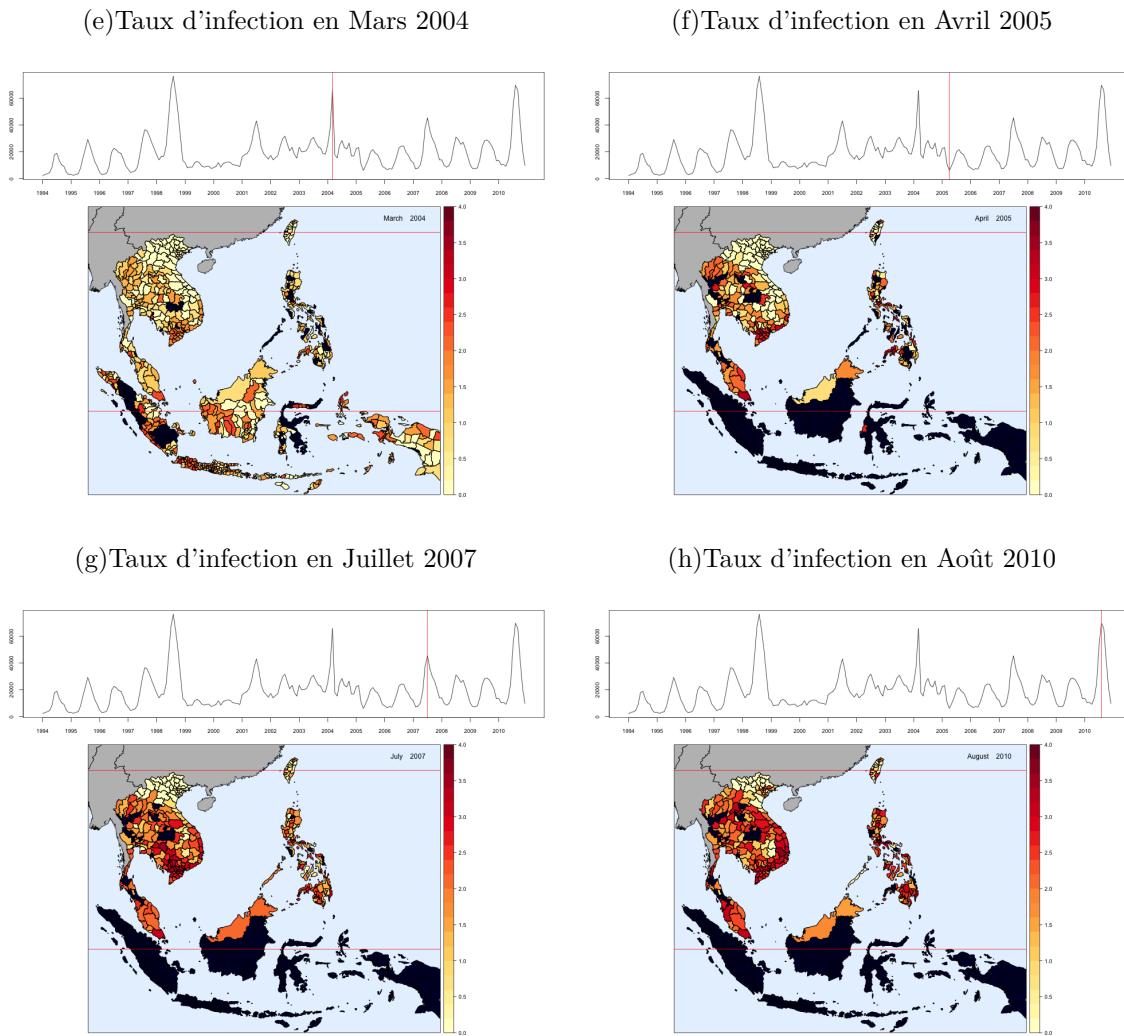


FIGURE 2.1 – Le taux d'infection des provinces des pays du Sud-Est d'Asie à partir Janvier 1994 juste qu'au Décembre 2010

## 2.3 Les données environnementaux

Les données mensuelle d'environnementaux sont collecté des 3 pays : Vietnam, Laos et Thaïlande. Ils contiennent les différentes facteurs climatiques aux différentes périodes. Les données environnementaux du Vietnam est été collecté par 67 station climatiques à partir Janvier 1960 au Décembre 2010. Ils comprennent les variables suivants : Température moyenne (TA), Température maximal (TX), Température minimal (TM), Pluviosité (RF), Humidité relative (RH), L'heure du soleil (SH) et Humidité absolue (AH). Ensuite, au Laos, les données environnementaux sont été collecté par 33 station climatique à partir Janvier 1998 au Décembre 2011. Ils comprennent les variables : Température maximal (TX), Température minimal (TM), Pluviosité (RF), Humidité relative maximal (RHX), Humidité relative minimal (RHM) et L'heure du soleil (SH). Tandis que les données environnementaux du Thaïlande contiennent qu'une seule variable : la pluviosité (RF) à partir Janvier 2000 juste qu'au Novembre 2011. Ces données sont été collecté par 76 station climatique du Thaïlande. Le tableau 2.1 représente la durée, le nombre des stations et les

variables du données environnementaux des 3 pays.

Pays	Nombre des stations climatiques	Période du données	Variables
Vietnam	67	01/1960 - 12/2010	Température moyenne (TA) Température maximal (TX) Température minimal (TM) Pluviosité (RF) Humidité relative (RH) L'heure du soleil (SH) Humidité absolue (AH)
Laos	33	01/1998 - 12/2011	Température maximal (TX) Température minimal (TM) Pluviosité (RF) Humidité relative maximal (RHX) Humidité relative minimal (RHM) L'heure du soleil (SH)
Thaïlande	76	01/2000 - 11/2011	Pluviosité (RF)

TABLE 2.1 – Données climatiques des 3 pays : Vietnam, Laos et Thaïlande

## 2.4 Les données du Vietnam

### 2.4.1 Le Vietnam

Parmi des pays en Asie du Sud-Est, le Vietnam est un pays typique avec les condition géographie particulière où le taux d'infection est relativement élevée. Le terrain du Vietnam est étroit et horizontal qui s'étend du nord au sud. Il borde la Chine au nord, le Laos et le Cambodge à l'ouest, et la mer de l'Est à l'est. C'est pourquoi la topographie et le climat du Vietnam est divisé en 3 régions du Nord, Centre et Sud. Les données au Vietnam sont collecté sur 64 provinces (pour la dengue) et 67 station climatiques (pour les facteurs environnementaux) et ils sont les données le plus complète durée la période. C'est la raison pour laquelle nous choisissons le Vietnam comme une échantillon pour appliquer des méthode d'analyse pour analyser la relation entre la dengue et les les facteurs environnementaux.

### 2.4.2 Les étapes pré-traitements des données du Vietnam

À partir des données brutes initiales, nous avons procédé des des étapes pré-traitements pour la synchronisation de deux types de données en termes de temps et l'élimination des données vierges pour optimiser les résultats obtenus. Avant d'effectue les étapes pré-traitement, il faut d'abord modifier les données de la dengue dans quelques provinces du Vietnam à cause de la séparation et la combinaison des provinces durant 1994 - 2010. Le tableau 2.2 représente les changements des provinces dans la période 1994 - 2010. La modification des données est calculé en fonction de la taux de population des provinces après la séparation / combinaison. Après avoir

Année	Avant	Après
1997	Bac Thai Vinh Phu Ha Bac Hai Hung Ha Nam Dinh Quang Nam - Da Nang Song Be Minh Hai	Thai Nguyen + Bac Can Vinh Phuc + Phu Tho Bac Giang + Bac Ninh Hai Duong + Hung Yen Ha Nam + Nam Dinh Quang Nam + Da Nang Binh Duong + Binh Phuoc Ca Mau + Bac Lieu
2004	Lai Chau Hau Giang Dack Lak	Dien Bien + Lai Chau Can Tho + Hau Giang Dak Lak + Dac Nong
2007	Ha Noi + Ha Tay	Ha Noi

TABLE 2.2 – Changement des noms des provinces au Vietnam durant 1994 - 2010

modifier les données de la dengue à cause de la changement le nom des provinces, nous avons divisé le nombre les cas de dengue de chaque mois de chaque provinces par leurs nombre de population typique pour obtenir le taux d'infection. Puis, nous avons effectué l'interpolation, la normalisation et le retranchement sur le taux d'infection pour obtenir des données prêtes pour les analyses. La figure 2.2 représente les étapes de pré-traitements la donnée de la dengue au Vietnam. D'après les étapes de pré-traitements, on obtenu la donnée prête à analyse dans la période de Janvier 1998 au Septembre 2010. Les données d'environnementaux du Vietnam sont aussi retranché dans le même période pour préparer à l'analyse. La figure 2.3 montre les

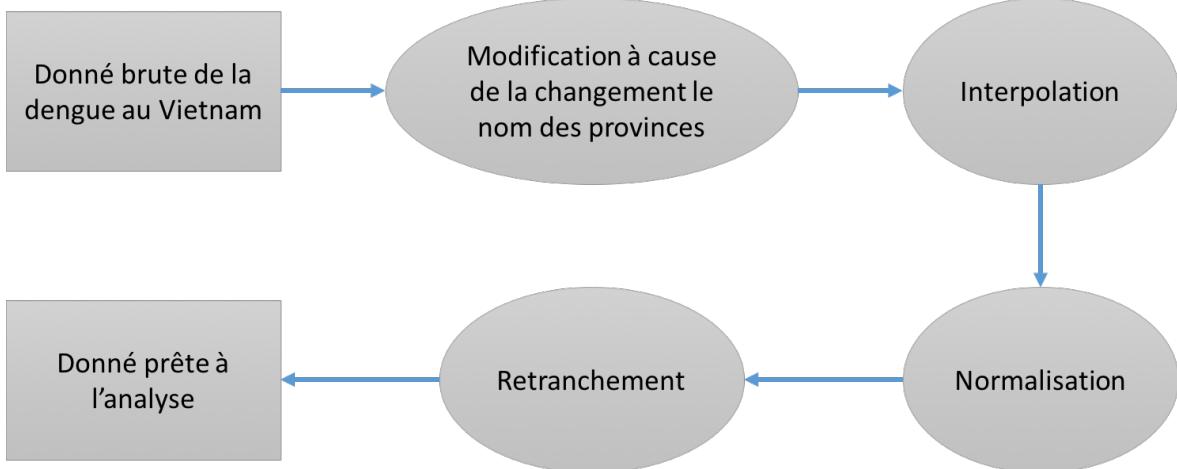


FIGURE 2.2 – Les étapes de pré-traitements la donnée de la dengue au Vietnam

taux d'infection de la dengue pour chaque mois, en moyenne sur les 13 années de la période d'étude et la figure 2.4 montre la même chose pour les 7 variables climatiques des 63 stations climatiques du Vietnam. Nous pouvons voir que le taux d'infection augmente fortement entre juin et novembre dans de nombreuses provinces des régions du sud et du centre-sud. Pendant cette période, les trois variables de température (maximale, minimale et moyenne) augmentent et atteignent leur nadir de façon spectaculaire en juin et juillet (milieu de la saison des pluies) avant de diminuer d'août à décembre (fin de la saison des pluies et début de saison sèche). Pendant la saison des pluies (de mai à octobre), les précipitations et l'humidité relative

augmentent rapidement tandis que le nombre d'heures d'ensoleillement diminue fortement. Le changement apparent des facteurs climatiques et l'augmentation de l'incidence de la dengue entre juin et novembre peuvent avoir une certaine relation. Nous appliquerons des méthodes pour analyser cette relation et représenterons dans les chapitres suivants.

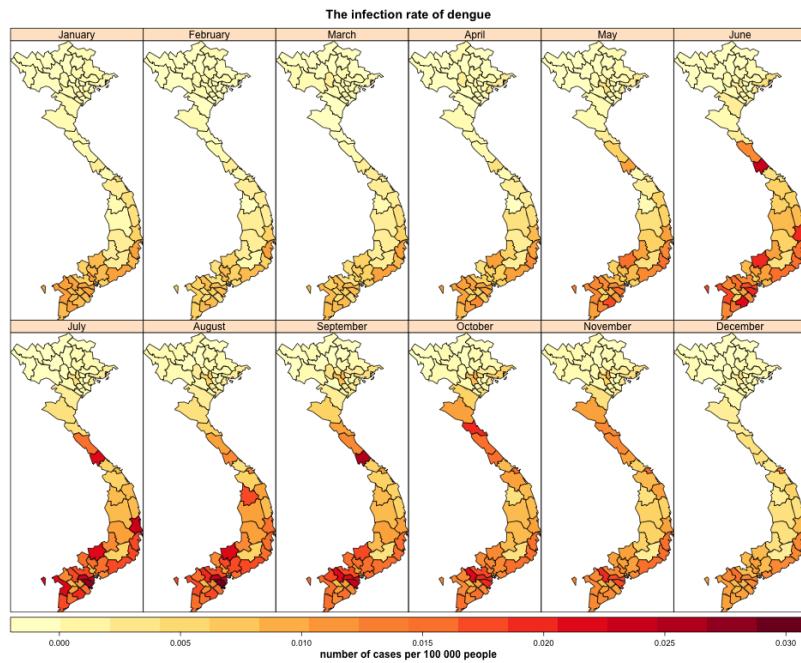


FIGURE 2.3 – Incidence de la dengue par mois et province au Vietnam en moyenne sur les 13 années de janvier 1998 à septembre 2010.

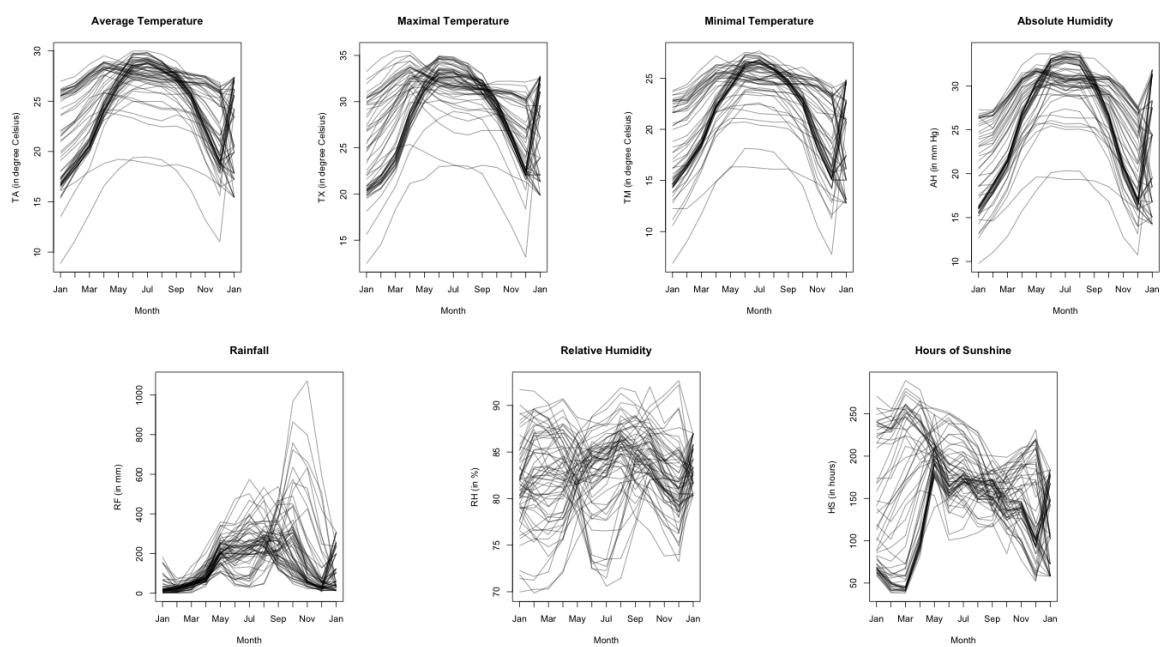


FIGURE 2.4 – Variations des 7 variables climatiques par mois et stations climatiques moyennes sur les 13 années de janvier 1998 à septembre 2010.

## **Chapitre 3**

# **Déetectez des zones épidémiologiques.**

Dans ce chapitre, nous allons procéder à la détection des régions épidémiologique à partir des données mensuelle de l'incidence de la dengue dans 273 provinces du région Asie du Sud-Est. La détection des zones épidémiques est effectué par la méthode clustering k-means sur les séries temporelles de l'incidence de la dengue des provinces observés. Nous représentons d'abord les méthodes pour calculer la distance entre les séries temporelles. Puis, la méthode clustering k-mean et la méthode d'analyse silhouette ont été représenté. Enfin, nous avons montré que la situation de la dengue des provinces sont corrélé dans l'espace en appliquant la méthode k-means sur les données de l'épidémie de la dengue dans la période de Janvier 1998 au Septembre 2010.

### **3.1 Distance entre les séries temporelles**

Les séries temporelles sont des données dans le temps et leur ordonnancement a une signification que l'on ne peut ignorer. Ainsi, on ne peut pas leur appliquer des méthodes de fouille de données classiques qui supposent l'indépendance entre les exemples mais bien des méthodes spécialement adaptées, qui respectent la temporalité de ce type de donnée.

Une mesure de distance entre deux séries temporelles peut être utilisée dans plusieurs tâches de data mining tel que l'apprentissage supervisé et l'apprentissage non supervisé. Dans les bases de données traditionnelles, les mesures de similarité sont basées sur un matching exact. Cependant, dans les données des séries temporelles, caractérisées par leur nature numérique et continue, la mesure de similarité est calculé d'une manière approximative. Nous présentons ici quelques mesures de similarités entre deux séries temporelles :

### 3.1.1 Distance Euclidienne

La distance Euclidienne est une distance géométrique dans cet espace multidimensionnel qui ont beaucoup utilisé dans l'espace Euclidien. Un espace euclidien est un objet algébrique permettant de généraliser de façon naturelle la géométrie traditionnelle développée par Euclide. La distance Euclidienne est calculé comme étant :

$$d_{ED(u,v)} = \sqrt{\frac{1}{N} \sum_{i=1}^N (u_i - v_i)^2}$$

donc  $u$  et  $v$  sont des vecteurs de taille  $N$ . Comme la mesure de similarité ED n'a pas de borne supérieure et que sa valeur augmente avec le nombre de descripteurs N, il est conseillé de calculer la distance ED normalisée. De plus, cette distance ignore les dépendances temporelles entre les différentes séries de données. Ces deux contraintes ne permettent pas de comparer la forme du signal.

### 3.1.2 Complexe invariance distance - CID

La distance CID introduit par Keogh en 2011[22], propose une notion plus robuste que la distance euclidienne, car elle comporte un facteur de correcteur que nous pouvons associer à une nouvelle mesure de dissimilarité. La distance CID est calculer par :

$$d_{CID(u,v)} = d_{ED(u,v)} \frac{\max\{CE(u), CE(v)\}}{\min\{CE(u), CE(v)\}}$$

avec

$$CE(x) = \sqrt{\sum_{i=1}^{N-1} (x_i - x_{i-1})^2}$$

$CE(x)$  est l'estimation de la complexité de la série temporelle x. Le facteur de complexité étant calculé en cumulant l'ensemble des variations locales du signal.

### 3.1.3 Dynamic times wrapping - DTW

Dans le domaine des données énergétiques, de nombreuses études portent sur la comparaison des mesures à différents endroits du bâtiment. A cause du phénomène de propagation de la chaleur, les signaux mesurant une même grandeur physique sont décalés dans le temps. Pour résoudre le problème de distorsion dans les séries temporelles, San-koffet Kurskal [23] ont présenté la distance de déformation temporelle "Dynamic Time Warping (DTW)" qui considère que le temps est élastique et non pas linéaire.

