REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ DES MATHEMATIQUES ET D'INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Réseau et Télécommunication

Par :

Belkheirat Zineb

Sur le thème

Prévision de l'intensité des cyberattaques par les séries chronologiques.

Soutenu publiquement le 24 /06 /2024 à Tiaret devant le jury composé de :

Mr Alem Abdelkader	Grade	Université MAA	Président
Mr Bekkar khaled	Grade	Université MAA	Encadrant
Mr Benaouda Habib	Grade	Université MAA	Examinateur

2023-2024

ملخص

أصبحت الهجمات الالكترونية شائعة جدا في عصر الإنترنت، حيث أنها تزداد كل عام وتزداد شدة الأضرار أيضًا. ضمان الأمن ضد الهجمات السيبرانية أصبح أهم في عالم الشبكات. ومع ذلك، فإن ضمان الأمن السيبراني أمرًا معقدًا للغاية لأنها تتطلب المعرفة في المجال حول الهجمات والقدرة على تحليل احتمالات التهديدات الأساسية، تحدي الأمن السيبراني هو التنبأ بكثافة الهجمات السيبرانية، هذ ه المذكرة توضح تحليل السلاسل البيانية باستخدام نموذج **ARIMA** من خلال خوارزميات التعلم الآلي التي يمكن تطبيقها على كشف الهجمات السيبرانية والتحليلات الإحصائية لشدة الهجوم وتحديث النموذج.

الكلمات المفتاحية: الأمن السيبراني، الهجوم، السلاسل الزمنية، ARIMA(المتوسط المتحرك المتكامل ذاتي الانحدار)، التنبؤ...

Résumé

Les cyberattaques sont devenues très courantes à l'ère d'Internet, car elles augmentent chaque année et la gravité des dommages augmente également.

Assurer la sécurité contre les cyberattaques est devenu plus important dans le monde des réseaux. Cependant, assurer la cybersécurité est très complexe car cela nécessite des connaissances approfondies sur les attaques et la capacité à analyser la probabilité des menaces sous-jacentes.

La cybersécurité est de prédire l'intensité des cyberattaques. Cette mémoire illustre l'analyse de séries chronologiques à l'aide du modèle ARIMA (AutoRegressive Integrated Moving Average) grâce à des algorithmes d'apprentissage automatique, qui peuvent être appliqués pour détecter les cyberattaques et effectuer des analyses statistiques de la gravité de l'attaque, ainsi que pour mettre à jour le modèle.

Mots clés : La cybersécurité, attaque, les séries chronologiques, ARIMA (AutoRegressive Integrated Moving Average), Prévision …

Remerciement

Je remercie tout d'abord dieu « ALLAH» de m'avoir permis de terminer ce travail dans les meilleures conditions et qui a éclairé mon chemin et m'a doué de la connaissance.

Je remercie mon encadreur Dr. Bekkar khaled qui m'a toujours soutenu par son aide et ses précieux conseils.

Le professeur Dr. Bouazza Abdelhamid, merci pour votre aide et vos conseils.

Je remercie les membres du jury pour avoir accepté de juger mon présent travail.

Je remercie aussi à ceux qui ont toujours cru en moi, qui m'ont toujours poussé et qui ont toujours été là pour moi…merci à toute ma famille,

J'exprime mon remerciement à tous les enseignants de département d'informatique qui m'ont suivi durant mon parcours académique et qui ont su me transmettre leurs savoirs faire.

Enfin, Je remercie tous mes amis GHALIA, NOURELHOUDA et tous ceux qui ont contribué à l'élaboration de ce travail de près ou de loin et qui méritent d'y trouver leurs noms.

Dédicace

Je dédie ce mémoire

A mes trés chers parents ma mère et mon père pour leur patience, leur amour et leur encouragements.

A mes sœurs Oum Elkheir, Houria, Khadra, Fatima et ses enfants.

A mes frère mohamed et Abd Elkhader et ses enfants.

En particulier a mes camarades Alia et Asmaa.

A toute personnes qui m'ont encouragé ou aide au long de mes études.

Table des matières

Chapitre 2 : Généralités sur les séries chronologiques

Liste des figures

Liste des tables

Liste des abréviations

X_t processus aléatoire indicé par le temps

AR Autorégressif

MA Moyennes Mobiles (Moving Average).

ARIMA Autorégressif moyenne mobile intégré

ARMA Autorégressif moyenne mobile

MAE Erreur absolue moyenne

MAPE Ecart absolu moyen en pourcentage

ADF Test de Dickey Fuller Augmenté

AIC Critère d'information d'Akaike

BIC Critère de Schwarz

MSE Erreur quadratique moyenne

RMSE Racine carrée de l.erreur quadratique moyenne

(ACF) La fonction d"autocorrélation

(PACF) La fonction d'autocorrélation partielle

p, d, q les valeurs des paramètres du modèle ARMA

Introduction générale

La **cybersécurité** est un domaine de recherche en constante évolution. Les méthodes d'attaque et de défense progressent simultanément, avec la découverte de nouvelles vulnérabilités et la correction des vulnérabilités précédentes. L'accent est généralement mis sur la détection et la prévention des attaques cybernétiques en identifiant les modèles de comportement d'attaque et en tentant de les reconnaître en temps réel. Bien que ces méthodes soient efficaces, elles nécessitent l'accès au réseau de la victime et ne peuvent détecter que les attaques en cours[1], Il serait extrêmement précieux pour une victime potentielle de savoir à l'avance qu'elle sera la cible d'une attaque cybernétique **[2].**

Nous utilisons des séries chronologiques pour enregistrer le nombre d'attaques et analyser les corrélations temporelles entre les attaques successives. En appliquant des techniques de prédiction des séries chronologiques aux données relatives aux cyberattaques, nous sommes en mesure de révéler des schémas temporels tels que des augmentations ou des baisses d'activité à court terme, ainsi que des variations saisonnières dans les taux d'attaques. Pour ce faire, nous avons développé un système basé sur des modèles de prévision ARIMA (moyenne mobile intégrée autorégressive) qui vise à prédire l'intensité du nombre d'attaques cybernétiques à une date future, en se basant sur le taux d'occurrence des attaques précédentes. Ce système est constamment mis à jour à mesure que de nouvelles données sont acquises, ce qui lui permet de détecter dynamiquement les changements dans la corrélation et la saisonnalité. Le mémoire s’articule autour de trois chapitres :

1. Le premier chapitre est consacré à la présentation générale de la Sécurité des réseaux informatiques
2. le deuxième chapitre présente les séries chronologiques et leurs modèles.
3. Le dernièr chapitre est consacre a l’implemetation du système de prediction et les resultats obtenus sur les cyberattaques enregistres au cours des l’annees 2020/2021/2022.

CHAPITRE 1

Sécurité des réseaux informatiques

Introduction

La cybersécurité est l'effort continu de protéger les individus, les organisations et les gouvernements contre les attaques numériques en sécurisant les systèmes d'information tels que les ordinateurs, les serveurs, les appareils mobiles, les réseaux et les données. Pour garantir des réseaux optimaux, cohérents, performants et sécurisés, il est essentiel d'avoir des solutions de sécurité fiables qui utilisent le concept de résilience. Ces mesures devraient être capables de protéger contre de nouvelles attaques et de s'adapter aux menaces constantes. En combinant les concepts de sécurité réseau et les attaques potentielles.

1.1. Définition : La sécurité informatique c'est l'ensemble des moyens mis en oeuvre pour réduire la vulnérabilité d'un système contre les menaces accidentelles ou intentionnelles. L'objectif de la sécurité informatique est d'assurer que les ressources matérielles et/ou logicielles d'un parc informatique sont uniquement utilisées dans le cadre prévu et par des personnes autorisées **[3].**

1.2. Objectives de sécurité

« Le système d'information représente un patrimoine essentiel de l'organisation, qu'il convient de protéger. La sécurité informatique consiste à garantir que les ressources matérielles ou logicielles d'une organisation sont uniquement utilisées dans le cadre prévu ».

La sécurité des systèmes d'information vise les objectifs suivants (C.A.I.D.)

- **La confidentialité :** Seule les personnes habilitées doivent avoir accès aux données. Toute interception ne doit pas être en mesure d'aboutir, les données doivent être cryptées, seuls les acteurs de la transaction possèdent la clé de compréhension.
- **L'intégrité :** Il faut garantir à chaque instant que les données qui circulent sont bien celles que l'on croit, qu'il n'y a pas eu d'altération (volontaire ou non) au cours de la communication. L'intégrité des données doit valider l'intégralité des données, leur précision, l'authenticité et la validité.
- **La disponibilité :** Il faut s'assurer du bon fonctionnement du système, de l'accès à un service et aux ressources à n'importe quel moment. La disponibilité d'un équipement se mesure en divisant la durée durant laquelle cet équipement est opérationnel par la durée durant laquelle il aurait dû être opérationnel.
- **La non-répudiation :** Une transaction ne peut être niée par aucun des correspondants. La non-répudiation de l'origine et de la réception des données prouve que les données ont bien été reçues. Cela se fait par le biais de certificats numériques grâce à une clé privée.

- **L'authentification :**Elle limite l'accès aux personnes autorisées. Il faut s'assurer de l'identité d'un utilisateur avant l'échange de données **[3].**

1.3. Les Mécanismes de sécurité

Le tableau **1.1** énumère les mécanismes de sécurité sont divisés en ceux implémentés dans une couche de protocole spécifique et ceux qui ne sont pas Spécifiques à une couche de protocole ou à un service de sécurité particulier.