La distance DTW permet de comparer deux séries temporelles de dimension différente. Le principe de la distance consiste à mettre en correspondance les sous-séquences qui "se ressemblent" même si elles ne correspondent pas à un même intervalle de temp. La particularité de la méthode Dynamic Time Warping (DTW) ou déformation dynamique temporelle, est de savoir gérer les décalages temporels qui peuvent éventuellement exister entre deux séries. Au lieu de comparer chaque point

d'une série avec celui de l'autre série qui intervient au même instant t, on permet à la mesure de comparer chaque point d'une série avec un ou plusieurs points de l'autre série, ceux-ci pouvant être décalés dans le temps.

La comparaison de deux séries temporelles U et V de dimensions respectives  $m$  et  $n$ , est basée sur la réPLICATION des valeurs jusqu'à l'obtention de la meilleure correspondance. Pour calculer la matrice de distance ( $mxn$ ), on initiale  $d_{cum}(1, 1) = |u_1 - v_1|$ . Après, on effectue l'initialisation de la premier ligne et la premier colonne de la matrice distance :

$$d_{cum}(1, j) = |u_j - v_1| + d_{cum}(1, j - 1), \text{ pour } j > 1$$

$$d_{cum}(i, 1) = |u_1 - v_i| + d_{cum}(i - 1, 1), \text{ pour } i > 1$$

Les valeurs des autres cases sont calculées comme suites pour  $i + j > 2$  :

$$d_{cum}(i, j) = d_{ED(u, v)} + \min\{d_{cum}(i - 1, j), d_{cum}(i, j - 1), d_{cum}(i - 1, j - 1)\}$$

Pour relier les point U et V, il faut trouver le chemin qui minimise la distance cumulé. Cet chemin W est formé de K point avec  $\max(m, n) \leq K \leq n + m - 1$ . La distance DWT optimale est :

$$d_{DWT(U, V)} = \min \sqrt{\sum_{k=1}^K d_{ED(k)}}$$

Selon ce principe, la DTW a tendance à expliquer les variations de l'axe des Y en déformant l'axe des X. Cela peut cependant induire à des alignements non désirables.

### 3.1.4 Derivative dynamic times wrapping - DDTW

La distance DTW a tendance à expliquer les variations de l'axe des Y en déformant l'axe des X. Cela peut cependant induire à des alignements non désirables. Pour répondre à cette problème, Keogh et Pazzani ont proposé en 2001 une modification de la DTW nommée "Derivative Dynamic Times Warping (DDTW)" [24].

Cette méthode prend en compte la forme des séries temporelles et de la premier dérivée des séquences. Le premier terme dans le calcul de la distance cumulée n'est plus celui de la distance Euclidienne  $d_{ED}(u_i - v_i)$ , mais celui de l'estimation de la dérivée de  $u_i$  et  $v_i$  :

$$d_x(v_i) = \frac{(v_i - v_{i-1}) + \frac{(v_{i+1} - v_{i-1})}{2}}{2}$$

La distance DDTW fournit des performances nettement supérieurs à celle de la DWT original en minimisant le nombre les point dupliqués.

### 3.1.5 Adaptive Feature Base Dynamic Time Warping

Les distances DTW et DDTW ne permettent pas de trouver un alignement dans le cas de données manquantes, ce qui est très fréquent dans les mesures réelles. Xie et Wiltgen a présenté la méthode "Adaptive Feature Base Dynamic Time Warping

(AFBDTW)" en 2010 [25] permet de prendre en compte à la fois le caractère local ainsi que global des séries pour les correspondances au lieu de la valeur elle même ou de sa dérivée.

Le caractère locale de  $u_i$  nommé  $f_{local}(i)$  est défini par :

$$f_{local}(i) = (u_i - u_{i-1}, u_i - u_{i+1})$$

Il semblerait que cette définition représente d'une meilleure façon le caractère global par rapport à la dérivée de la DDTW. Le caractère global est défini par :

$$f_{global}(i) = \left( u_i - \frac{1}{i-1} \sum_{k=1}^{i-1} u_k, u_i - \frac{1}{m-i} \sum_{k=i+1}^m u_k \right)$$

Pour évaluer la distance entre  $u_i$  et  $v_j$ , on définit  $dist(u_i, v_j)$  comme suit. Le calcul de  $d_{cum}$  restera le même que celui défini précédemment.

$$dist(u_i, v_j) = W_1 \bullet dist_{local}(u_i, v_j) + W_2 \bullet dist_{global}(u_i, v_j)$$

avec

$$dist_{local}(u_i, v_j) = |(f_{local}(u_i))_1 - (f_{local}(v_j))_1| + |(f_{local}(u_i))_2 - (f_{local}(v_j))_2|$$

$$dist_{global}(u_i, v_j) = |(f_{global}(u_i))_1 - (f_{global}(v_j))_1| + |(f_{global}(u_i))_2 - (f_{global}(v_j))_2|$$

$$W_1 + W_2 = 1 \quad 0 \leq W_1 \leq 1 \quad 0 \leq W_2 \leq 1$$

Les poids  $W_1$  et  $W_2$  permettent de régler le pourcentage d'influence du critère local et global. La classification des séries temporelles basée sur la mesure de similarité avec la AFBDTW, est le meilleur compromis, permettant de prendre en compte à la fois le critère local et global des séries temporelles.

### 3.1.6 Fast time series evaluation - FTSE

La méthode FTSE a été présenté par M. Morse et Jignesh M. Patel [26] est une technique pour évaluer la seconde classe de fonctions de comparaison de la série de temps qui calcule un score de similarité basé sur un seuil correspondant.

Cette méthode identifie les éléments d'adaptation  $u_i$  et  $v_j$  entre deux séries temporelles  $U$  et  $V$  sans utilisant les grandes matrices à deux dimension. Ceci est réalisé en traitant  $U$  et  $V$  de manière non uniforme, plutôt que de les traiter de la même façon que dans la programmation dynamique.

Pour trouver les paires entre  $U$  et  $V$  sans comparer chaque  $u_i$  avec chaque  $v_j$ , la méthode FTSE indices les éléments de  $U$  sur une grille. Chaque élément de  $U$  est placée dans une cellule de la grille. Maintenant, pour trouver les éléments de  $U$  qui correspondent à  $v_j$ , la grille est sondée avec  $v_j$ . Seuls les éléments de  $U$  qui résident dans la même cellule de grille comme  $v_j$  doivent être comparés avec elle pour voir si elles correspondent.

FTSE mesure la similarité entre les séries temporelles  $U$  et  $V$  avec une valeur de seuil  $\varepsilon$ . En utilisant  $\varepsilon$ , chaque paire d'élément  $u_i \in U$  et  $v_j \in V$  peut être classer comme un match ou un décalage. On peut dire qu'un élément  $u_i$  et  $v_j$  est match si  $|u_i - v_j| < \varepsilon$  dans tous dimensions.

On a montré que la traitement de FTSE a beaucoup plus vite que la traitement des méthodes traditionnelles qui utilisent la programmation dynamique (comme DTW).

### 3.1.7 K-means clustering

K-means clustering est une méthode de quantification vectorielle, à l'origine du traitement du signal, qui est populaire pour l'analyse de cluster dans l'exploration de données. K-means clustering vise à partitionner  $n$  observations en  $k$  clusters dans lesquels chaque observation appartient au cluster avec la moyenne la plus proche, servant de prototype du cluster. L'idée principale est la minimisation d'une fonction objective, qui est normalement choisie comme étant la distance totale entre tous les modèles de leurs centres de cluster respectifs. Sa solution repose sur un schéma itératif, qui commence par des appartenances ou des centres de cluster initiaux choisis arbitrairement. La répartition des objets entre les clusters et la mise à jour des centres de clusters sont les deux principales étapes de l'algorithme c-means. L'algorithme alterne entre ces deux étapes jusqu'à ce que la valeur de la fonction objectif ne puisse plus être réduite.

Étant donné  $n$  motifs  $\{x_k | k = 1, \dots, n\}$ , k-means détermine  $c$  centres de cluster  $\{v_i | i = 1, \dots, c\}$ , en minimisant la fonction objectif donnée comme :

$$\text{Min} J_1(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik} \| x_k - v_i \|^2$$

avec

$$u_{ik} \in \{0, 1\} \forall i, k \quad \sum_{i=1, c} u_{ik} = 1, \forall k$$

et  $\| \bullet \|$  dans l'équation ci-dessus est normalement la mesure de distance euclidienne entre les séries temporelles. Cependant, d'autres mesures de distance pourraient également être utilisées. La procédure de solution itérative comporte généralement les étapes suivantes :

- (1) Choisis  $c$  ( $2 \leq c \leq n$ ) et  $\varepsilon$  (un petit nombre pour arrêter la procédure itérative). Régler le compteur  $l = 0$  et initialiser les centres des cluster  $V^{(0)}$  arbitrairement.
- (2) Distribué  $x_k, \forall k$  pour déterminer  $U^{(l)}$  de sorte que  $J_1$  soit minimisé. Ceci est réalisé normalement en réaffectant  $x_k$  à un nouveau cluster qui en est le plus proche.
- (3) Révisé les centres des cluster  $V^{(l)}$
- (4) Arrêté si le changement de  $V$  est plus petit que  $\varepsilon$ . Sinon, augmenté  $l$  et répétez les étapes 2 et 3.

Bien qu'il puisse être prouvé que la procédure se terminera toujours, l'algorithme k-means ne trouve pas nécessairement la configuration la plus optimale, correspondant à la fonction objective globale minimum. L'algorithme est également sensible aux centres de clusters sélectionnés au hasard. L'algorithme k-means peut être exécuté plusieurs fois pour réduire cet effet.

K-means est un algorithme simple qui a été adapté à de nombreux domaines problématiques. Comme nous allons le voir, c'est un bon candidat pour que l'extension fonctionne avec des vecteurs de fonctions flous.

## 3.2 Analysis de Silhouette

L'analyse de Silhouette est une méthode d'interprétation et de validation de la cohérence au sein de groupes de données [27]. Cette technique fournit une représentation graphique succincte de la façon dont chaque objet se trouve dans son groupe. La valeur de silhouette est une mesure de la similitude entre un objet et son propre cluster (cohésion) par rapport aux autres clusters (séparation). La silhouette valeur varie de -1 à +1, où une valeur élevée indique que l'objet est bien adapté à son propre cluster et mal adapté aux clusters voisins. Si la plupart des objets ont une valeur élevée, la configuration de cluster est appropriée. Si de nombreux points ont une valeur faible ou négative, la configuration de clustering peut comporter trop ou trop peu de clusters.

Supposer que les données ont été regroupées par n'importe quelle technique, tel que k-means, en  $k$  clusters. Pour chaque donnée  $i$ , soit  $x(i)$  la distance moyenne entre  $i$  et toutes les autres données dans le même groupes. Nous définissons ensuite la dissimilarité moyenne du point  $i$  à un cluster  $c$  comme la moyenne de la distance de  $i$  à tous les points de  $c$ .

Soit  $y(i)$  la distance moyenne la plus basse de  $i$  à tous les points de tout autre groupe, dont  $i$  n'est pas membre. Le cluster avec cette dissimilarité moyenne la plus basse est dit le "cluster voisin" de  $i$  parce que c'est le cluster suivant le plus approprié pour le point  $i$ . Nous définissons maintenant une silhouette :

$$s(i) = \frac{y(i) - x(i)}{\max\{x(i), y(i)\}}$$

On peut aussi écrire comme :

$$s(i) = \begin{cases} 1 - x(i)/y(i) & \text{si } x(i) < y(i) \\ 0, & \text{si } x(i) = y(i) \\ y(i)/x(i) - 1 & \text{si } x(i) > y(i) \end{cases}$$

Il est très clair que  $-1 < s(i) < 1$ . Si  $s(i)$  soit plus proche 1, alors  $x(i) \ll y(i)$ . Comme  $x(i)$  est la mesure de la distance moyenne de  $i$  à son cluster, une petite valeur signifie qu'il est bien groupé. De plus, un grand  $y(i)$  implique que  $i$  est mal groupé à son cluster voisin. Ainsi, un  $s(i)$  proche de 1 signifie que les données sont regroupées de manière appropriée. Si  $s(i)$  est proche de négatif, alors par la même logique nous voyons que je serais plus approprié s'il était groupé dans son cluster voisin. Un  $s(i)$  proche de zéro signifie que la donnée est à la frontière de deux groupes naturels.

## 3.3 Clustering les données de la dengue

Dans cette partie, nous avons appliqué la méthode k-means pour grouper les données de la dengue des provinces des pays dans la région Asie du Sud-Est. Nous utilisons les

taux d'inflection des pays Vietnam, Laos, Cambodge, Malaisie, Philippines, Thaïlande, Singapour et Taiwan à partir de Janvier 1998 jusqu'au Septembre 2010 pour effectuer le k-means. Les données des provinces de l'Indonésie n'était pas être utilisés à cause de la carence et de la discontinuité. La distance entre les séries temporelles des provinces sont calculés par la méthode DTW. La méthode d'analyse silhouette a été utilisé pour évaluer le nombre des cluster de la méthode k-means. Le tableau 3.1 montre les silhouettes coefficients (SE) des différents nombres de cluster (NC) de la méthode k-means. On trouve que le nombre de cluster qui a eu la valeur silhouette le plus mieux est égal à 3. Nous avons appliqué la méthode k-means pour grouper les taux d'infections des provinces des pays et leurs visualisation sont présenté sur la figure 3.1. On a traité un test spatial pour évaluer le résultat de la méthode k-means. On a calculé la distance moyenne entre les provinces dans le même groupe en utilisant leurs latitude et longitude. Puis on va créer des séquences de résultats générée aléatoirement en classant les provinces dans les groupes aléatoire. Ensuite on calcule les distances moyennes de ces séquences et leur compare avec celui de la méthode k-means pour obtenir valeur statistique. La figure 3.2 montre la résultat du test statistique de la méthode k-means avec  $k = 3$  et le nombre de répétition des séquence aléatoires est 1.000.000 fois. On a trouvé le p-valeur :  $p < 1e - 6$  et ceci est une résultat très significatif.

Ce résultat suggère que la catégorisation des provinces basée sur la situation de la dengue entre janvier 1998 et septembre 2010 par l'algorithme k-means a donné un très bon résultat lorsque la distance moyenne entre les provinces d'un même groupe est révélée optimale. On peut facilement détecter les région commune comme : Le Nord du Vietnam et de Laos, le centre du Vietnam et de Laos, l'Ouest de Laos et de Thaïlande, le Sud du Vietnam et du Cambodge, le Taiwan, le Philippines, L'est du Malaisie. En raison de la proximité de géographique, ces zones peuvent avoir les mêmes conditions générales de climat. Si les facteurs météorologiques influencent le développement de la dengue, la distribution des provinces sur les cluster dans la figure 3.1 est également relativement appropriée. Cela signifie que la situation épidémique des villes voisines peut être influencée par les autres ou affectée par les mêmes conditions météorologiques et géographiques. Dans les chapitre après nous allons identifier la relation entre les facteurs environnementaux et l'infection de la dengue par les différentes méthodes.

Nombre de cluster	Silhouette coefficient
2	0.5734565
3	0.581752
4	0.5420058
5	0.5307371
6	0.5148082
7	0.4907725
8	0.4794107
9	0.4964826
10	0.4867265
11	0.4732487
12	0.472837
13	0.484026
14	0.473134
15	0.4924166
16	0.4496414
17	0.4425588
18	0.4652832
19	0.4520348
20	0.4081333

TABLE 3.1 – Silhouette coefficient des différents valeurs de clusters

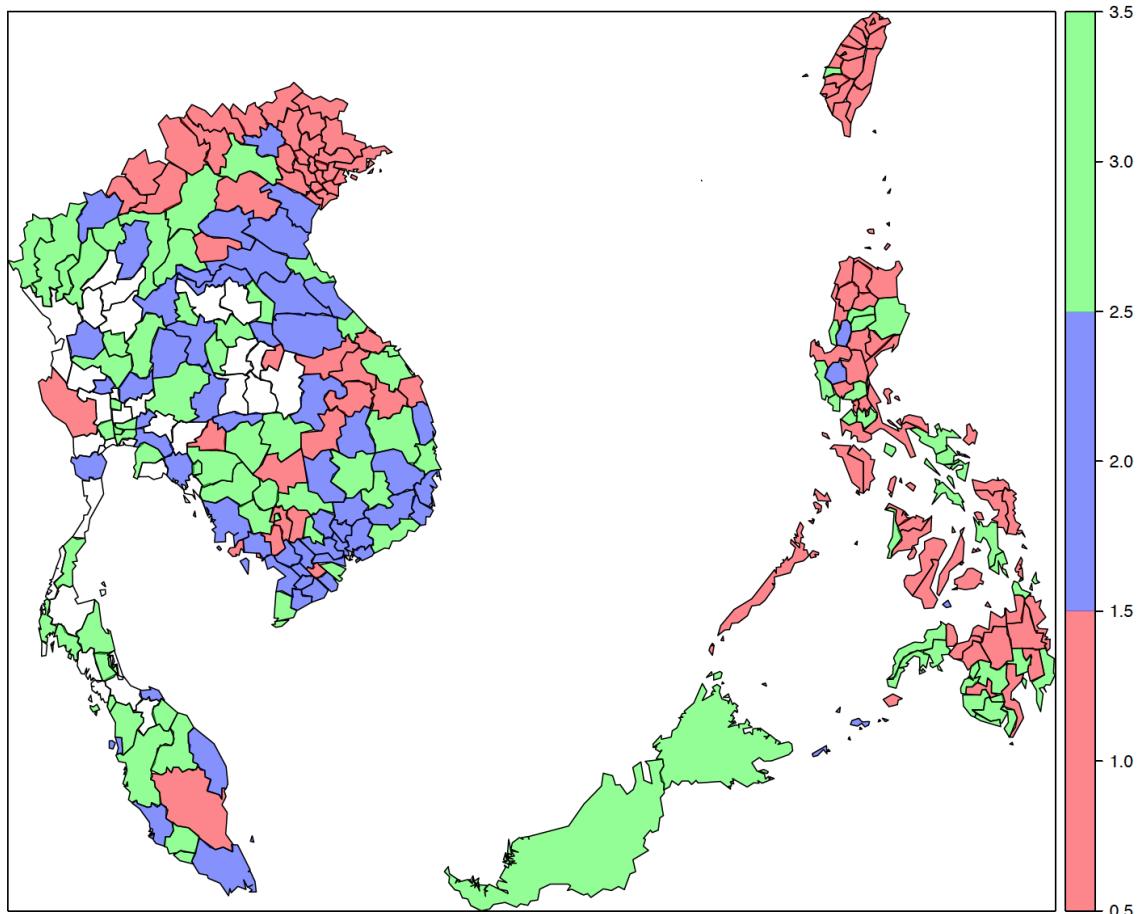


FIGURE 3.1 – Résultat de la méthode k-means avec k = 3, méthode validation : silhouette

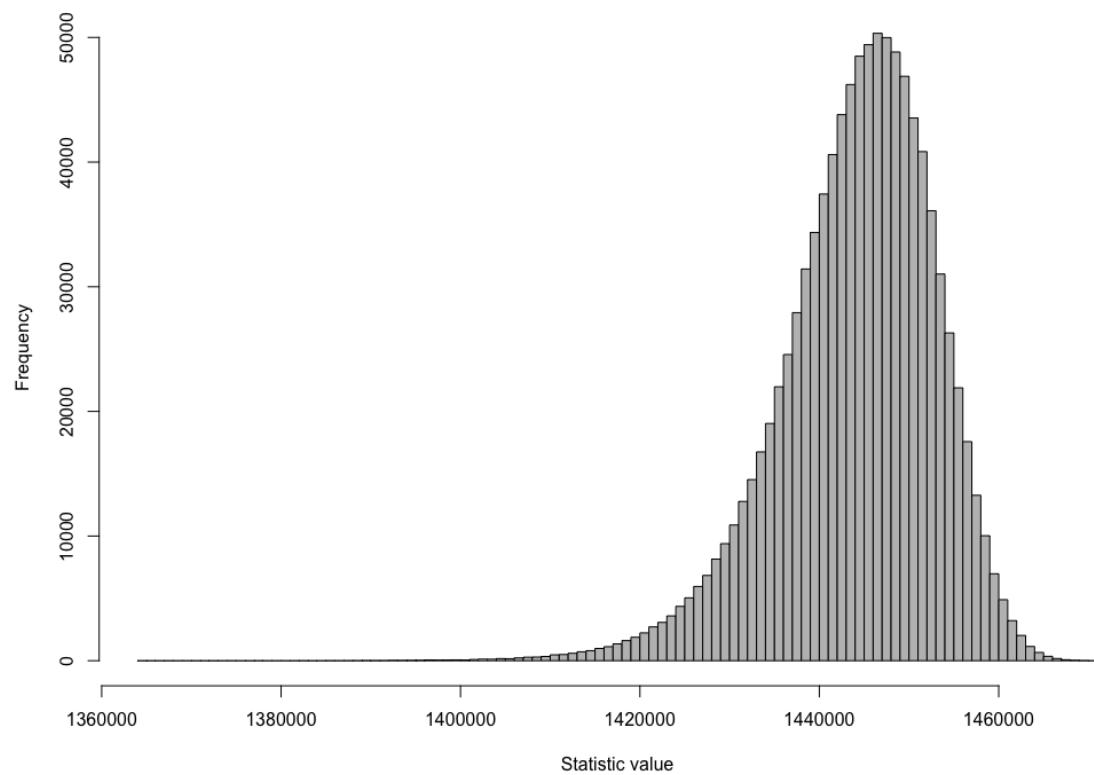


FIGURE 3.2 – Résultat du test statistique de la méthode k-means avec  $k = 3$

## Chapitre 4

# Identifier la relation entre les facteurs climatiques et l'influence de la dengue.

Dans le chapitre 3, nous avons trouvé des zones épidémiques en utilisant la méthode clustering sur la donnée de l'épidémie de la dengue des 273 provinces dans la région Asie du Sud-Est dans la période de Janvier 1998 au Septembre 2010. Les provinces voisins dans des zones d'épidémie peuvent partager les mêmes conditions météorologiques et climatique. C'est la raison pour laquelle nous serons identifier la relation directe des facteurs climatiques sur la développement de la dengue dans cette période en utilisant les méthodes d'analyse ondelette, .

Nous représentons d'abord les théories sur la transformé de Fourier - un outil mathématique important dans la domaine traitement du signal. Puis, la méthodologie de la méthode d'analyse ondelette sera introduit dans la section suivant. Ensuite, les résultats d'expérimentaux doit être représenter dans la dernier section de cette chapitre.

### 4.1 Méthodologie

Dans le domaine traitement du signal, la transformée de Fourier (FT) est un outil mathématique important car elle est un pont de connexion très important pour la représentation du signal entre le domaine spatial et le domaine fréquentiel. La transformée de Fourier est appelée la représentation du domaine de fréquence du signal origine. Le terme transformée de Fourier fait référence à la fois à la représentation dans le domaine fréquentiel et à l'opération mathématique qui associe la représentation du domaine fréquentiel à une fonction du temps. La figure 4.1 montre une exemple sur la représentation les signals de la taux d'infection et le température moyenne du province Cantho - Vietnam et leurs transformée de Fourier.

On peut trouver que les spectre de Fourier montre les composantes de fréquence du signale mais n'indique pas où les fréquences apparaissent. La transformée de Fourier fournit uniquement des information globales et ne convient qu'aux signaux circulaire.

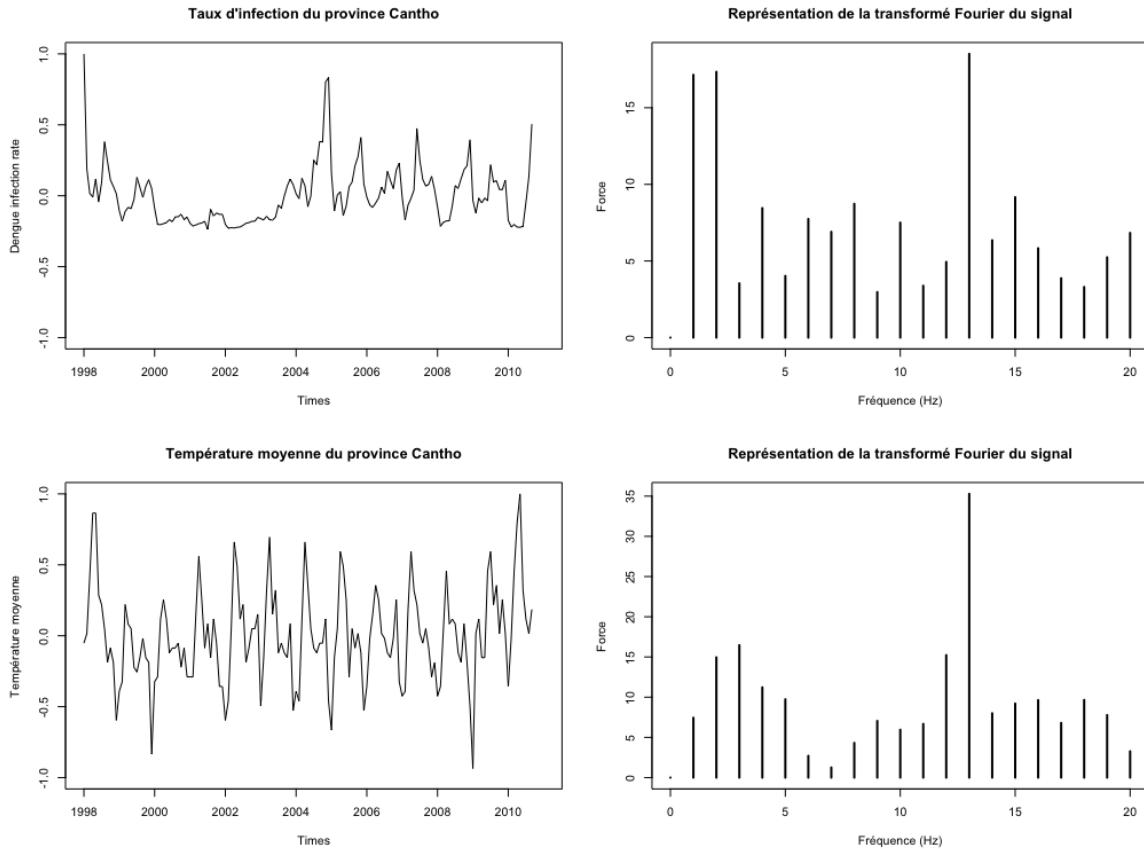


FIGURE 4.1 – Exemple de la transformée de Fourier

Elle n'est pas contenu des mutation ou de changement imprévisibles. Pour surmonter cette défaut, Gabor [28] a appliqué la transformée de Fourier (Windowed Fourier Transform - WFT) par fenêtre à chaque petit segment du signal. Pour surmonter cet inconvénient, Gabor a appliqué la transformée de Fourier de la fenêtre à chaque petit segment du signal (fenêtre) ; Cette transformation montre la relation entre l'espace et la fréquence mais est régie par le principe d'incertitude de Heisenberg pour les composantes à haute fréquence et à basse fréquence dans le signal [29]. En 1975, Zweig a développé une méthode de multirésolution, qui utilise une impulsion oscillante pour redimensionner et comparer des signaux dans les segments individuels. Cette impulsion est appelée une "ondelette" (qui traduit à partir de son origine une petite onde). Cette technique utilise une ondelette contenant des oscillations de basse fréquence et le compare avec le signal analytique pour trouver des information globale du signal dans la résolution brute. Ensuite, on compresse les petites ondes (cette étape est appelé mise à échelle) pour augmenter progressivement la fréquence d'oscillation. Dans les étapes suivant, le signal sera étudié en détail à des résolution plus élevées pour détecter les composants rapides qui reste cachés dans le signal.

La transformée en ondelette est plus flexible par rapport à la transformée de Fourier car il n'est pas nécessaire d'utiliser une fonction d'ondelette fixe et il peut sélectionner les différentes fonctions d'ondelettes pour être adapté au problème qui convient. Les ondelettes qui constituent une famille de fonction dérivée d'une seule fonction qui s'appelle "ondelette mère" dénoté par  $\Psi_{\alpha, \tau}(t)$ . Cette "ondelette mère" a été exprimé sous forme une fonction avec deux paramètre, l'un est le paramètre qui exprimer la position du temps  $\tau$ , l'autre est l'échelle de l'ondelette  $\alpha$ , qui correspond

au fréquence. Plus précisement, l'ondelette est définit comme

$$\Psi_{\alpha,\tau} = \frac{1}{\sqrt{\alpha}} \Psi \left( \frac{t - \tau}{\alpha} \right)$$

Une des formes ondelette qu'on utilise très souvent est ondelette Morlet. L'ondelette Morlet est définie comme

$$\Psi(t) = \pi^{-1/4} \exp(-i2\pi f_0 t) \exp \left( \frac{-t^2}{2} \right)$$

Une transforme d'ondelette d'une série temporelle  $x(t)$  avec une "ondelette mère" est performé comme

$$W_x(\alpha, \tau) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} x(t) \Psi^* \left( \frac{t - \tau}{\alpha} \right) dt = \int_{-\infty}^{\infty} x(t) \Psi_{\alpha,\tau}^*(t) dt$$

où l'astérisque indique la forme conjugué complexe. Le coefficient d'ondelette  $W_x(\alpha, \tau)$  représente le contribution de l'échelle  $\alpha$  du signal donc le temp était dans une différente position  $\tau$ . Le calcul de la transformée en ondelette du signal  $x(t)$  se fait en augmentant le paramètre  $\tau$  sur une rayon d'échelle  $\alpha$  jusqu'à ce que toutes les structures cohérentes dans le signal peuvent être identifiés. Avec l'approche ondelette, on peut estimer la répartition de la variance entre l'échelle  $\alpha$  et la différence emplacement du temps  $\tau$ . Il est connu comme le spectre de puissance ondelette :  $S_x(f, \tau) = |W_x(f, \tau)|^2$ . Pour quantifier les relations statistiques entre les deux séries temporelles, la cohérence des ondelette peut être calculer

$$R_{x,y}(f, \tau) = \frac{| \langle W_{x,y}(f, \tau) \rangle |^2}{| \langle W_x(f, \tau) \rangle |^2 | \langle W_y(f, \tau) \rangle |^2}$$

où les crochets autour des termes indiquent le lissage dans le temps et la fréquence,  $W_x(f, \tau)$  est la transforme ondelette de la série  $x(t)$ ,  $W_y(f, \tau)$  est la transforme ondelette de la série  $y(t)$  et  $W_{x,y}(f, \tau) = W_x(f, \tau) * W_y(f, \tau)$  est la transforme de la coupe d'ondelette. La cohérence des ondelettes fournit des informations locales sur l'endroi où deux signaux non stationnaires,  $x(t)$  et  $y(t)$ , sont linéaire corrélés à une fréquence particulière(ou période).  $R_{x,y}(f, \tau)$  est égale à un quand il y aura une relation linéaire parfaite à un moment donné et la fréquence entre les deux signaux.

## 4.2 Résultats d'expérimentaux

Dans notre recherche, nous appliquons cette méthode pour analyser la relation entre les taux d'inflections et les variables environnementaux des 64 provinces du pays Vietnam dans la période Janvier 1998 au Septembre 2010. La raison pour laquelle nous choisissons le Vietnam et cette période est due à l'exhaustivité et la continuité des données. Pour chaque province au Vietnam, nous calculons la latitude et la longitude moyenne et leurs comparé avec ceux des stations climatiques. Puis nous choisissons le station climatique le plus proche de cette province et le prendre leurs variables climatiques pour calculer la cohérence des ondelette avec le taux d'inflection de la province observé. La figure 4.2 montre la schéma d'application la méthode d'analyse ondelette sur les données du Vietnam.

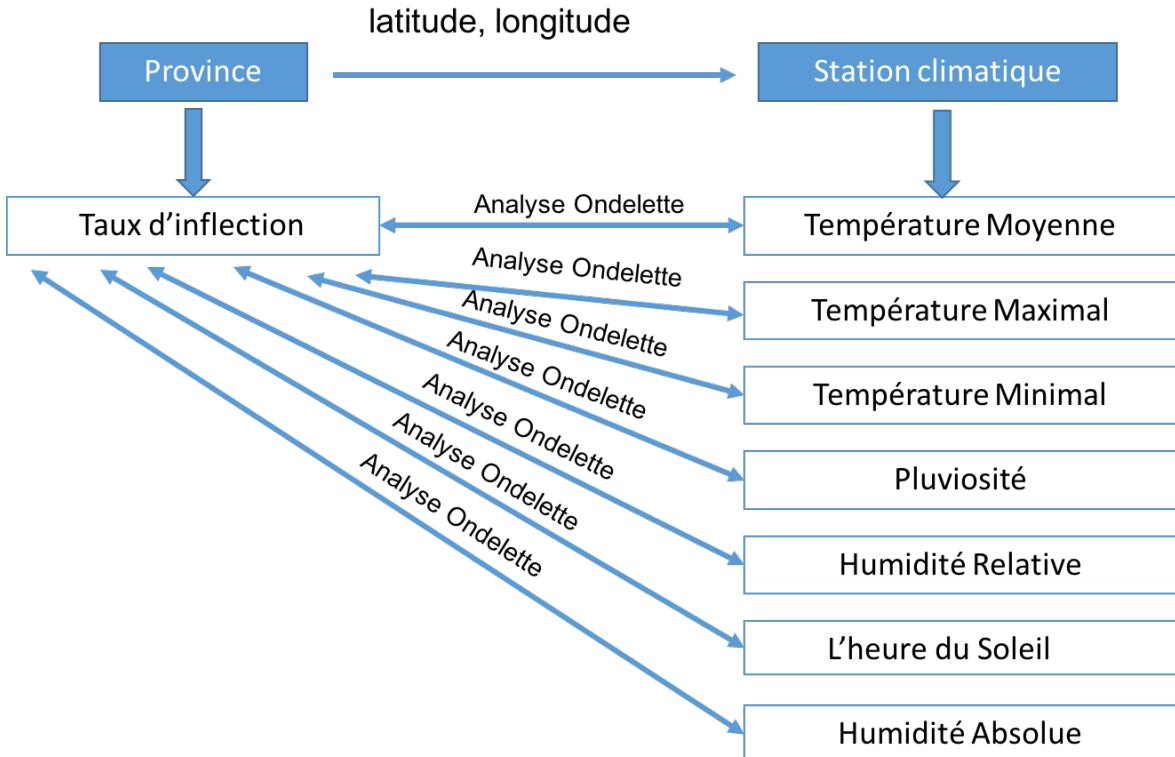


FIGURE 4.2 – La processus d’appliquer la méthode analyse ondelette sur les données du Vietnam

Nous utilisons R version 3.3.1 sur RStudio version 1.1.442 pour calculer les résultats expérimentaux. La méthode d’analyse ondelette est définie dans le package "biwavelet" [?]. Nous choisissons les 3 provinces typiques : Hanoi, Da Nang et Ho Chi Minh qui traversent les 3 régions du Vietnam pour montrer leurs résultats expérimentaux.

La figure 4.3, 4.4, 4.5 montre le résultat expérimental du province Hanoi et la relative humidité, Da Nang et l’heure du soleil, Ho Chi Minh et la température moyenne . Les figures 4.3 (a), 4.4 (a), 4.5 (a) montrent les deux signaux qui ont été normalisés. Le signal de la dengue est en couleur rouge et les facteurs climatiques sont en couleur bleu. La figure 4.3 (b), 4.4 (b), 4.5 (b) montrent la cohérence des deux signaux observés. Les régions en couleurs rouges sont les régions significatives entre les deux signaux. Les flèches avec la direction à gauche qui indiquent que les deux signaux sont en phase l’inverse, à droite qui indiquent que les deux signaux sont en phase. Les flèches pointant vers le haut signifient que le taux d’inflection de la dengue conduit le facteur climatique par  $\pi/2$  et si elles pointent vers le bas signifient que le facteur climatique conduit le taux d’inflection par  $\pi/2$ .

On peut facilement détecter des grandes corrélations entre Hanoi et l’humidité relative. Pendant la période 1998 - 2003 et 2005 - 2010, le taux d’inflection à Hanoi et l’humidité relative sont toujours en phase l’inverse dans la période 2 mois. Ceci est dû au fait que le climat du nord du Vietnam est relativement élevé et humide toute l’année, et que le climat est fortement influencé par la Chine continentale et qu’il a un climat continental.

Dans le cas de Da Nang, une grande province au Centre du Vietnam avec le climat tropical mousson, on trouve une serrée relation entre l’épidémie de la dengue à Da Nang avec l’heure de soleil. Ici le climat de Da Nang avait la transition entre le

climat xavan du Sud et le climat subtropical du Nord du Vietnam. Puis, il existe aussi la mousson du Sud-ouest qui souffle du golfe de Thaïlande et au-dessus des montagnes de Truong Son, il fera un temps chaud et sec pour toute la région. Entre 2003 - 2010, les deux signaux sont corrélé aux période 1 mois et l'heure du soleil conduit la taux d'infection par  $\pi/2$ . Tandis que 1998 - 2010 ils sont totalement corrélé en phase aux période 4 mois.

Le troisième cas est Ho Chi Minh ville, qui est considérée comme la capital du Sud-Vietnam, qui appartient dans le région le climat tropical xavan avec des fortes précipitations et des températures élevées tout au long de l'année. C'est la raison pour laquelle on peut détecter une grande corrélation entre l'épidémie de la dengue et la température moyenne dans cette provinces. Les deux signaux montrent une association dans entre 1998 - 2000 et 2003 - 2010 sur la période 1 mois. La température moyenne a totalement conduit le taux d'infection par  $\pi/2$  dans ces deux temps.

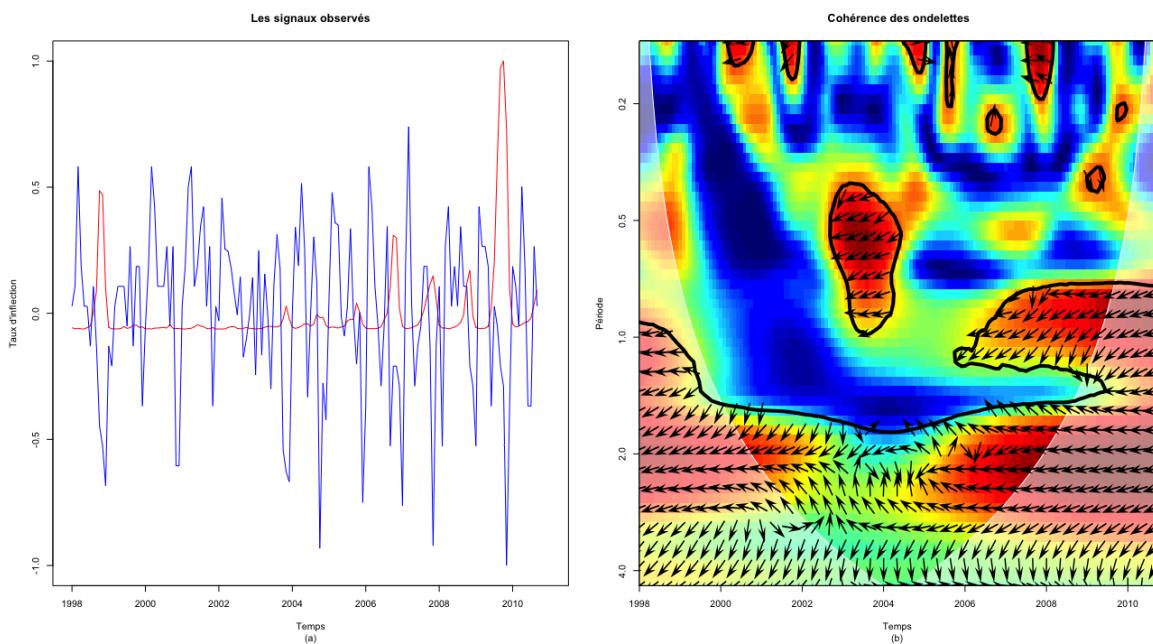


FIGURE 4.3 – Cohérence des ondelettes entre le taux d'inflexion de Hanoi et l'humidité relative. (a) Les signaux observé : le taux d'inflexion est représenté en couleur rouge et l'humidité relative est représenté en couleur bleu. (b) Cohérence des ondelettes entre le taux d'inflexion de la dengue et l'humidité relative de Hanoi, calculée à l'aide de la fonction d'ondelette de Morlet. Les couleurs codent les valeurs de puissance du bleu foncé pour une faible cohérence au rouge foncé pour une cohérence élevée. Les lignes noir imbriquées montrent les niveaux de signification  $\alpha = 5\%$  calculés sur la base de 1 000 séries amorcées. Le cône d'influence indique la région non influencée par les effets de bord.