- Mécanismes de sécurité spécifiques : Peuvent être incorporé dans la couche de Protocole appropriée afin de fournir certains des services de sécurité OSI.
- Mécanismes de sécurité omniprésents : Mécanismes qui ne sont pas spécifiques à un Service de sécurité OSI ou à une couche de protocole particulier **[4].**

	Mécanisme	**Description**
Mécanismes De Sécurité Spécifiques	Le chiffrement	L'utilisation d'algorithmes mathématiques pour transformer les données en une forme qui n'est pas facilement intelligible. La transformation et la récupération ultérieure des données dépendent d'un algorithme et de clés de cryptage.
	Signature Numérique	Les données sont ajoutées, ou transformation cryptographique d'une unité de données.
	Contrôle d'accès	Divèrs mécanismes sont utilisés pour imposer les droits d'accès aux ressources. Ils garantissent que seules les personnes autorisées peuvent utiliser les ressources matérielles et logicielles d'un réseau.
	Intégrité des Données	Divèrs mécanismes utilisés pour assurer l'intégrité d'une unité de données ou d'un flux d'unités de données.
	Echange d'authentification	Un mécanisme destiné à assurer l'identité d'une entité au moyen d'un échange d'informations.
	Remplissage du trafic	L'insertion de bits dans des intervalles dans un flux de données pour éviter les tentatives d'analyse de trafic.
	Contrôle de routage	Permet de sélectionner des routes physiquement sécurisées particulières pour certaines données et permet des modifications de routage, en particulier lorsqu'une violation de sécurité est suspectée.
	Notarisation	L'utilisation d'un tièrs de confiance pour assurer certaines

		propriétés d'un échange de données.
Mécanismes de sécurité omniprésents	Fonctionnalité de Confiance	Ce qui est perçu comme correct par rapport à certains critères (par exemple, tel qu'établi par une politique de sécurité).
	Étiquette de sécurité	Le marquage lié à une ressource (qui peut être une unité de données) qui nomme ou désigne les attributs de sécurité de cette ressource.
	Détection d'événement	Détection des événements liés à la sécurité.
	Sentier d'audit de sécurité	Les données recueillies et utilisées pour faciliter une vérification de la sécurité, qui est un examen et un examen indépendants des dossiers et des activités du système.
	Récupération de sécurité	Aborde les demandes de mécanismes, telles que la gestion des événements et prend des mesures de récupération.

Tableau 1.1Les mécanismes de sécurité

1.4.Les attaques d'un réseau

1.4.1 Définition d'une attaque

Les attaques de réseau sont un ensemble d'activités malveillantes qui perturbent, refusent, dégradent ou détruisent les données et les services des réseaux informatiques. Une attaque réseau cible l'intégrité, la confidentialité ou la disponibilité des systèmes de réseau informatique en exploitant le flux de données sur les réseaux. **[5].**

1.4.2 Anatomie d'une attaque

Une attaque est souvent décrite à l'aide des 5 "p" **:**

- **Probe(Analyser) :** c'est la collecte d'informations sur le système cible, elle peut s'effectuer de plusieurs manières. Comme par exemple un scan des ports grâce au programme Nmap pour déterminer la version des logiciels utilisés, et des outils comme firewalk, hping ou SNMP Walk permettent quant à eux de découvrir la nature d'un réseau.
- **Penetrate (Pénétrer):** Utilisation des données collectées afin de pénétrer un réseau informationnel. Des méthodes telles que le brute force ou les attaques par dictionnaires peuvent être employées afin de dépasser les protections par mot de passe.

- **Persist (Peréniser) :** Créez un compte avec les droits de super utilisateur afin de pouvoir effectuer une réinfiltration ultérieurement. Une autre méthode implique l'installation d'une application de contrôle à distance qui puisse faire face à un redémarrage.
- **Propagate(Propager) :** le réseau est infiltré, l'accès est pérein. Le pirate pourra alors explorer le réseau et trouver de nouvelles cibles qui l'interresseraient.
- **Paralyse (Paralyser) :** Cette étape peut impliquer différentes actions. Le pirate a la possibilité d'utiliser le serveur afin d'attaquer une autre machine, de détruire des données ou de causer des dommages au système d'exploitation dans le but de faire planter le serveur **[6].**

1.4.3. Différent types d'attaques : Une attaque peut être active ou passive.

- Une **«attaque active»** tente de modifier les ressources du système ou d'affecter leur fonctionnement.
- Une **«attaque passive»** tente d'apprendre ou d'utiliser des informations du système mais n'affecte pas les ressources du système. (P. Ex., Écoute téléphoniques).

  Une attaque peut être perpétrée de l'intérieur ou de l'extérieur de l'organisation.
- Une **«attaque interne»** Il s'agit d'une attaque lancée par une entité située dans le domaine de la sécurité, c'est-à-dire une entité qui a obtenu l'autorisation d'accéder aux ressources du système mais qui les utilise de manière non approuvée par ceux qui ont donné quelle autorisation
- Une **«attaque extérieure»** Un utilisateur non autorisé ou illégitime du système l'initie depuis l'extérieur du périmètre **[4].**

1. 5. Les types attaques d'un réseau existent plusieurs types des attaques très connues dans le monde de l'informatique, nous détaillons ici quelque exemple d'attaques recueillies à partir de la base de données Hackmageddon.