Les trois résultats ci-dessus montrent une serré relation entre les facteur climatiques les le situation d'infection de la dengue dans des trois grandes provinces du Vietnam. Nous allons ensuite procéder une analyse complète des liens entre les 7 facteurs climatiques et la situation épidémique de la dengue sur 64 provinces du Vietnam pour avoir une point vue global de l'influence des facteurs climatiques sur l'épidémie de la dengue. Pour chaque provinces au Vietnam, on applique la méthode d'analyse ondelette pour calculer la cohérence entre la taux d'infection de cette province avec chaque facteurs climatique. Puis, on calcule le taux significatif en divisant la somme

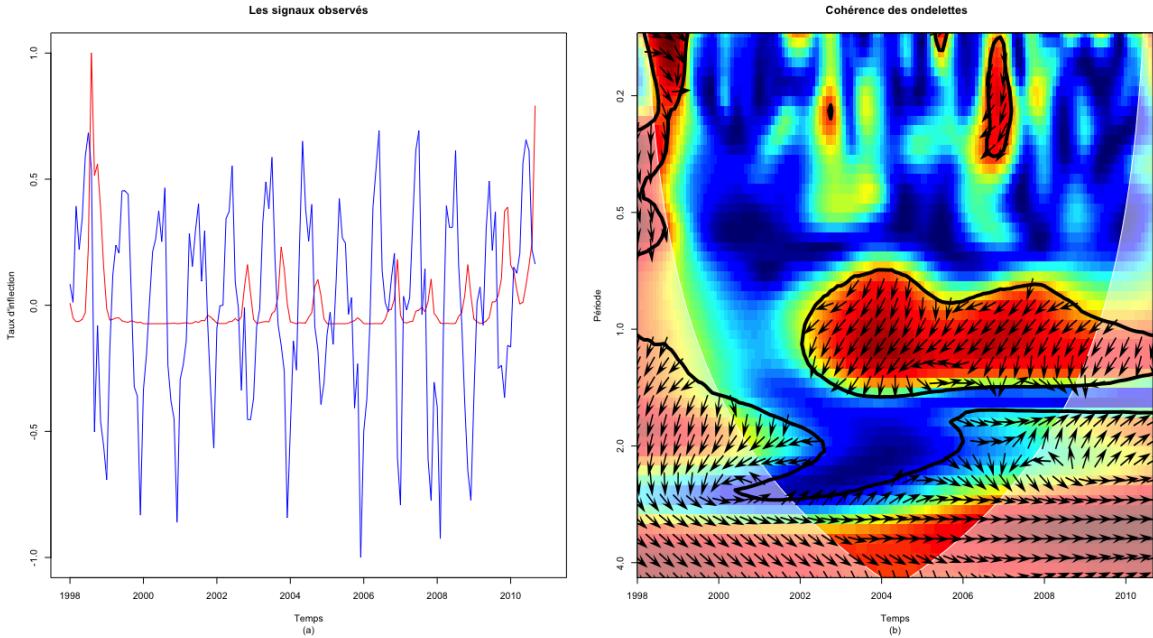


FIGURE 4.4 – Cohérence des ondelettes entre le taux d’inflection de Da Nang et l’heure du soleil.  
(a) Les signaux observé : le taux d’inflection est représenté en couleur rouge et l’heure du soleil est représenté en couleur bleu. (b) Cohérence des ondelettes entre le taux d’inflection de la dengue et l’heure du soleil de Da Nang, calculée à l’aide de la fonction d’ondelette de Morlet. Les couleurs codent les valeurs de puissance du bleu foncé pour une faible cohérence au rouge foncé pour une cohérence élevée. Les lignes noir imbriquées montrent les niveaux de signification  $\alpha = 5\%$  calculés sur la base de 1 000 séries amorcées. Le cône d’influence indique la région non influencée par les effets de bord.

des régions significatifs par l’ensemble des cohérences des 2 signaux. On définit

$$\tau = \frac{R}{C}$$

où  $R$  est la somme total des région significatif parmi des cohérence des 2 signaux et  $C$  est le nombre total des cohérences des 2 signaux. On conclus que la relation entre la situation de la dengue et le facteur climatique d’une province est significatif si son taux significatif est supérieur au seuil  $s$  :

$$\tau \geq s, 0 < s < 1$$

. On a montré les résultats expérimentaux avec les différentes valeurs de  $s$  dans des figures 4.6, 4.7, 4.8, 4.9 avec les seuil  $s = 0.2, 0.3, 0.4, 0.5$ . Un province est été coloré en couleur vert si son taux significatif  $\tau$  est supérieur au seuil  $s$  et en couleur gris si non. Les provinces en couleurs blanches sont des provinces avec les données manquantes. En observant la figure 4.6, on trouve que la plupart des provinces au Vietnam ont le lien entre l’épidémie de la dengue et les facteurs climatiques avec le taux significatif supérieur 20%. Sur la figure 4.7 avec le seuil  $s = 30\%$ , on trouve que la température et l’humidité absolue sont influencé sur les provinces au Sud et au Centre du Vietnam, la pluviosité, l’humidité relative et l’heure du soleil sont influencé sur les provinces au Sud et au Centre et aussi sur les provinces au Nord du Vietnam. Dans le cas où le seuil  $s$  est égale à 40% et 50% dans la figure 4.8, 4.9 , seulement quelques provinces au Sud du Vietnam qui ont le couleur verts. Cela indique que la situation de l’épidémie avaient une serrée relation avec les facteurs climatique sur les provinces au Sud du Vietnam.

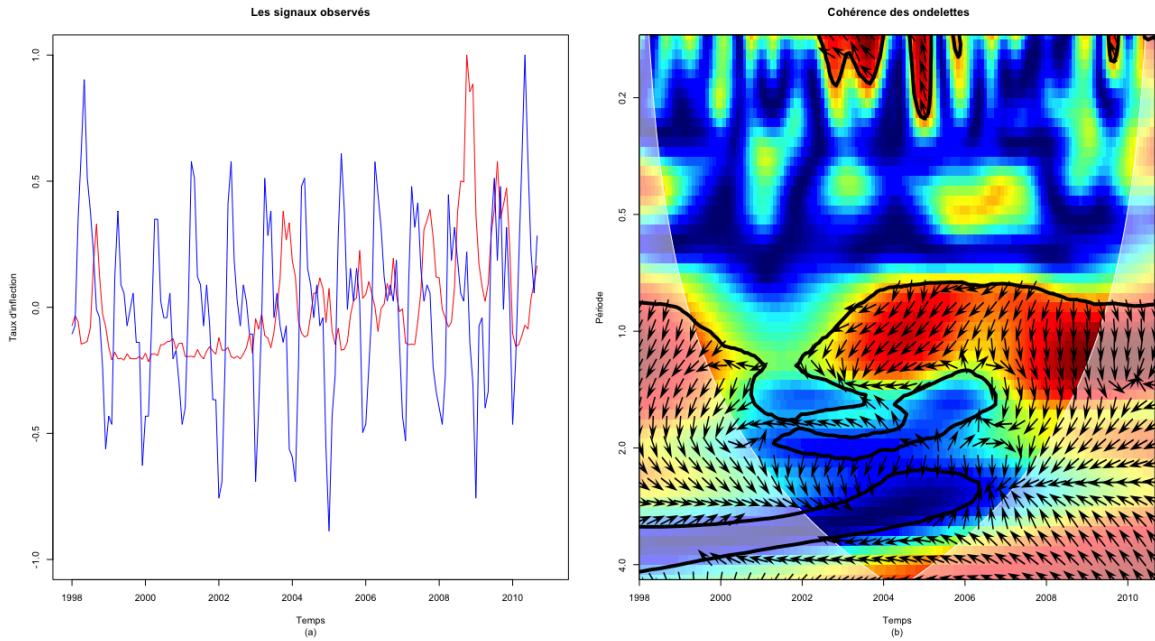


FIGURE 4.5 – Cohérence des ondelettes entre le taux d’inflection de Ho Chi Minh ville et la température moyenne. (a) Les signaux observé : le taux d’inflection est représenté en couleur rouge et la température moyenne est représenté en couleur bleu. (b) Cohérence des ondelettes entre le taux d’inflection de la dengue et la température moyenne de Ho Chi Minh ville, calculée à l'aide de la fonction d'ondelette de Morlet. Les couleurs codent les valeurs de puissance du bleu foncé pour une faible cohérence au rouge foncé pour une cohérence élevée. Les lignes noir imbriquées montrent les niveaux de signification  $\alpha = 5\%$  calculés sur la base de 1 000 séries amorcées. Le cône d'influence indique la région non influencée par les effets de bord.

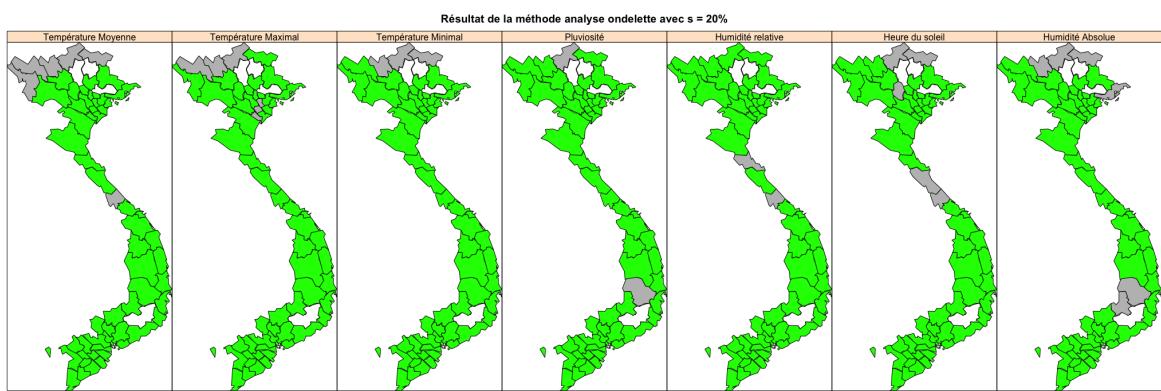


FIGURE 4.6 – Résultat de la méthode d’analyse ondelette sur 64 provinces du Vietnam avec  $s = 20\%$

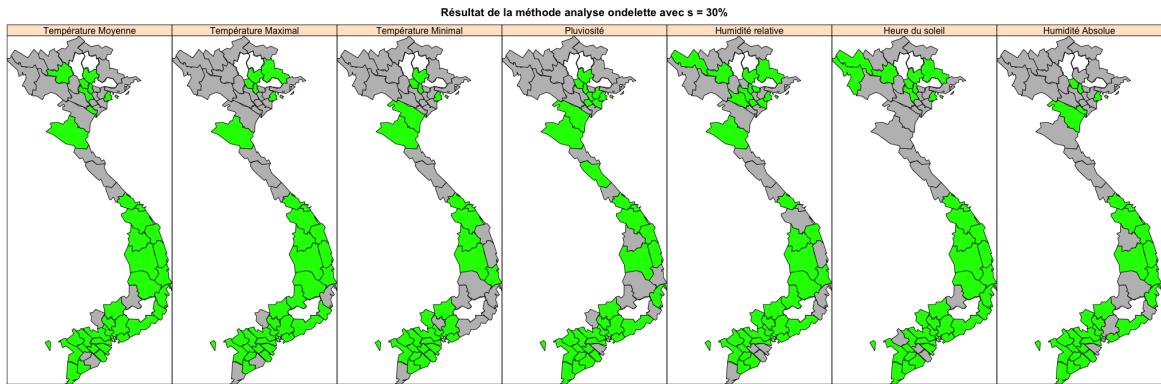


FIGURE 4.7 – Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec  $s = 30\%$

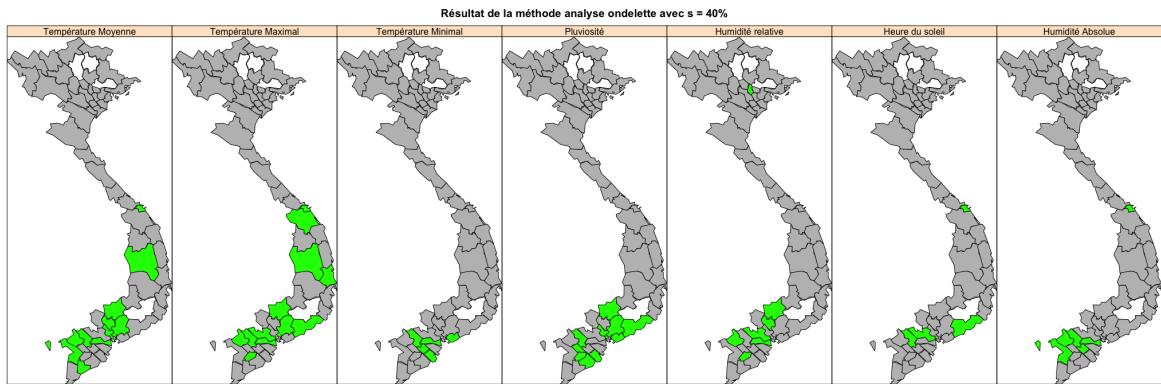


FIGURE 4.8 – Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec  $s = 40\%$

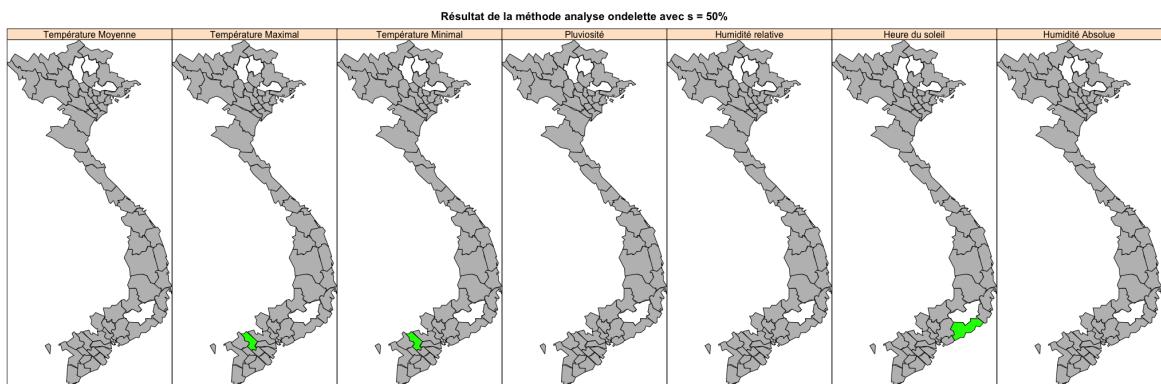


FIGURE 4.9 – Résultat de la méthode d'analyse ondelette sur 64 provinces du Vietnam avec  $s = 50\%$

# Chapitre 5

## Analyse approfondie de la relation entre l'épidémie de la dengue et les facteurs environnement

Dans la chapitre 4, la relation directe entre l'incidence de la dengue et les facteurs climatiques a été montrée par la méthode d'analyse ondelettes. Dans ce chapitre, nous introduisons une méthode d'analyse pour examiner les anciens résultats. Ensuite, nous effectuons des analyses multivariables sur nos données par ce méthode pour montrer les effets latentes entre les facteurs environnementaux sur l'influence de l'épidémie. La méthode GCA consiste en un test statistique permettant de déterminer les relations entre les séries temporelles en mesurant la capacité de prédire les valeurs futures d'une série temporelle à l'aide des valeurs antérieures d'une autre série temporelle. Les résultats du test univariable montre que la température et l'absolue humidité est fortement influence l'incidence de la dengue dans les provinces du Vietnam. La pluviosité, l'humidité relative et l'heure du soleil ont des faible influence sur certains provinces du Vietnam. Cependant, dans les multivariable, l'humidité relative, la pluviosité et l'heure de soleil a montré des latent corrélations lorsque leurs combinaisons avait eu les nombre de provinces significatif plus grandes que ceux du test univariable. De plus, nous avons lancé les test de sous-séquence en appliquant la méthode GCA univariable et multivariable avec les différents périodes. Leurs résultats montre que les effets des facteurs environnementaux sur l'incidence de la dengue a augmenté avec le temps durant le période 1998 - 2010??.

### 5.1 Méthodologie

#### 5.1.1 Le modèle autorégressive - AR

Dans les domaines de statistiques et du traitement du signal, le modèle autorégressive (AR) est une représentation d'un type de processus aléatoire. il est utilisé pour décrire certains processus qui varie suivre le temps dans la nature, l'économie, etc. Dans ce modèle de régression, la variable de réponse dans la période de temps précédente est devenue le prédicteur et les erreurs ont nos hypothèses habituelles sur

les erreurs dans un modèle de régression linéaire simple. L'ordre d'une autorégression est le nombre de valeurs immédiatement précédentes dans la série qui sont utilisées pour prédire la valeur à l'heure actuelle.

La notation  $AR(p)$  indique un modèle autorégressif avec l'ordre  $p$ , le modèle  $AR(p)$  est défini comme :

$$X_t = c + \sigma_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t$$

où  $\varphi_1, \dots, \varphi_p$  sont les paramètres du modèle,  $c$  est un nombre constant, et  $\varepsilon_t$  est le bruit blanc.

Certaines contraintes de paramètres sont nécessaires pour que le modèle reste stationnaire. Par exemple, le processus du modèle  $AR(1)$  avec  $|\varphi_1| \geq 1$  n'est pas stationnaire. Le processus AR le plus simple est  $AR(0)$  qui n'a aucune dépendance entre les termes. Seul le terme erreur/ bruit contribue à la sortie du processus.

Pour un processus  $AR(1)$  avec un  $\varphi$  positif, seul le terme précédent dans le processus et le terme de bruit contribuent à la sortie. Si  $\varphi$  est proche de 0, alors le processus ressemble toujours au bruit blanc, mais lorsque  $\varphi$  s'approche de 1, la sortie obtient une grande contribution du terme précédent par rapport au bruit. Cela entraîne un lissage ou une intégration de la sortie, similaire à un filtre passe-bas.

Pour un processus  $AR(2)$ , les deux termes précédents et le terme de bruit contribuent à la sortie. Si  $\varphi_1$  et  $\varphi_2$  sont tous deux positifs, la sortie ressemblera à un filtre passe-bas, la partie haute fréquence du bruit diminuant. Si  $\varphi_1$  est positif alors que  $\varphi_2$  est négatif, le processus favorise les changements de signe entre les termes du processus. La sortie oscille. Cela peut être assimilé à une détection de bord ou à une détection de changement de direction.

La performance prédictive du modèle autorégressif peut être évaluée dès que l'estimation a été effectuée si la validation croisée est utilisée. Alternativement, après un certain temps donc des paramètres sont estimés, davantage de données seront disponibles et les performances prédictives pourront être évaluées à l'aide des nouvelles données.

### 5.1.2 Le modèle autoregressive vectorielle - VAR

L'autorégression vectorielle (VAR) est un modèle de processus stochastique utilisé pour capturer les interdépendances linéaires entre plusieurs séries temporelles. Les modèles VAR généralisent le modèle autorégressif univarié (modèle AR) en permettant plusieurs variables évolutives. Toutes les variables d'un VAR entrent dans le modèle de la même manière : chaque variable a une équation expliquant son évolution basée sur ses propres valeurs décalées, les valeurs décalées des autres variables du modèle et un terme d'erreur. La modélisation VAR ne nécessite pas autant de connaissances sur les forces influençant une variable que les modèles structurels à équations simultanées : la seule connaissance préalable requise est une liste de variables qui peuvent être supposées s'affecter intertemporellement.

Le modèle VAR décrit l'évolution d'un ensemble de  $k$  variables sur la même période d'échantillonnage ( $t = 1, \dots, T$ ) en tant que fonction linéaire de leur seules valeurs passées. Les variables sont collectées dans des vecteurs  $y_t$  où  $y_{i,t}$  est l'élément  $i^{\text{ème}}$  au

temps observé  $t$ . Un modèle VAR d'ordre  $p$ , noté VAR(p) est défini par :

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t$$

où les arrière observation des  $i$ -période  $y_{t-i}$  s'appelle le  $i^{ème}$  retard de  $y$ ,  $c$  est un vecteur  $k \times 1$  de constance,  $A_i$  est une matrice  $k \times k$  invariance dans le temps et  $e_t$  est un vecteur  $k \times 1$  de terme d'erreur.

Le modèle VAR combine l'avantage du modèle autoregressif (AR) et des équations simultanées (SE) en utilisant la méthode minimisé les résidus et l'estimation de plusieurs variables dans un même système. De même, il surmonte l'inconvénient des SE qu'il ne se soucie pas de l'endogénéité des variable économiques. Les variables macroéconomiques sont souvent endogènes lorsqu'elles interagissent les unes avec les autres. Cet attribut rend la méthode de régression multiple utilisant une équation de régression multiple lors de la déviation de l'estimation. Ce sont les raisons fondamentales qui rendent le modèle VAR si populaires dans la recherche macroéconomique.

### 5.1.3 La méthode causalité de Granger

Le test de causalité de Granger (CA) est un test d'hypothèse statistique permettant de déterminer si une série temporelle est utile en prévoir une autre série temporelle, proposé pour la première fois en 1969 [30]. Cette méthode est reconnu comme le principal progrès sur le problème de causalité. L'idée principal de la méthode CA est :  $Y_t$  cause  $X_t$  si l'on est mieux capable de prédire  $X_t$  à partir de l'ensemble des informations disponibles qu'à partir de ce même ensemble privé de  $Y_t$  [30, 31]. Si cette condition est satisfaite, cela signifie que  $Y_t$  contient des informations exclusive qui ne sont pas présentes dans toutes les autres séries temporelles et que cette information est utile pour prédire la série temporelle  $X_t$  d'intérêt.

La causalité de Granger est basée sur deux principes :

- La cause se produit avant son effet ;
- La cause produit des changements uniques dans l'effet. En d'autres termes, la série temporelle causale contient des informations uniques sur la série temporelle d'effet qui ne sont pas disponibles autrement.

Soit  $I(t)$  l'ensemble des informations contenant toutes les informations pertinentes dans l'univers disponibles jusqu'à l'instant  $t$  et  $I_{-X}(t)$  est l'ensemble des informations de l'univers à l'exclusion de  $X$  jusqu'au moment  $t$ . Donnant deux séries temporelles  $X$  et  $Y$ , Granger a proposé de tester l'hypothèse suivante pour l'identification d'un effet causal de  $X$  sur  $Y$  :

$$\mathbb{P}[Y(t+1) \in A | I(t)] \neq \mathbb{P}[Y(t+1) \in A | I_{-X}(t)]$$

où  $\mathbb{P}$  dénote une probabilité et  $A$  est un ensemble arbitraire non vide. Si nous pouvons prouver l'hypothèse ci-dessus ou si nous pouvons rejeter l'hypothèse nulle :

$$\mathbb{P}[Y(t+1) \in A | I(t)] = \mathbb{P}[Y(t+1) \in A | I_{-X}(t)]$$

nous disons que  $X$  Granger-cause  $Y$ .

Plus précisément, considérons deux séries chronologiques stationnaires  $X_1$  et  $X_2$  avec un modèle autorégressif linéaire :

$$\begin{aligned} X_1(t) &= \sum_{j=1}^p A_{11j}X_1(t-j) + \sum_{j=1}^p A_{12j}X_2(t-j) + E_1(t) \\ X_2(t) &= \sum_{j=1}^p A_{21j}X_1(t-j) + \sum_{j=1}^p A_{22j}X_2(t-j) + E_2(t) \end{aligned} \quad (5.1)$$

où  $p$  est le nombre maximum d'observations retardées incluses dans le modèle,  $A$  est la matrice contenant les coefficients du modèle (par exemple les contributions de chaque observation retardée) des valeurs prédictives de  $X_1(t)$  et  $X_2(t)$ .  $E_1$  et  $E_2$  sont des erreurs de prédictions pour chaque série temporelle. Si l'écart de  $E_1$  (ou  $E_2$ ) est réduit par l'inclusion de  $X_2$  (ou  $X_1$ ), alors il est dit que  $X_2$  (ou  $X_1$ ) Granger-cause  $X_1$  (ou  $X_2$ ). En d'autres termes,  $X_2$  Granger-cause  $X_1$  si le coefficient  $A_{12}$  est significativement différent de zéro. Cela peut être testé en effectuant un F-test de l'hypothèse nulle que  $A_{12} = 0$ , étant donné l'hypothèse de la stationnarité de la covariance sur  $X_1(t)$  et  $X_2(t)$ . L'ampleur d'une interaction Granger-causalité peut être estimée par le logarithme correspondante du F-statistique [32].

Le cadre de causalité de Granger peut être étendu pour résoudre le problème multivariable : étant donné les séries de temps  $p$ ,  $X_1, X_2, \dots, X_{p-1}, X_p$ , nous sommes intéressés à identifier quelle série temporelle Granger cause  $X_i$ . Il y a deux façons de résoudre ce problème :

- Examiner les relations par paires dans chaque paire de séries temporelles ;
- Effectuer une autorégression vectorielle (VAR modèle) comme :

$$X_i(t) = \sum_{j=1}^p A_{i,j}^T X_j^{t,lagged} + E \quad (5.2)$$

où  $X_j^{t,lagged} = [X_j(t-L), \dots, X_j(t-1)]$  est la série temporelle retardée et  $A_{i,j}$  est le vecteur de coefficient. Ensuite, nous testons si  $A_{i,j}$  sont significativement différents de zéro. Une série temporelle  $X_i$  est appelée cause Granger d'une autre série temporelle  $X_j$  si au moins un des éléments  $A_L(i, j)$  pour  $L = 1, \dots, p$  est significativement plus grand que zéro.

Nous avons appliqué la causalité de Granger pour analyser la relation entre l'incidence de la DHF et les variables climatiques dans chacune des 64 provinces du Vietnam. Le test de causalité de Granger a été effectué en utilisant le package R **vars** [33]. Chaque provinces a été associée à la station climatique la plus proche de son centre de population comme la façon de celui de la méthode d'analyse ondelette au chapitre 4. Dans le premier test, nous avons effectué un test de causalité de Granger pour déterminer l'influence indépendance des facteurs climatiques sur l'incidence de la dengue dans chaque provinces au Vietnam. Puis, nous avons testé l'influence d'ensemble des facteur climatiques sur l'incidence de la dengue sur des provinces au Vietnam. Enfin, nous avons divisé la période en sub-séquence et nous testons la méthode GCA sur ces sub-séquence pour trouver la tendance d'influence des facteurs climatiques avec l'incidence de la dengue.

## 5.2 Résultats

### 5.2.1 Résultats d'analyse univariable

Nous avons d'abord analysé la relation entre chaque facteur climatique et l'incidence de la dengue pour chaque province du Vietnam. Pour chaque province, nous avons appliqué la méthode GC et testé l'hypothèse nulle que le facteur climatique n'est pas la cause d'épidémiologie de la dengue. La figure 5.1 montre le résultat expérimentale de la méthode GCA sur notre données. Le vert indique les provinces pour lesquelles le test est significatif au niveau de  $\alpha = 0,05$ , le gris indique les provinces pour lesquelles le test n'est pas significatif et le blanc indique les provinces pour lesquelles le test n'a pu être effectué faute de données. On a calculé le nombre total des provinces qui avaient la relation entre DHF et chacun des facteurs climatiques et leurs représente dans le tableau 5.1.

Climatic factor	ta	tx	tm	rf	rh	sh	ah
Number of provinces	33	34	34	13	12	10	39
Percentage	51.56	53.13	53.13	20.31	18.75	15.63	60.94

TABLE 5.1 – Le nombre des provinces qui ont la relation entre le DHF et chacun des facteurs climatiques.

Les résultats de la figure 5.1 et le tableau 5.1 suggèrent que :

- la température et l'humidité absolue fortement influence l'incidence de la dengue, principalement dans les régions du centre et du sud du pays ;
- la pluviosité et l'humidité relative ont un impact faible sur l'incidence de la dengue dans certaines provinces ;
- les heures d'ensoleillement ont un impact important sur l'influence de la zone centrale.

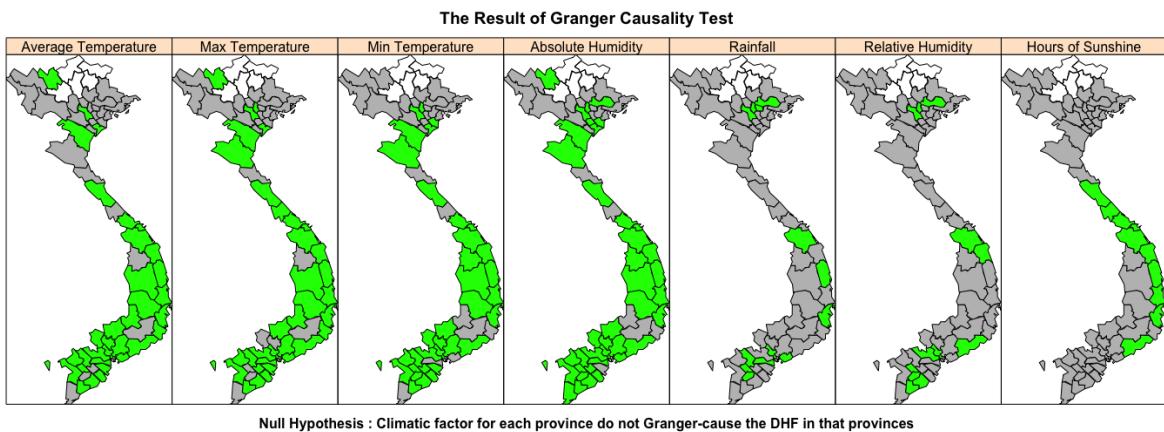


FIGURE 5.1 – Résultats du test de causalité Granger pour chaque variable climatique et chaque province.

### 5.2.2 Résultats d'analyse multivariable

Dans cette étude, nous utilisons le GC multivariable pour analyser la relation entre DHF et l'impact de l'ensemble des facteurs climatiques au même moment. Nous avons combiné des paires de variable pour appliquer la méthode GCA multivariable. Cependant, parmi les trois variable de température : température moyenne (TA), température maximal (TX), température minimal (TM), nous choisissons uniquement la température moyenne pour combiner avec les autres variables. Pour chaque cas de test, nous avons compté le nombre des provinces qui avaient la relation entre DHF et chacun des paires des facteurs climatiques et leurs représentante dans le tableau 5.2.

Set of climatic factors	ta + rf	ta + rh	ta + sh	ta + ah	rf + rh	rf + sh	rf + ah	rh + sh	rh + ah	sh + ah
Number of provinces	33	31	29	35	13	20	34	18	33	33
Percentage	51.56	48.44	45.31	54.69	20.31	31.25	53.13	28.13	51.56	51.56

TABLE 5.2 – Le nombre de provinces a été conclu qui était la relation entre DHF et l'ensemble des paires des facteur climatique

Nous pouvons voir clairement que les variables température et humidité absolue ont une grande influence sur la situation de DHF, que ce soit avec une analyse indépendant ou multivariée. Ce qui nous a surpris, c'est la combinaison entre les heures d'ensoleillement et les précipitations, les heures d'ensoleillement et l'humidité relative nous donnent un meilleur résultat qu'une analyse indépendante. Cela peut certainement arriver parce que les heures d'ensoleillement et les précipitations sont négativement corrélées.

### 5.2.3 Résultats d'analyse sous-séquence

Dans cette analyse, nous effectuons une analyse sous-séquence avec la méthode GCA dans les deux cas univariable et multivariable pour analyser le développement des corrélation d'épidémie et les facteurs climatiques sur des périodes de temps. À partir de janvier 1998, nous avons effectué l'analyse sur une période de 6, 7, 8 et 9 ans au lieu de 12 ans. Ensuite, nous répétons le processus avec le temps de début de 3 mois plus tard que le test précédent. La figure 5.2 montre que les résultats de l'analyse de sous-séquence avec la longueur de la période sont égaux à 6 ans. Chaque point de l'axe abscisse représente une période d'analyse de six ans et l'axe ordonnée indique le nombre de villes dont on a conclu qu'elles étaient liées à la dengue et aux facteurs climatiques correspondants. Les figures 5.3, 5.4, 5.5 montrent le résultat de la sous-séquence univariable avec la longueur de la période égale à 7, 8 et 9 ans. Nous pouvons voir que l'influence des facteurs climatiques sur la DHF a tendance à augmenter avec le temps. Cela se reflète dans la valeur de croissance sur l'axe ordonné de la plupart des variables météorologiques, où l'augmentation de la pluviométrie variable et de l'humidité absolue est la plus évidente.

Les figures 5.6, 5.7, 5.8, 5.9 représentent le résultat des tests de sous-séquence multivariable dans ces périodes. Nous constatons que la température et l'humidité absolue ont toujours des effets importants sur l'incident de la dengue lorsque leur

combinaison avec d'autres variables augmente fortement avec le temps. De plus, la combinaison des précipitations et des heures d'ensoleillement, des précipitations et de l'humidité relative montre également leur effet latent lorsqu'elles sont combinées.

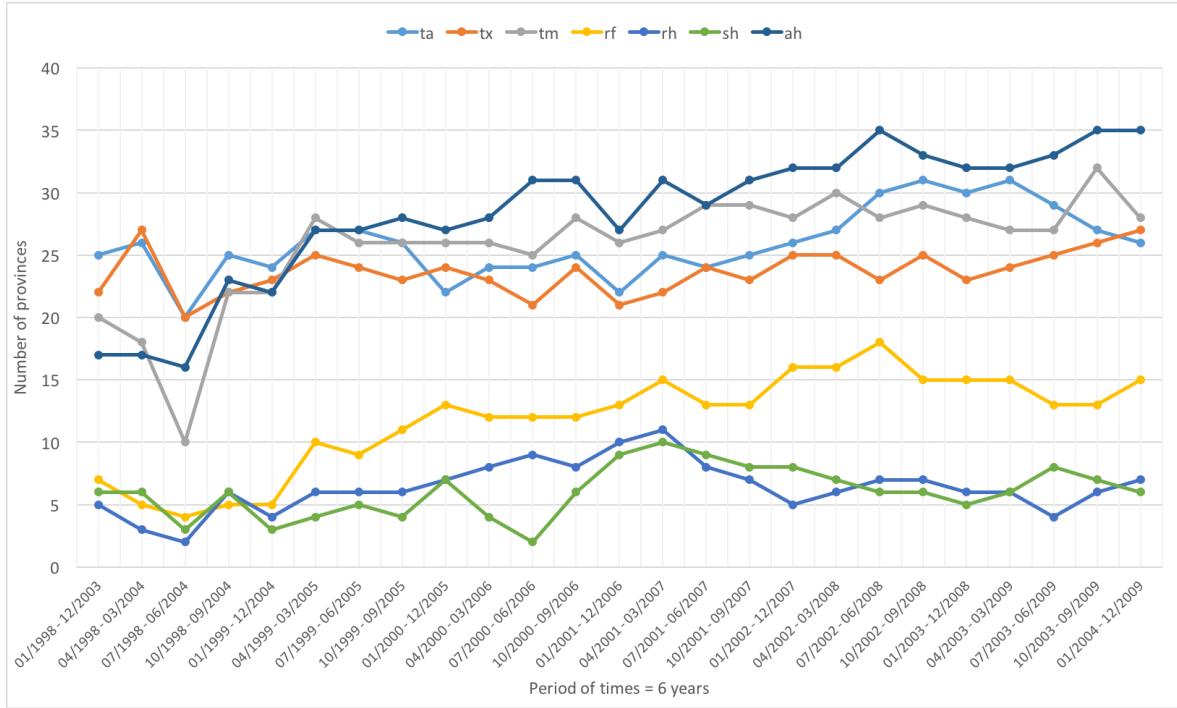


FIGURE 5.2 – Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période L = 6 ans

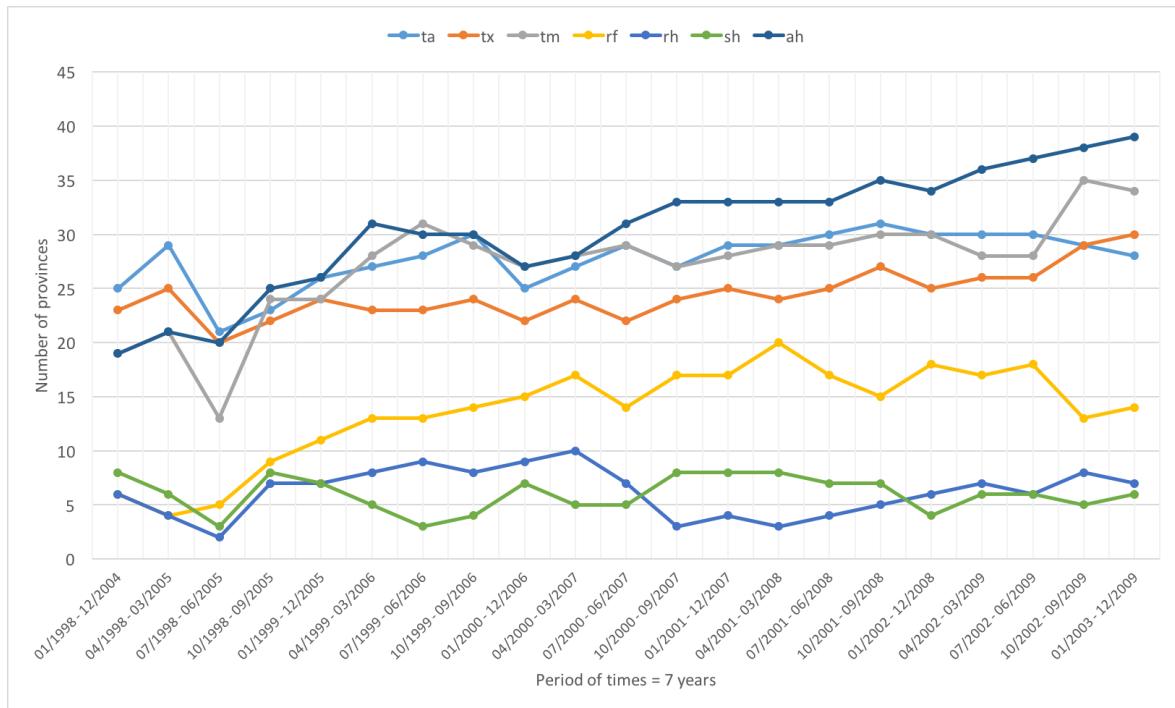


FIGURE 5.3 – Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période  $L = 7$  ans