1.5.1 Les attaques s'appuyant sur les faiblesses d'authentification

1.5.1.1 IP Spoofing : est une méthode dans laquelle un pirate falsifie son adresse IP pour se connecter à un autre système. Ils choisissent un système cible, déterminent les adresses IP autorisées et envoient plusieurs paquets au serveur cible. Le pirate rend alors la machine inopérable, remplace l'adresse IP invalide et envoie une demande de connexion. Cette attaque est difficile à exécuter car elle ne reçoit pas de données sur le serveur **[7].**

1.5.1.2 L'attaque man-in-the-middle : L'attaquant pénètre entre deux systèmes sans que l'un d'eux réalise qu'il existe un troisième système qui permet de passer les échanges réseau. Pour que cette attaque soit efficace, il est nécessaire que la machine de l'attaquant soit physiquement située entre les deux machines.

Les victimes peuvent être victimes ou l'attaquant peut modifier le routage du réseau pour que sa machine devienne l'un des points de passage **[07].** Le schéma ci-dessous montre comment l'attaque fonctionne. L'homme dans le milieu (Voir **Figure 1.1**)

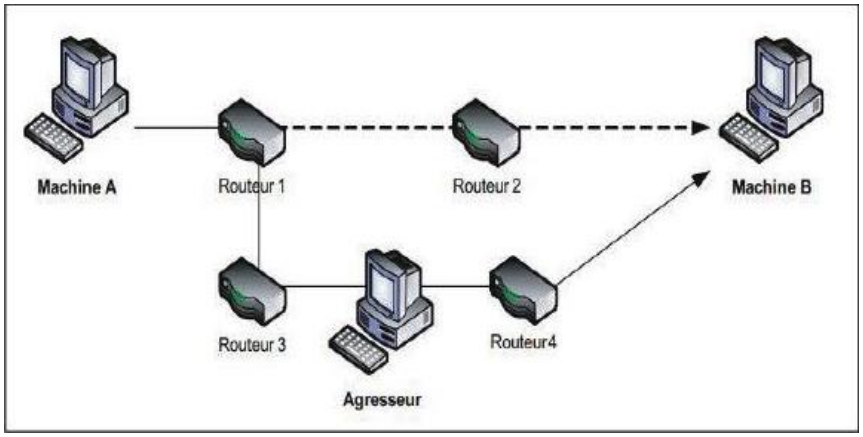


FIGURE 1.1 – L'attaque man-in-the-middle [07]

1.5.2 Les autres formes d'attaques

1.5.2.1 DNS Spoofing

Le terme Spoofing, qui signifie « usurpation » ou « falsification », englobe différents scénarios dans lesquels une manipulation est opérée sur la résolution de nom DNS.

Le DNS spoofing est une technique utilisée par les attaquants pour corrompre les serveurs DNS afin de rediriger le trafic d'un utilisateur vers un site Web malveillant.

- Cette attaque fonctionne en modifiant les données du cache DNS. Ainsi, lorsque

L'utilisateur tente d'accéder au site Web légitime, son navigateur est redirigé vers le site Web malveillant contrôlé par l'attaquant.

- Cette attaque peut être utilisée pour voler des informations sensibles telles que des identifiants de connexion ou des informations de carte de crédit **[9]**

1.5.2.2 ARP Spoofing

La mystification ARP, particulièrement utilisable sur les réseaux Ethernet LAN, utilisant le protocole TCP/IP. Cette attaque est basée sur l'accouplement d'une adresse IP avec une adresse MAC non correspondante, pour que la victime soit redirigée vers une autre machine **[10].**

1.5.2.3 TCP Session Hijacking

Le hijacking de session est une attaque dans laquelle un attaquant prend le contrôle d'une session de communication valide entre deux ordinateurs utilisant le protocole TCP (Transmission Control Protocol).

- La majorité des types d'authentification ne sont effectué qu'au début d'une session TCP, ce qui permet à un attaquant d'accéder à une machine pendant qu'une session est en cours.
- Les attaquants peuvent intercepter tout le trafic des sessions TCP établies et effectuer des vols d'identité ou d'informations, des fraudes, etc.
- Le hijacking de session est une attaque qui utilise un mécanisme de génération de jetons de session ou des contrôles de sécurité de jetons pour permettre à l'attaquant de créer une connexion non autorisée avec un serveur cible.
- Les cybercriminels ont la capacité d'intercepter l'intégralité du trafic des sessions TCP établies et de commettre des vols d'identité ou d'informations, des fraudes, etc.
- L'attaquant peut deviner ou voler un identifiant de session valide, qui identifie les utilisateurs authentifiés, et l'utiliser pour établir une session avec le serveur. Le serveur web répond aux demandes de l'attaquant en croyant qu'il communique avec un utilisateur authentifié.
- Les attaquants peuvent utiliser le hijacking de session pour lancer différents types d'attaques, telles que des attaques de type man-in-the-middle (MITM) et des attaques de déni de service (DoS).