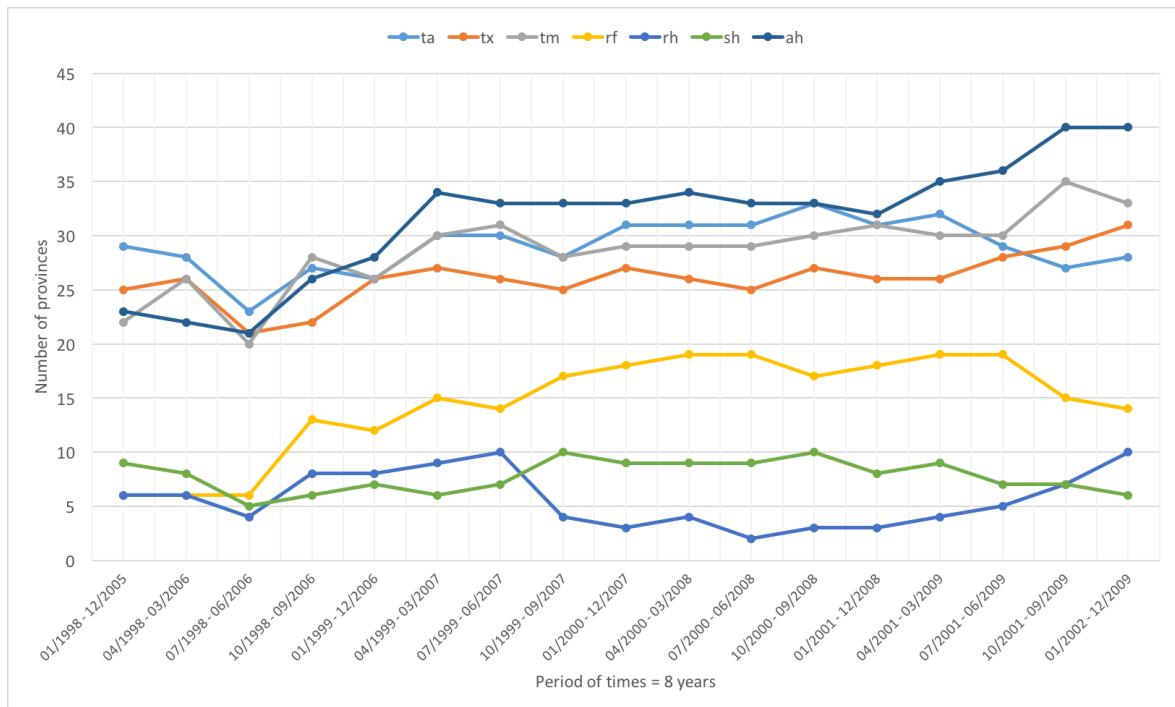


FIGURE 5.4 – Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période  $L = 8$  ans

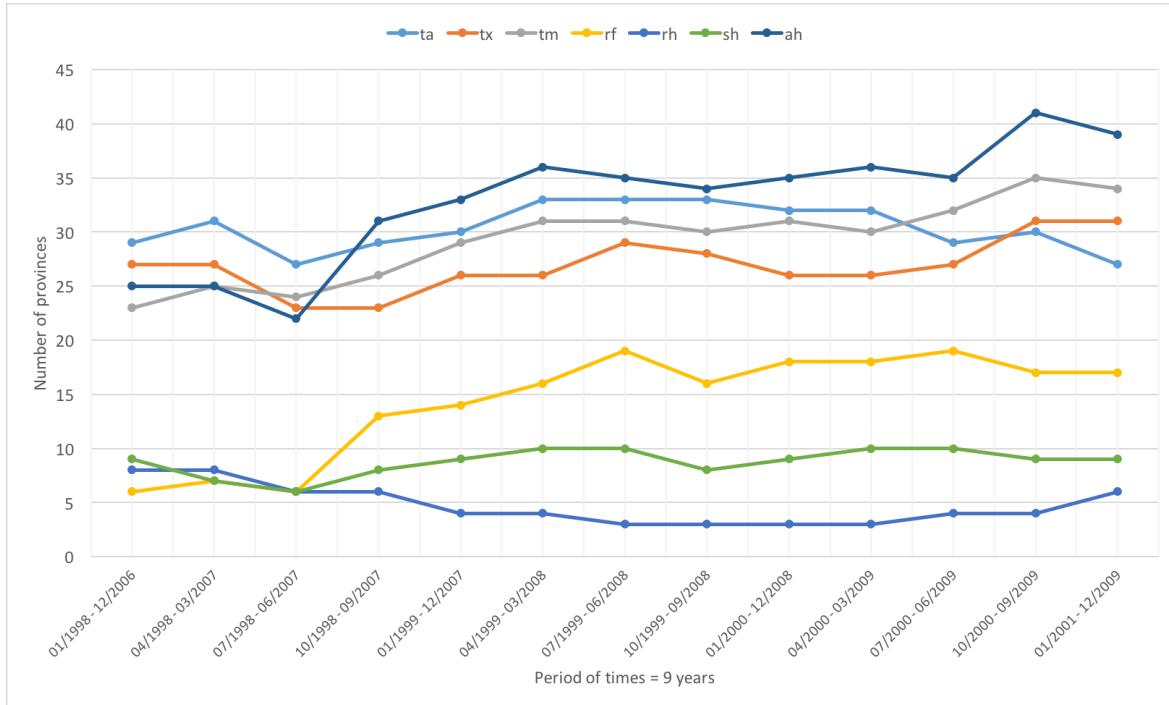


FIGURE 5.5 – Le résultat du test sous-séquence avec la méthode GCA univariée. La durée de chaque période L = 8 ans

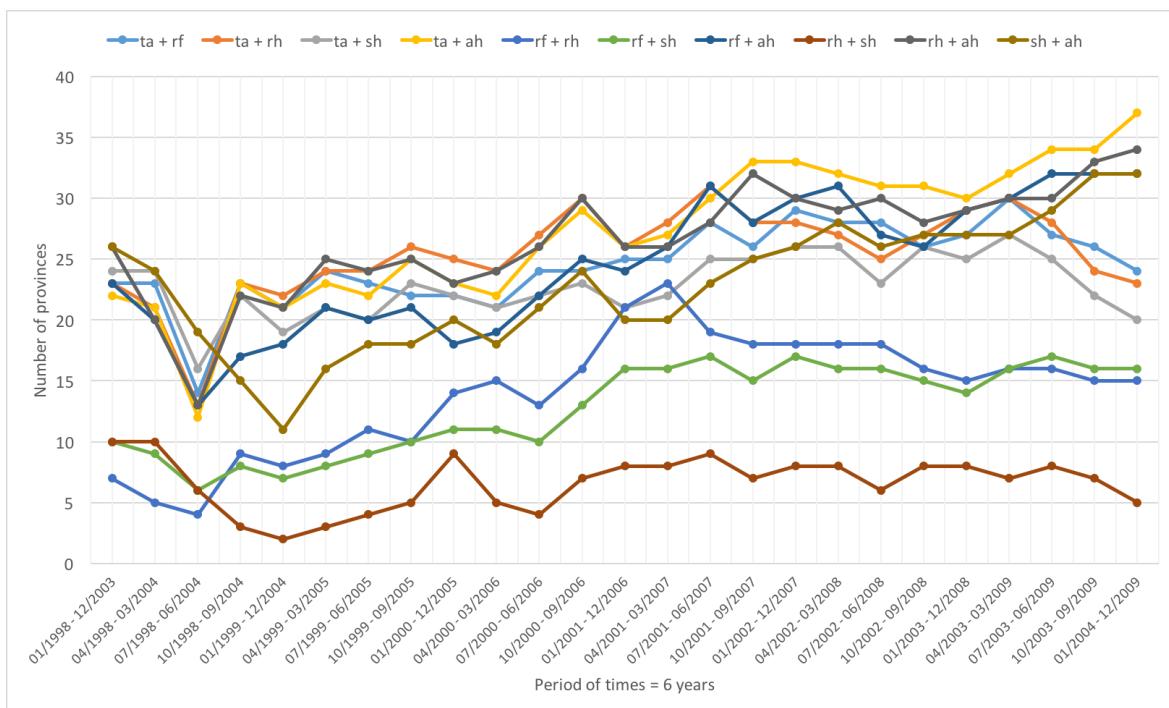


FIGURE 5.6 – Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période L = 6 ans

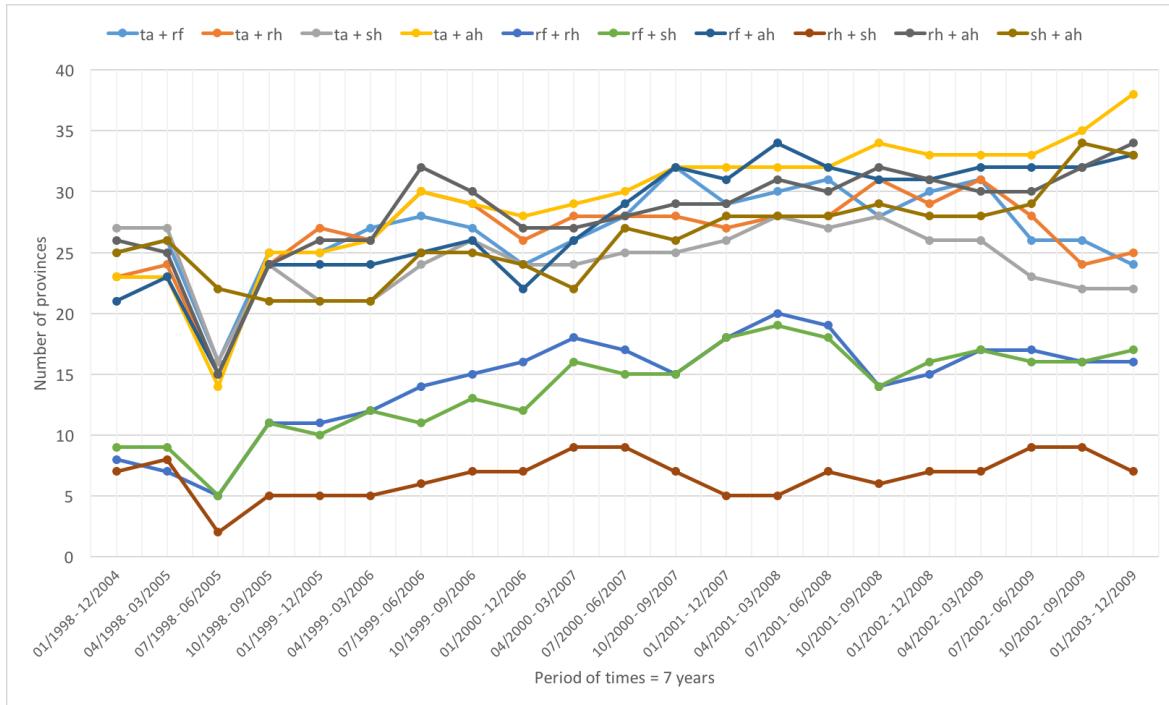


FIGURE 5.7 – Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période  $L = 7$  ans

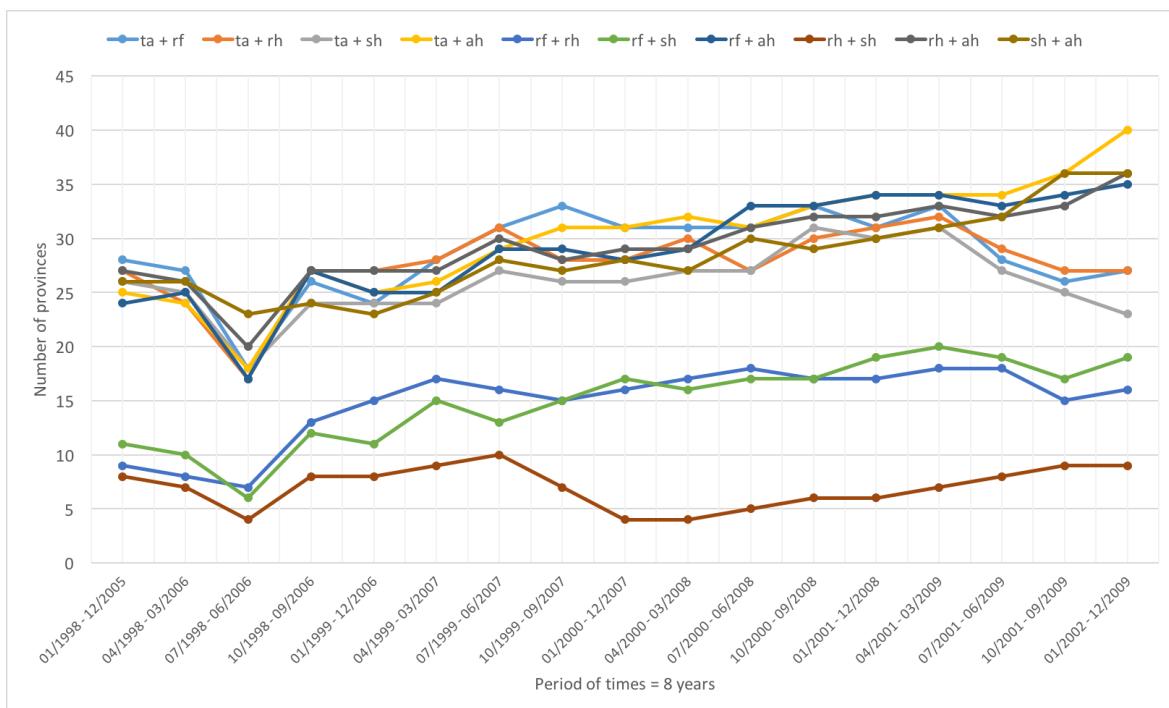


FIGURE 5.8 – Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période  $L = 8$  ans

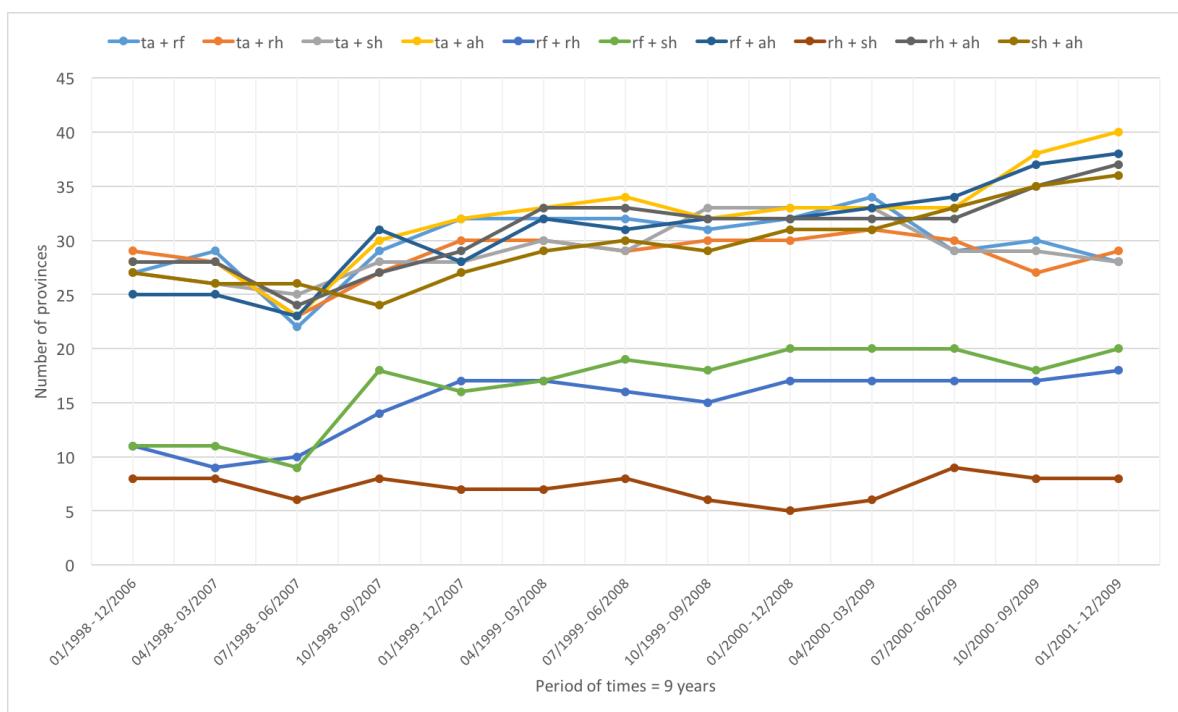


FIGURE 5.9 – Le résultat du test sous-séquence avec la méthode GCA multivariée. La durée de chaque période  $L = 9$  ans

# Chapitre 6

## Conclusion

### 6.1 Résultats de la thèse

Nous avons présenté dans cette thèse nos contribution pour le problème de détection les relations entre les facteurs climatiques et la maladie infectieuse en utilisant les données réels des provinces des pays en Asie du Sud-Est.

La structure et les caractéristiques des données sont été présenté dans le chapitre 2. Afin de faciliter le traitement des analyse, nous avons effectué des étape de pré-traitement sur les données comme l'interpolation, la normalisation et le retranchement sur les séries temporelles. Dans le chapitre 3, nous avons appliqué la méthode k-means pour classifier les taux d'inflections de 273 provinces au 9 pays du région Sud-Est d'Asie. Les taux d'infection des provinces sont classifiés dans les régions qui ont des situation géographique proche les uns des autres. Ces régions peuvent être affectés par des même conditions environnementales, ce qui peut conduire à une similarité des taux d'infection.

Dans les chapitre 4, 5 et 6, nous avons utilisés les différentes méthodes basé sur les différentes contexte pour effectuer l'analyse sur la relation entre le taux d'infection et les facteurs environnementaux du 64 provinces du Vietnam à partir Janvier 1998 jusqu'au Septembre 2010. Les résultats de ces méthodes ont montré des même conclusion :

- La température et l'humidité absolue sont fortement affecté sur l'influence de la dengue dans plupart des provinces du Vietnam.
- La pluviosité, l'humidité relative et l'heure d'ensoleillement ont un impact faible sur l'influence de la dengue dans certaines provinces du Vietnam.
- Dans la plupart des provinces, la dengue est plus répandue pendant la saison des pluies, mais à l'échelle interannuelle, les épidémies de dengue ont également été associées à la sécheresse.

Cependant, Il existe des différentes avantages des algorithmes :

- La méthode analyse ondelette dans le chapitre 4 permet de trouver les corrélation positive et négative dans les cycles spécifique de la maladie.
- La méthode GCA univariable permet de trouver les relation directement entre les facteurs climatiques et l'incidence de la dengue. De plus, les résultats du GCA multivariable montre des implications latentes entre des facteurs

climatiques dans la mesure où ils contribuent ensemble à l'épidémie de la dengue.

- Le modèle EDM a montré une relation non-linéaire entre l'humidité absolue et la température. Cet effet est masqué dans la région tempérée par leur forte corrélation entre eux.

## 6.2 Perspectives

Les relations entre l'incidence de la dengue et les facteurs environnementaux du 64 provinces au Vietnam a été analysée très détails en utilisant des différentes méthodes. Cependant, nous devons approfondir l'analyse des données des provinces d'autres pays dans la région Asie du Sud-Est pour obtenir un aperçu plus complet et détaillé de l'influence des facteurs environnementaux sur l'évolution de l'épidémie de la dengue.

Les connaissances sur l'influence des facteurs environnementaux et l'épidémie nous permettent de construire un système d'alerte visant à renforcer les mesures de prévention de l'épidémie dans les périodes où l'éclosion de l'épidémie est venue (Par exemple : au début la saison de la pluie où la température et l'absolue humidité fortement augmentent).

De plus, nous pouvons développer un logiciel en utilisant la méthode d'analyse dans cette thèse pour aider les utilisateurs ordinaires à décomposer les relations entre des séries temporelles.

# Bibliographie

- [1] SNOW, John. On the mode of communication of cholera. John Churchill, 1855.
- [2] KERMACK, M. ; MCKENDRICK, A. Contributions to the mathematical theory of epidemics. Part I. Proc. r. soc. a, 1927, 115.5 : 700-721.
- [3] KERMACK, William Ogilvy ; MCKENDRICK, Anderson G. Contributions to the mathematical theory of epidemics. II.—The problem of endemicity. Proc. R. Soc. Lond. A, 1932, 138.834 : 55-83.
- [4] KERMACK, William Ogilvy ; MCKENDRICK, Anderson G. Contributions to the mathematical theory of epidemics. III.—Further studies of the problem of endemicity. Proc. R. Soc. Lond. A, 1933, 141.843 : 94-122.
- [5] DOLL, Richard ; HILL, Austin Bradford. Mortality in relation to smoking : ten years' observations of British doctors. British medical journal, 1964, 1.5395 : 1399.
- [6] BHATT, Samir, et al. The global distribution and burden of dengue. Nature, 2013, 496.7446 : 504.
- [7] ZHANG, Ying ; BI, Peng ; HILLER, Janet E. Climate change and the transmission of vector-borne diseases : a review. Asia Pacific Journal of Public Health, 2008, 20.1 : 64-76.
- [8] CUMMINGS, Derek AT, et al. The impact of the demographic transition on dengue in Thailand : insights from a statistical analysis and mathematical modeling. PLoS medicine, 2009, 6.9 : e1000139.
- [9] BARRETO, Florisneide R., et al. Spread pattern of the first dengue epidemic in the city of Salvador, Brazil. BMC Public Health, 2008, 8.1 : 51.
- [10] HALES, S. ; WEINSTEIN, P. ; WOODWARD, A. Dengue fever epidemics in the South Pacific : driven by El Niño southern oscillation ?. The Lancet, 1996, 348.9042 : 1664-1665.
- [11] HALES, Simon, et al. El Niño and the dynamics of vectorborne disease transmission. Environmental Health Perspectives, 1999, 107.2 : 99.
- [12] CAZELLES, Bernard, et al. Nonstationary influence of El Nino on the synchronous dengue epidemics in Thailand. PLoS medicine, 2005, 2.4 : e106.
- [13] BARRETO, Mauricio L. ; TEIXEIRA, Maria Gloria. Dengue fever : a call for local national and international action. Lancet, 2008, 372.9634 : 205.
- [14] GUBLER, Duane J. Cities spawn epidemic dengue viruses. Nature medicine, 2004, 10.2 : 129.
- [15] CUMMINGS, Derek AT, et al. Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. Nature, 2004, 427.6972 : 344.

- [16] VAN PANHUIS, Willem G., et al. Region-wide synchrony and traveling waves of dengue across eight countries in Southeast Asia. *Proceedings of the National Academy of Sciences*, 2015, 112.42 : 13069-13074.
- [17] SCHMIDT, Wolf-Peter, et al. Population density, water supply, and the risk of dengue fever in Vietnam : cohort study and spatial analysis. *PLoS medicine*, 2011, 8.8 : e1001082.
- [18] THANH TOAN, Do Thi, et al. Hot spot detection and spatio-temporal dispersion of dengue fever in Hanoi, Vietnam. *Global health action*, 2013, 6.1 : 18632.
- [19] DO, Thi Thanh Toan, et al. Climatic-driven seasonality of emerging dengue fever in Hanoi, Vietnam. *BMC public health*, 2014, 14.1 : 1078.
- [20] LE VIET, Thanh, et al. A dengue outbreak on a floating village at Cat Ba Island in Vietnam. *BMC public health*, 2015, 15.1 : 940.
- [21] THAI, Khoa TD, et al. Dengue dynamics in Binh Thuan province, southern Vietnam : periodicity, synchronicity and climate variability. *PLoS neglected tropical diseases*, 2010, 4.7 : e747.
- [22] BATISTA, Gustavo EAPA ; WANG, Xiaoyue ; KEOGH, Eamonn J. A complexity-invariant distance measure for time series. In : *Proceedings of the 2011 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2011. p. 699-710.
- [23] KRUSKAL, Joseph B. ; SANKOFF, David (ed.). *Time Warps, String Edits, and Macromolecules : The Theory and Practice of Sequence Comparison*. Addison-Wesley, 1983.
- [24] KEOGH, Eamonn J. ; PAZZANI, Michael J. Derivative dynamic time warping. In : *Proceedings of the 2001 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2001. p. 1-11.
- [25] XIE, Ying ; WILTGEN, Bryan. Adaptive feature based dynamic time warping. *International Journal of Computer Science and Network Security*, 2010, 10.1 : 264-273.
- [26] MORSE, Michael D. ; PATEL, Jignesh M. An efficient and accurate method for evaluating time series similarity. In : *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007. p. 569-580.
- [27] ROUSSEEUW, Peter J. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 1987, 20 : 53-65.
- [28] GABOR, Dennis. Theory of communication. Part 1 : The analysis of information. *Journal of the Institution of Electrical Engineers-Part III : Radio and Communication Engineering*, 1946, 93.26 : 429-441.
- [29] KAISER, W. A. False-positive results in dynamic MR mammography. Causes, frequency, and methods to avoid. *Magnetic resonance imaging clinics of North America*, 1994, 2.4 : 539-555.
- [30] GRANGER, Clive WJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica : Journal of the Econometric Society*, 1969, 424-438.
- [31] GRANGER, Clive WJ. Testing for causality : a personal viewpoint. *Journal of Economic Dynamics and control*, 1980, 2 : 329-352.

- [32] GEWEKE, John. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 1982, 77.378 : 304-313.
- [33] PFAFF, Bernhard ; STIGLER, Matthieu ; PFAFF, Maintainer Bernhard. Package ‘vars’. Online] <https://cran.r-project.org/web/packages/vars/vars.pdf>, 2018.
- [34] HOANG, Pham Nguyen, et al. Causality analysis between climatic factors and dengue fever using the Granger causality. In : Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2016 IEEE RIVF International Conference on. IEEE, 2016. p. 49-54.

## Annexe A

# Les résultats complémentaires des méthodes utilisent

### A.1 Exploration des résultats d'analyse sous-séquence de la méthode Causalité de Granger uni-variable et multivariable

Plus précisément, nous allons mettre en place les diagrammes montrant la fréquence des provinces importantes pour chaque période. Les figures A.1, A.2, A.3, A.4, A.5, A.6, A.7 illustrent le diagramme de fréquence des provinces importantes par la fréquence avec la longueur de la période est égale à 6. Les provinces ayant une position basse sur l'axe des Y représentent une forte corrélation avec le facteur climatique tout au long de leur période.

Les figures A.8, A.9, A.10, A.11, A.12, A.13, A.14 montrent le diagramme de fréquence des provinces importantes classées par latitude de chaque province avec la durée de la période est égale à 6. Nous pouvons voir que quand une épidémie de dengue sur une province montre que la relation avec le facteur climatique dans une période continue à montrer sa relation dans les périodes plus récentes.

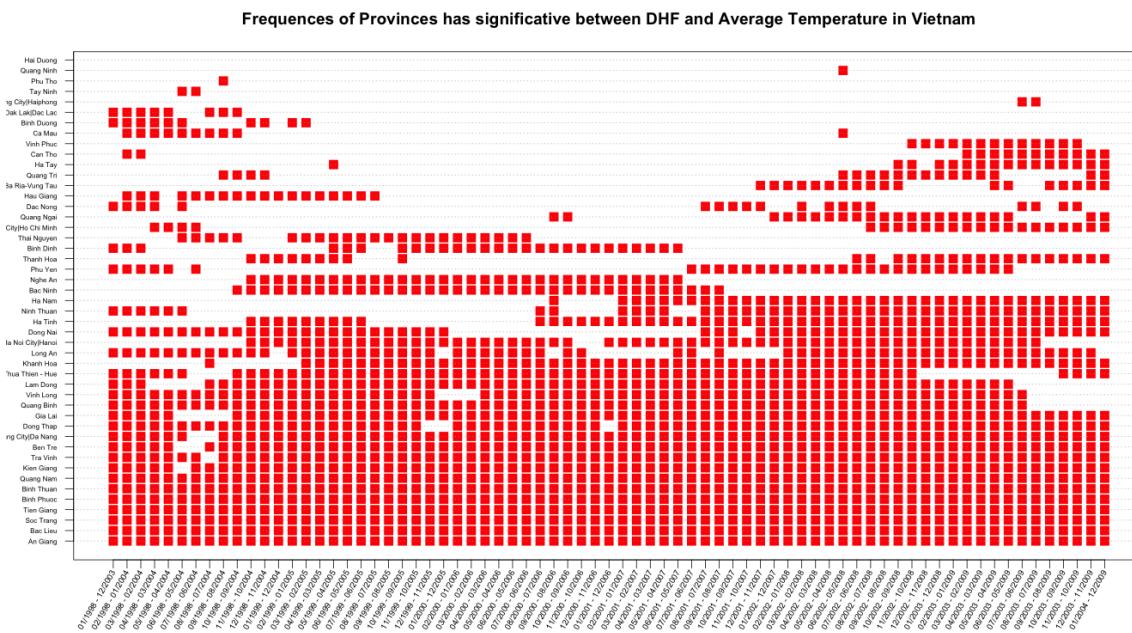


FIGURE A.1 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and average temperature in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

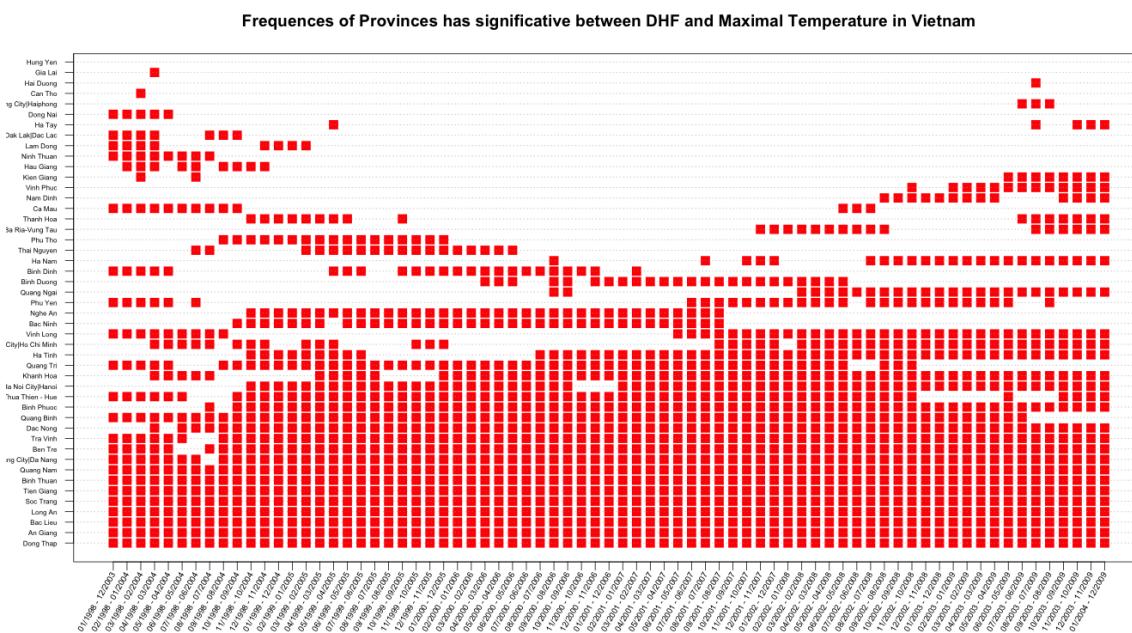


FIGURE A.2 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and maximal temperature in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

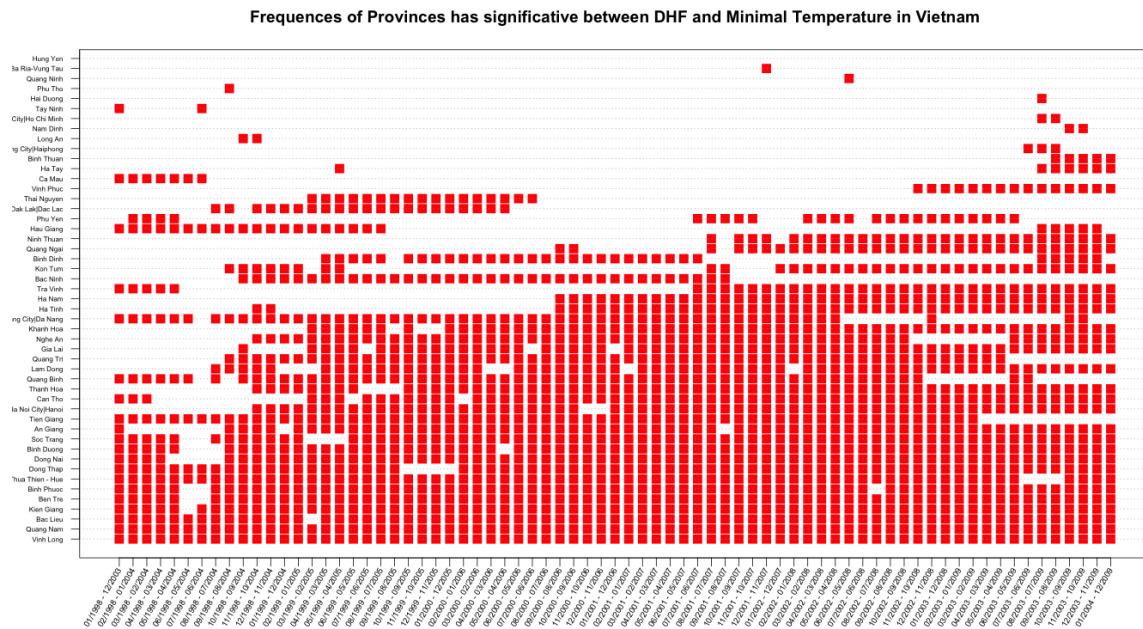


FIGURE A.3 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and minimal temperature in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

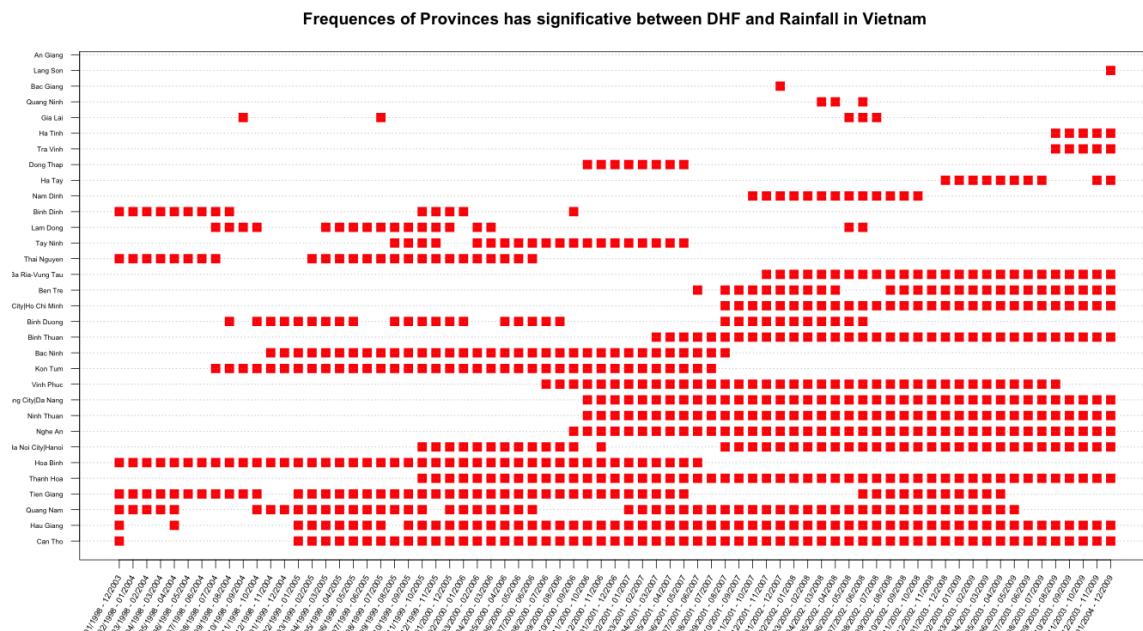


FIGURE A.4 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and rainfall in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

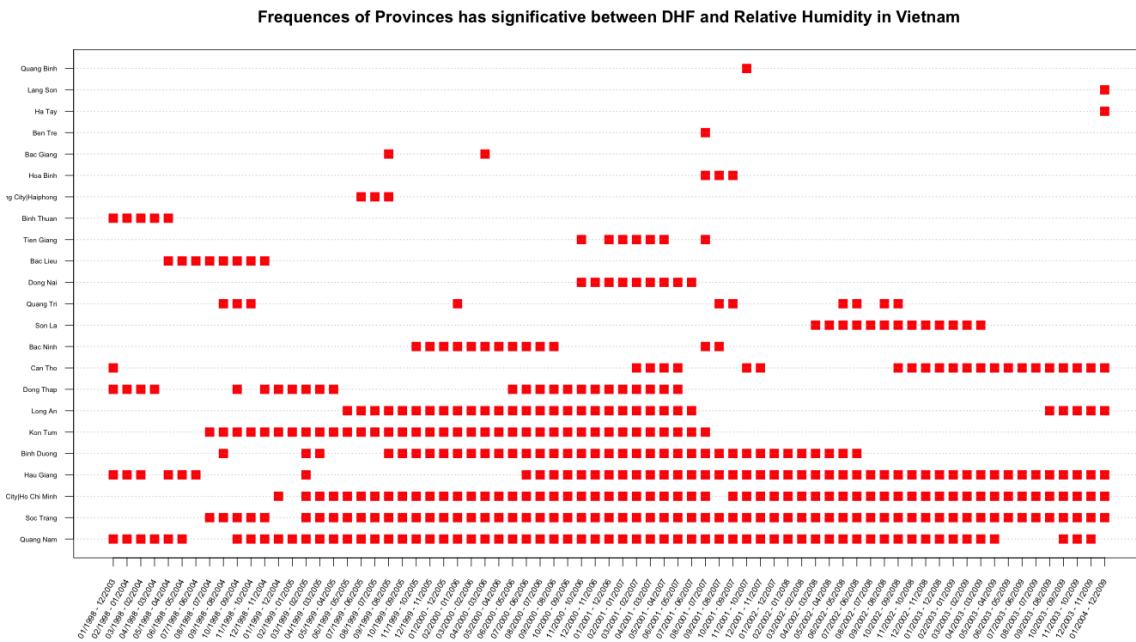


FIGURE A.5 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and relative humidity in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

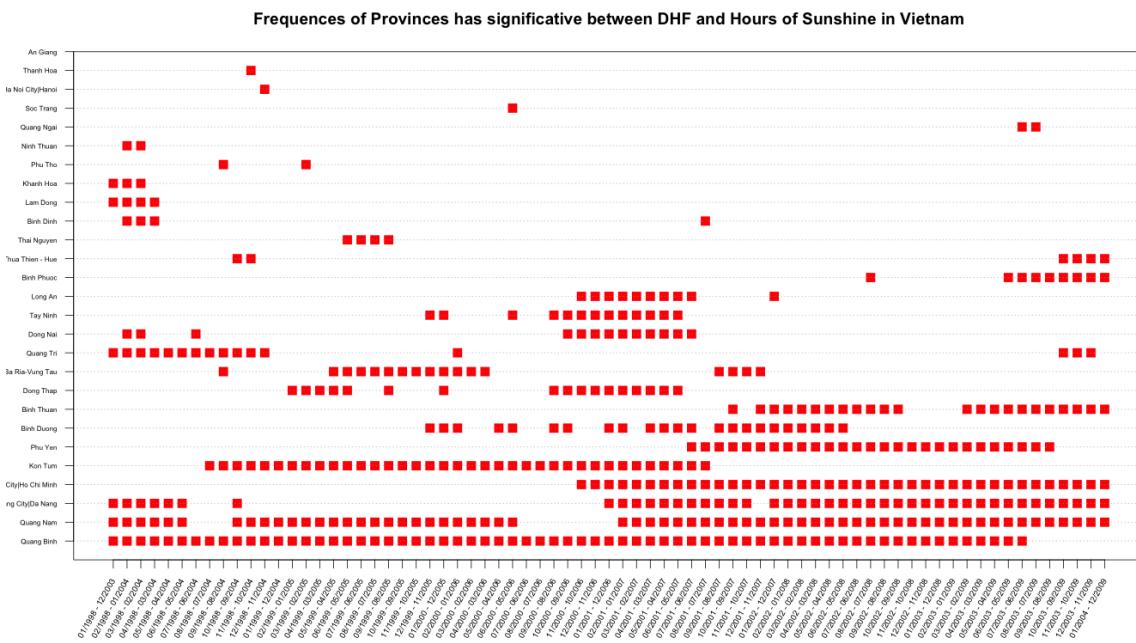


FIGURE A.6 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and hours of sunshine in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

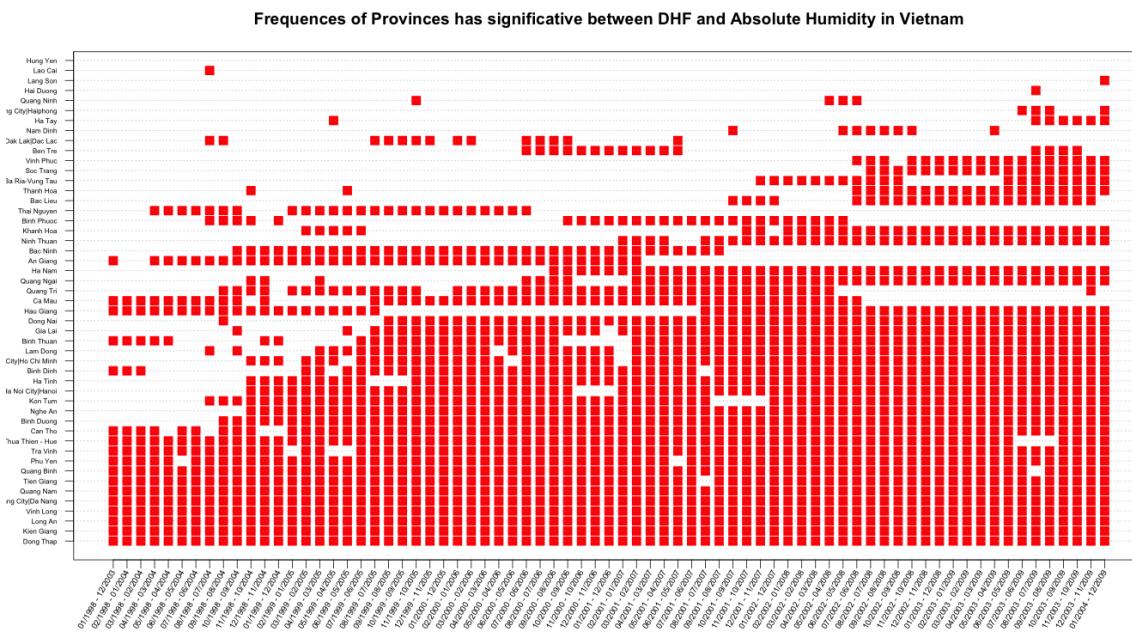


FIGURE A.7 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and absolute humidity in this province. The Y axis is sorted by the total number of significant for every provinces over the period.