Le hijacking de session peut être divisé en trois grandes phases :

1. Suivi de la connexion.
2. Désynchroniser la connexion.
3. Injecter le paquet de l'attaquant **[8]**

I.5.2.4 Injection SQL

L'injection SQL est une faille de sécurité web qui permet à un pirate d'entrer en contact avec les demandes effectuées par une application dans sa base de données. En règle générale, cela donne à un attaquant la possibilité de visualiser des données qu'il ne peut normalement pas récupérer. Il peut s'agir d'informations provenant d'autres utilisateurs, ou de toute autre information à laquelle l'application peut accéder elle-même. Dans de nombreuses situations, un pirate informatique a la possibilité de modifier ou de supprimer ces informations, ce qui entraîne des modifications continues dans le contenu ou le comportement de l'application.

Il arrive parfois qu'un pirate informatique puisse renforcer une attaque par injection SQL pour mettre en péril le serveur sous-jacent ou une autre infrastructure dorsale, ou encore mener une attaque par déni de service **[8]**

Les Attaque par logiciel malveillant

Un logiciel malveillant ou maliciel (ou « malware » en anglais) est un programme conçu pour causer des dommages à un système informatique, sans le consentement de l'utilisateur dont l'ordinateur est infecté.

De nos jours, le terme « virus » est souvent employé, à tort, pour désigner toutes sortes de logiciels malveillants. En effet, les malicieux englobent les virus, les vers, les chevaux de Troie, ainsi que d'autres menaces **[10].**

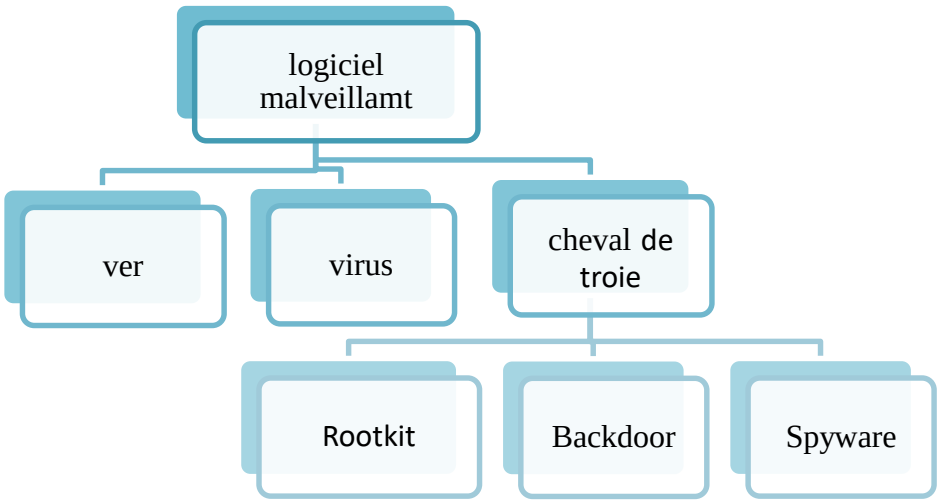


Figure 1.2 type de Logiciel malveillant.

1.6.1 Malware : Ransomware

- Ce malware est conçu pour maintenir captif le système d'un ordinateur ou sés données, jusqu'à Ce qu'un paiement soit effectué. Les rançongiciels chiffrent généralement vos données pour vous empêcher d'y accéder.
- D'autres versions de rançongiciel (Ransomware) peuvent tirer parti de vulnérabilités spécifiques du système pour le verrouiller.
- Les ransomware se propagent souvent par le biais d'e-mails de phishing qui vous encouragent à télécharger une pièce jointe malveillante ou par le biais d'une vulnérabilité logicielle

1.6.2 Malware: Scareware

- Il s'agit d'un type de malware qui utilise des tactiques de « peur » pour vous inciter à effectuer une action spécifique.
- Les Scarewares se composent principalement de fenêtres de type système d'exploitation qui s'affiche pour vous avertir que votre système est menacé et qu'il doit exécuter un programme spécifique pour revenir à un fonctionnement normal.

Si vous acceptez d'exécuter le programme en question, votre système sera infecté par un malware[8].

1.6.3 Malware : Spyware

- les logiciels espions (Spyware) sont capables de surveiller votre activité en ligne et de sauvegarder chaque touche de votre clavier.
- En outre, ils enregistrent quasiment toutes vos informations, y compris les données personnelles confidentielles comme vos coordonnées bancaires en ligne.Les pirates informatiques altèrent les réglages de sécurité de vos appareils.
- Souvent, le Spyware est associé à des logiciels légitimes ou à des chevaux de Troie

1.6.4 Malware : Adware

Logiciel qui affiche des bannières publicitaires dans les navigateurs comme Internet Explorer et Mozilla [10].