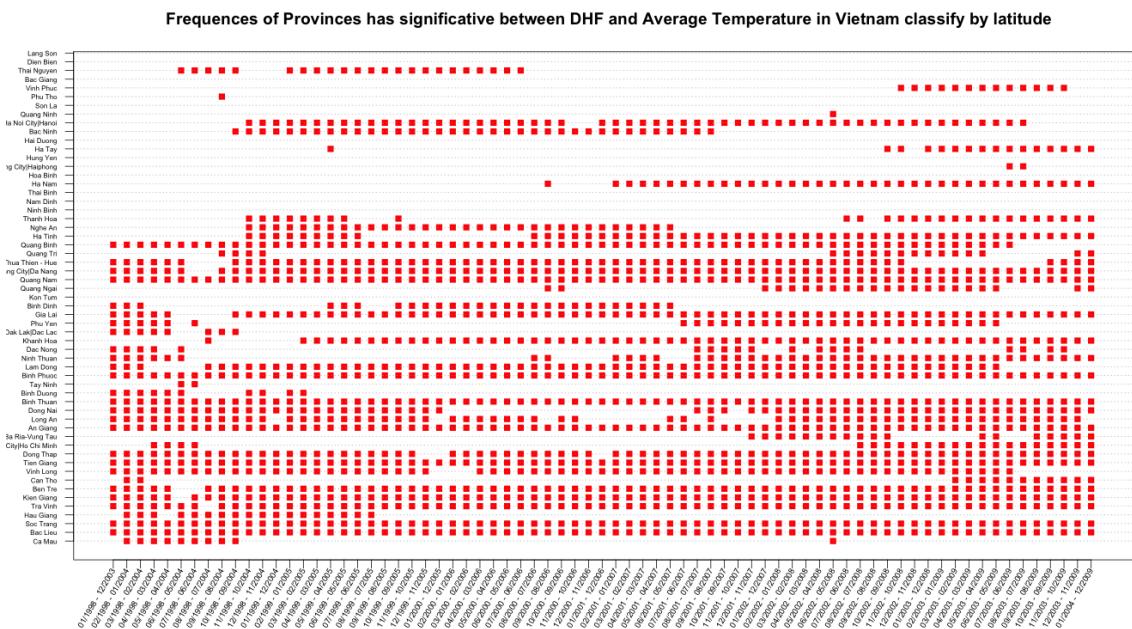


FIGURE A.8 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and average temperature in this province. Provinces on Y axis are ordered according to their latitude.

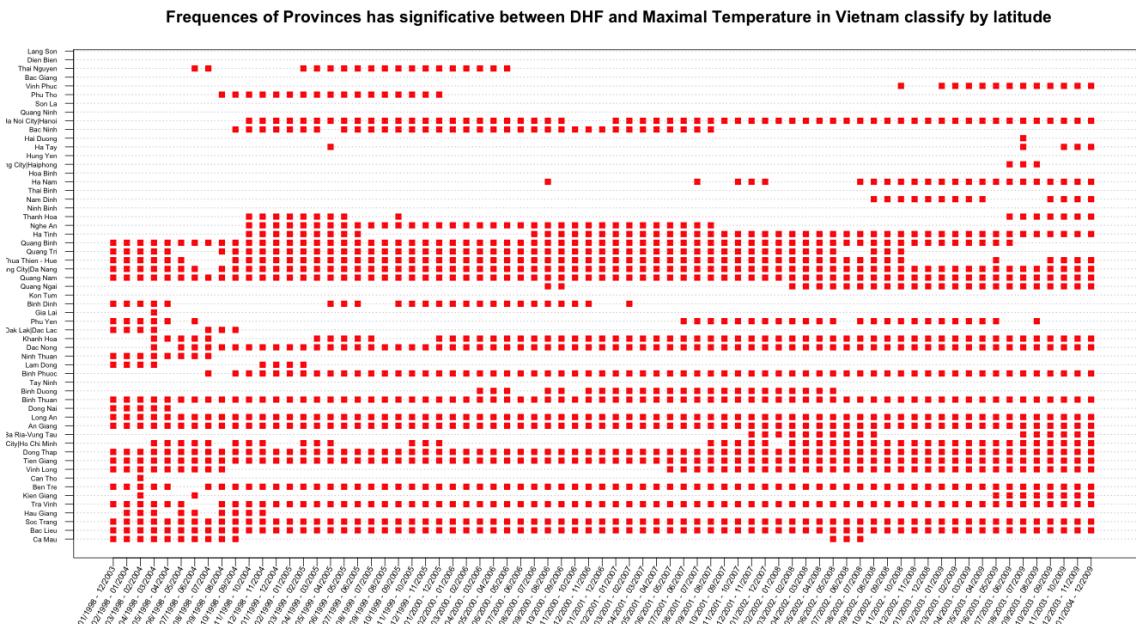


FIGURE A.9 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and maximal temperature in this province. Provinces on Y axis are ordered according to their latitude.

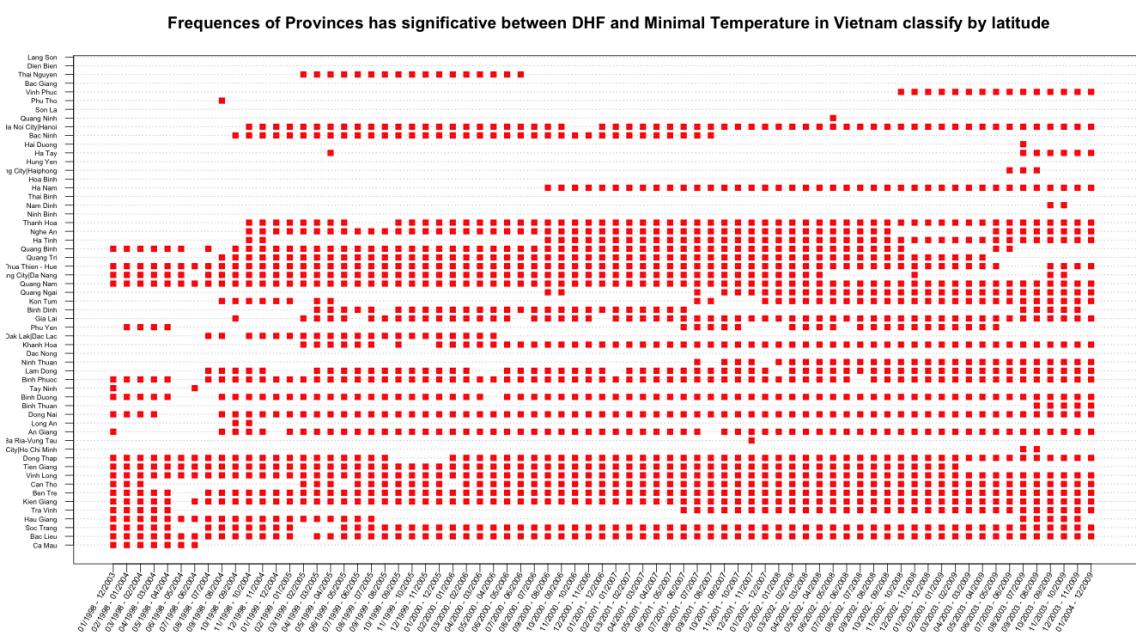


FIGURE A.10 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and minimal temperature in this province. Provinces on Y axis are ordered according to their latitude.

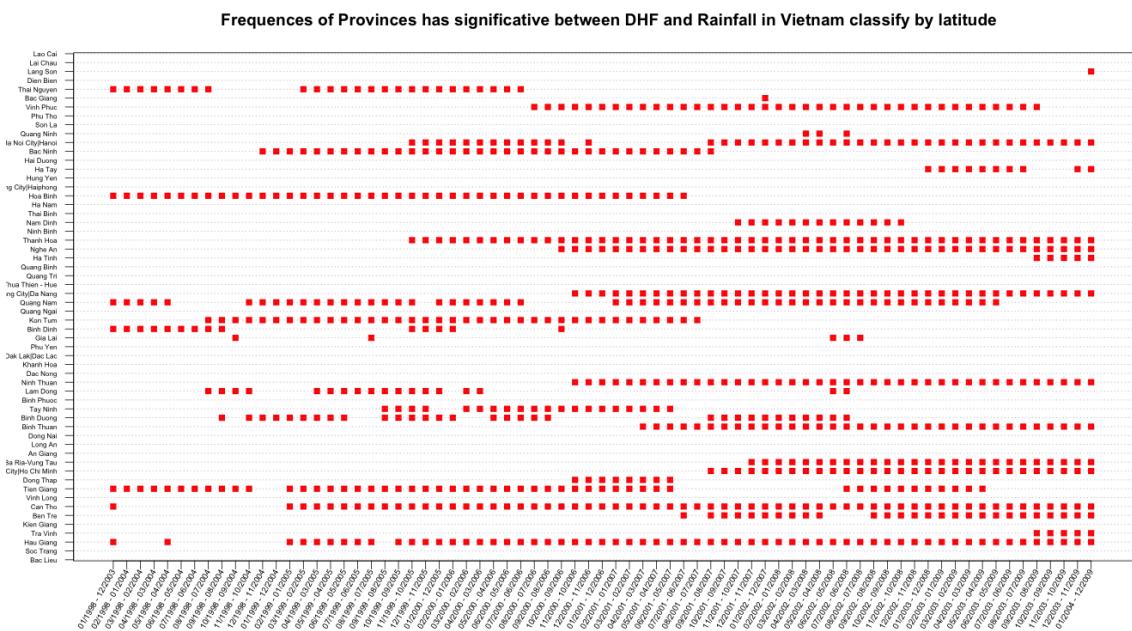


FIGURE A.11 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and rainfall in this province. Provinces on Y axis are ordered according to their latitude.

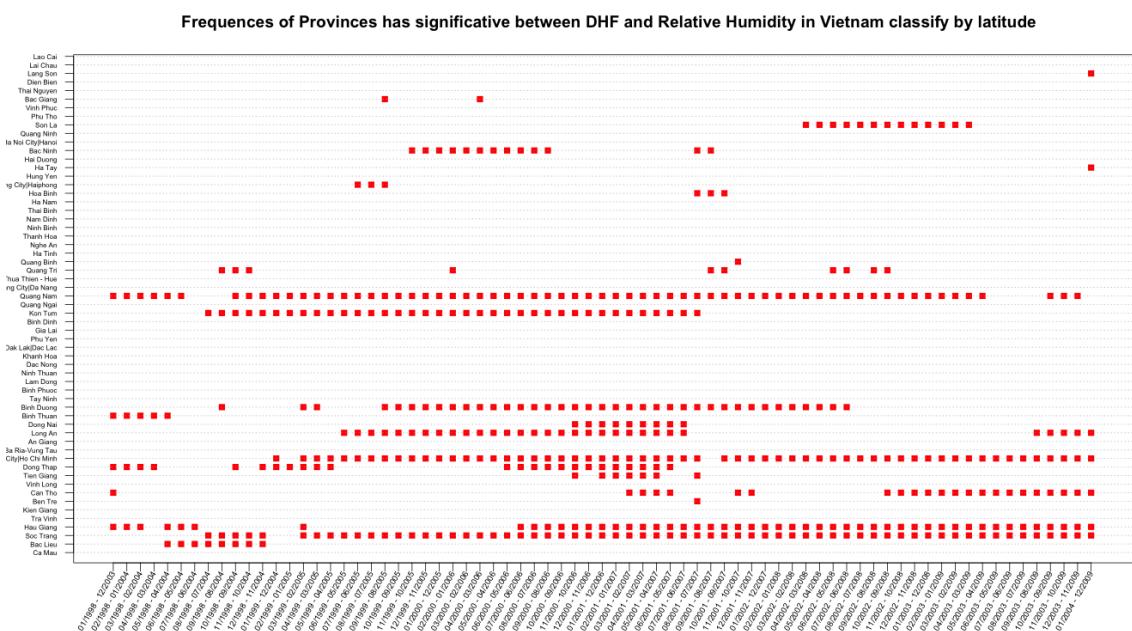


FIGURE A.12 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and relative humidity in this province. Provinces on Y axis are ordered according to their latitude.

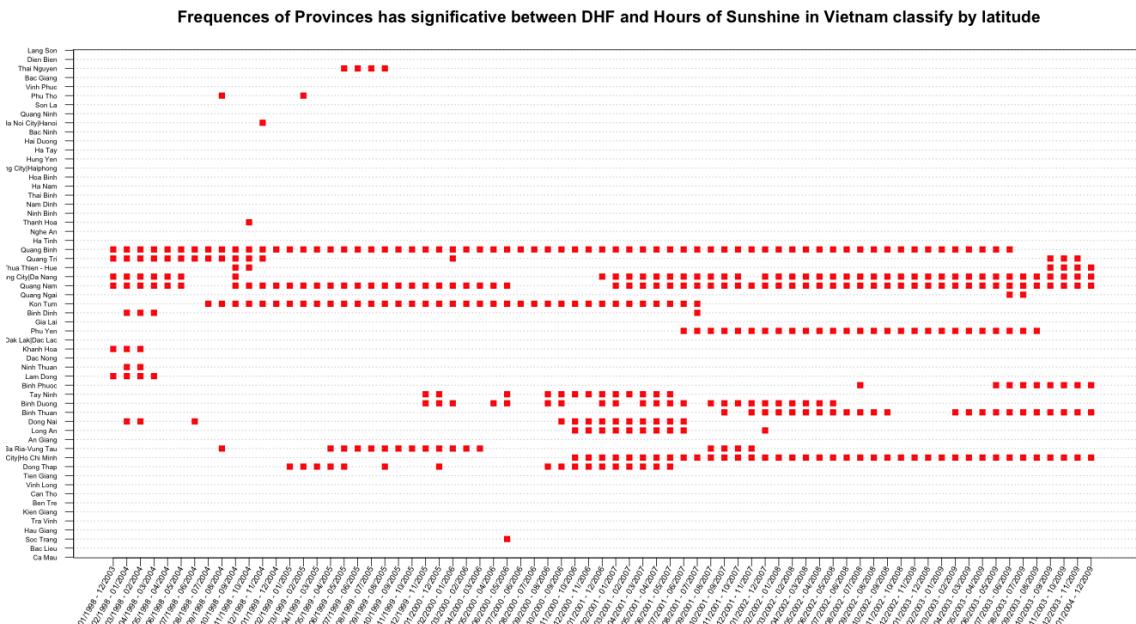


FIGURE A.13 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and hours of sunshine in this province. Provinces on Y axis are ordered according to their latitude..

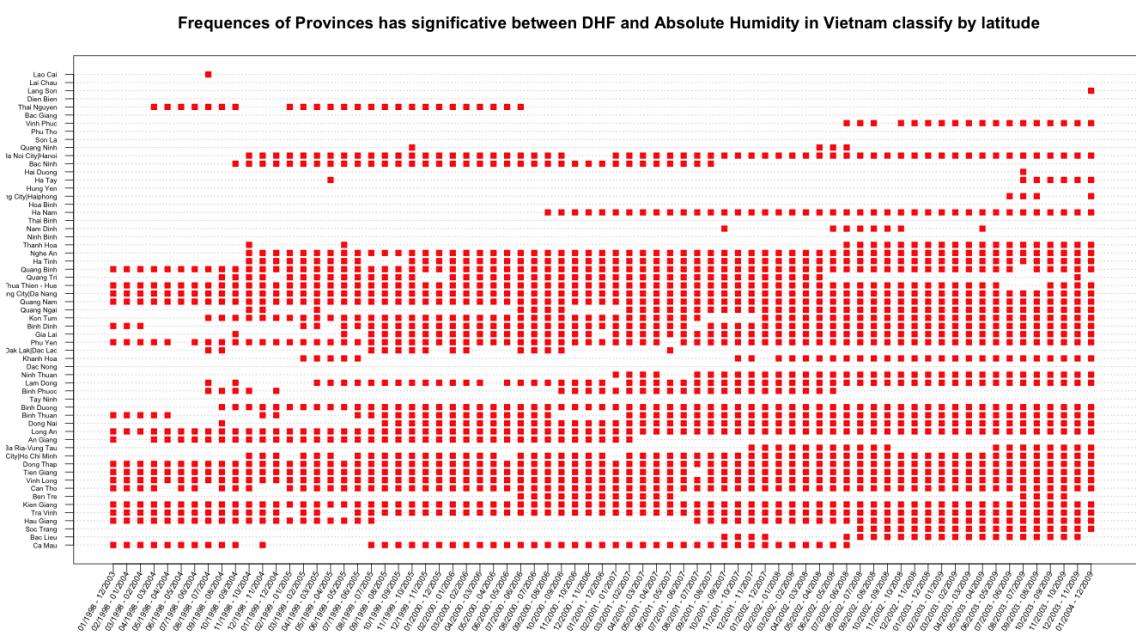


FIGURE A.14 – The exploration of result subsequence test with GC univariable method. The length of each period  $L = 6$  years. The red square in each period present the relationship between dengue incidence and absolute humidity in this province. Provinces on Y axis are ordered according to their latitude.