Le but principal de l'Adware est de générer des revenus pour les créateurs de logiciels malveillants

1.6.5 Malware : Backdoor

En utilisant des portes dérobées (Backdoors), Une application qui permet à des systèmes distants d'avoir accès aux ordinateurs **[12]**

1.6.6. Les ROOTKITS: Les rootkits sont un ensemble de techniques utilisées par le logiciel pour obtenir un accès non autorisé à une machine cible. Ils peuvent être utilisés pour l'espionnage, l'accès aux données stockées, ou la manipulation de la machine cible. Bien que souvent considéré comme un logiciel malveillant, ce n'est pas toujours le cas. ils peuvent utiliser des "techniques virales" pour se propager. (par exemple, en utilisant un virus ou un cheval de Troie) **[10]**

1.6.7 Malware : Virus

Les virus informatiques sont tous les programmes qui peuvent se reproduire, ce qui en fait le type d'attaque le plus courant. Une fois mis en marche, il peut prendre la forme d'une routine ou d'un programme, utilisant toutes les méthodes pour perturber le système. Plusieurs types de virus peuvent être évoqués

- Virus de secteur d'amorçage.
- Virus d'infection des fichiers (parasites).
- Virus non-résidents mémoire.
- Virus résidents mémoire.
- Bombes logiques.

1.6.8 Malware : Cheval de troi (Trojan horse en anglais)

On ne doit pas confondre ce type de logiciel malveillant avec les virus ou autres parasites. En apparence légitime, le cheval de Troie est un logiciel qui renferme un programme malveillant. Il a pour mission d'introduire ce parasite sur l'ordinateur et de l'installer sans que l'utilisateur le sache **[10].**

1.6.9 Malware Vers: est un code autonome particulièrement dangereux, qui se réplique et propage sans Aucune action de la part de l'utilisateur. La plupart des virus se dissimulent dans les pièces jointes de courriel et retournent à l'ordinateur lors de l'exploitation. Ils recherchent des fichiers comme des carnets d'adresses ou des pages Web temporaires contenant des adresses électroniques infectées. Ceux-ci sont utilisés pour envoyer des messages infectés, souvent envahissant le système avant d'être réparés **[10].**

1.7. Les Attaques DoS

1.7.1. **Attaque DoS (denial of service attack):** Dans les attaques par déni de service, un noeud malveillant envoie le message à d'autres noeuds et utilise la bande passante du réseau. L'objectif principal du noeud malveillant est de rendre le réseau occupé. Si un message de nœud non authentifié arrive, le récepteur ne recevra pas cet avis car il est occupé et l'initiateur doit attendre cinq avertissements pour la réponse du récepteur **[10].**

1.7.2. Attaque DDoS (distributed denial of service attack) :

L'attaquant utilise un réseau d'ordinateurs et d'objets connectés sous son contrôle pour multiplier les sources et la force de l'attaque. Plusieurs techniques différentes sont possibles pour amplifier l'attaque, comme illustré à la figure (**1. 3)** suivante.

La première étape consiste à pénétrer par diverses méthodes des systèmes dits handlers, ou maîtres (masters), et agents, ou esclaves (slaves). Le pirate contrôle ensuite dans une deuxième étape directement un ensemble desystèmes handlers, qui contrôlent eux-mêmes un ensemble de systèmes agents. La troisième étape consiste pour le pirate à déclencher son attaque vers un ou plusieurs systèmes cibles donnés. Cet ordre d'attaque aura été donné par les systèmes handlers, qui eux-mêmes auront reçu cet ordre du pirate **[11].**

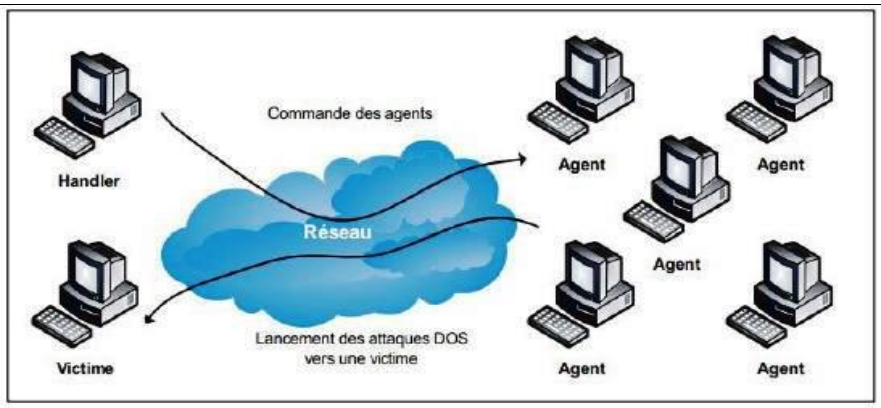


FIGURE 1.3 – L'attaque DDoS [11]

1.7.3. Types d'attaques DDoS

Voici quelques exemples spécifiques d'attaques DDoS mémorables :

- **SYN flood :** un attaque exploite une communication TCP (SYN-ACK) en envoyant une grande quantité de paquets SYN, ce qui consomme les ressources du système ciblé.
- **Spoofing :** un hacker se fait passer pour un utilisateur ou un appareil et, après avoir gagné sa confiance, utilise despaquets usurpés pour lancer une cyberattaque.
- **Attaque DDoS de la couche application :** Comme son nom l'indique, une fois déployée, cette attaque exploitera une vulnérabilité ou une configuration incorrecte dans une application et refusera à un utilisateur d'accéder ou d'utiliser l'application.
- **Domain name system (DNS) flood :** avec cette attaque également connue sous le nom d'attaque par amplification DNS, une attaque perturbe la résolution DNS d'un nom de domaine donné en inondant ses serveurs.
- **Internet control message protocol (ICMP) flood :** Aussi appelé Ping Flood, envoi massif de Paquets (ping) impliquant la réponse de la victime (pong) ayant le même contenu quele paquet d'origine.
- **User datagram protocol (UDP) flood :** consiste en l'envoi d'une grande quantité de paquets UDP sur des ports aléatoires de la machine victime. Ainsi, celle-ci sera obligée de répondre par l'envoi de nombreux paquets ICMP, la rendant inaccessible par d'autres clients.
- **IP Packet Fragment Attack** Ressources Envoi de paquets IP référençant volontairement d'autres paquets qui ne seront jamais envoyés, saturant ainsi la mémoire de la victime**. [11]**

1.8. Outils de défense informatique

Afin de remédier, aux différentes attaques informatiques citées auparavant, ou au moins réduire leur risque, différents moyens peuvent être utilisés, parmi ces systèmes on peut citer :

1.8.1. Pare-feu

Est un dispositif dont le rôle est de bloquer tout trafic non autorisé entre deux réseaux, un réseau de confiance tel que le réseau local, et un réseau public (suspect), tel qu'internet. Un firewall permet aussi, d'isoler des sous-réseaux internes nécessitant des critères de sécurité déférents comme, par exemple, un sous-réseau de développement, un sous-réseau de teste et le sous-réseau de production **[13].**

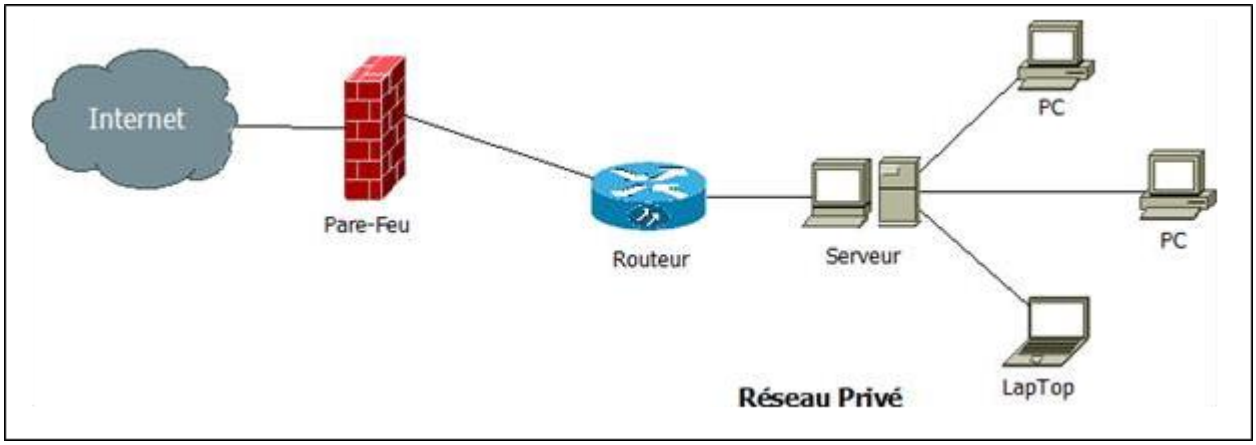


Figure1. 4 Exemple d'un pare-feu installé entre un réseau privé et Internet **[14].**

Selon le type de service de sécurité offert pour chaque service dans le modèle TCP/IP, lespare-feu peuvent être divisés en quatre types qui sont :

- Pare-feu de la couche application : offre des services tels que le cryptage, et la passerelleniveau application.
  - Pare-feu de la couche transport : offre principalement la fonctionnalité de filtrage depaquets TCP, UDP (User Datagram Protocol), ICMP (Internet Control Message Protocol).
  - Pare-feu de la couche réseau : Offre la fonctionnalité de filtrage NAT (Network Adressa Translation) et IP.
  - Pare-feu de la couche liaison de données : offre la fonctionnalité de filtrage des adresses MAC (Media Access Control) **[14].**

1.8.2 Les Antivirus

L'antivirus sont des programmes prévus pour détecter la présence de virus sur un système d'exploitation, ainsi que de les mette en quarantaine « les isoler » ou les neutraliser, sans endommager les fichiers infectés. Mais parfois, ce nettoyage simple n'est pas possible **[15].**

Ces derniers peuvent se baser sur l'exploitation de failles de sécurité, mais il peut également s'agir de logiciels modifiant ou supprimant des fichiers, que ce soit des documents de l'utilisateurstockés sur l'ordinateur infecté, ou des fichiers nécessaires au bon fonctionnement du système d'exploitation.

Il est intéressant de noter qu'une fois un fichier infecté, il ne l'est jamais deux fois. En effet, un virus est programmé de telle sorte qu'il signe le fichier dès qu'il est contaminé. On parle ainside signature de virus. Cette signature consiste en une suite de bits apposée au fichier. Cette suite, une fois décelée, permettra de reconnaitre le virus. Lorsque le virus est détecté par l'antivirus, plusieurs possibilités sont offertes pour l'éradiquer :

- Supprimer le fichier infecté ;
- Supprimer le code malicieux du fichier infecté ;
- Le placer en "quarantaine" pour un traitement futur **.**

1.8.3 VPN

Les VPN (Virtual Private Network), également connus sous le nom de réseaux privés virtuels, sont un ensemble de ressources qui peuvent être partagées par des flux de paquets ou de trames provenant de machines autorisées. Les VPN peuvent utiliser des technologies et des protocoles quelconques. La gestion de ces ressources nécessite un haut niveau d'automatisation pour obtenirla dynamique nécessaire au fonctionnement d'un VPN. Pour obtenir cette dynamique, les ressources permettant d'acheminer les paquets au destinataire doivent être gérés avec efficacité, utilisant des outils d'authentification et des systèmes cryptographiques, afin d'acheminer des paquets d'informations, à savoir sensibles, via des réseaux généralement publiques **[16].**

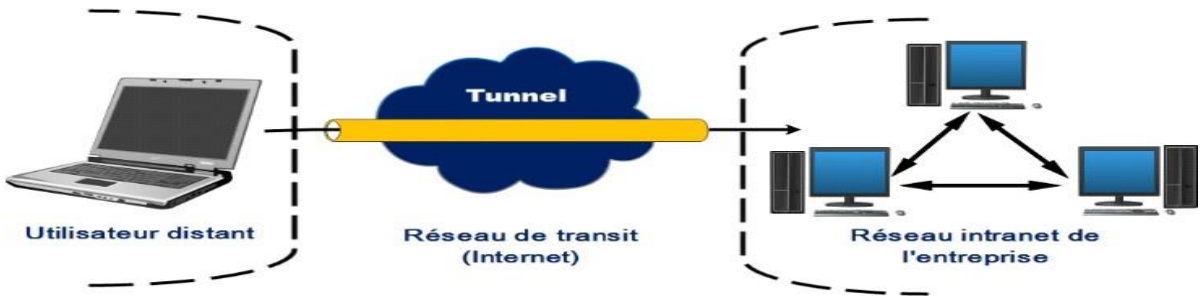


Figure 1. 5 Exemple de VPN (Ghost Warrior).

1.8.4. Définition de la cryptographie :La cryptographie est un mécanisme qui permet de mettre en œuvre le chiffrement et les signatures électroniques. Le terme lui-même provient de deux mots grecs : "Kruptus", qui signifie secret, et "Graphein", qui signifie écriture. La cryptographie est l'art de cacher l'information de manière à ce qu'elle soit incompréhensible, et englobe l'ensemble des techniques permettant de déchiffrer les messages. Son objectif principal est d'assurer la confidentialité des données. Cela implique l'utilisation d'algorithmes de chiffrement qui sont paramétrés par des clés. Il utilise Deux méthodes :

- Cryptographie symétrique ou à clé secrète
- Cryptographie asymétrique ou à clé publique **[16]**

1.8.5 Les systèmes de détection d'intrusions (IDS

Un système de détection d'intrusion (ou IDS : Intrusion Detection System) est un mécanisme destiné à repérer des activités anormales ou suspectes sur la cible analysée (un réseau ou un hôte). Il permet ainsi d'avoir une connaissance sur les tentatives réussies comme échouées des intrusions. Les IDS, les plus connus selon leurs différentes catégories sont

- **les NIDS** (Network Based Intrusion Detection System), qui surveillent l'état de la sécurité au niveau du réseau

- **les HIDS** (HostBased Intrusion Detection System), surveille le trafic sur une seule machine. Il analyse les journaux systèmes, les appels, et enfin vérifie l'intégrité des fichiers. Un HIDS a besoin d'un système sain pour vérifier l'intégrité des données. Si le système a été compromis par un pirate, les HIDS du système ont été extrêmement rares, ce qui a facilité la tâche du système de détection d'intrusion
- **les IDS hybrides**, qui utilisent les NIDS et HIDS Ils permettent, de surveiller le réseau et les terminaux. Les sondes sont placées en des points stratégiques, et agissent comme NIDS et/ou HIDS suivant leurs emplacements **[6].**

Conclusion:

Nous avons analysé divèrs aspects de la sécurité des réseaux informatiques, en définissant la sécurité informatique et en examinant ses objectifs. Et exploré les mécanismes de sécurité ainsi que les différentes formes d'attaques pouvant cibler un réseau, tout en abordant les outils de défense informatique.

Ce chapitre nous a permis de prendre conscience de l'importance des défis liés à la sécurité des réseaux informatiques et de recenser les différentes mesures de protection qui peuvent être appliquées.

CHAPITRE II

Généralités sur les séries chronologiques

2. Les Séries Chronologiques

Introduction

Dans ce chapitre, nous aborderons plusieurs concepts importants liés à l'analyse des séries chronologiques.Tout d'abord nous examinerons les principaux aspects des séries chronologiques, en commençant par leur définition Ensuite, nous étudierons les différentes composantes des séries chronologiques, à savoir la composante tendancielle, la composante saisonnière et la composante résiduelle. Nous présenterons également les raisons qui motivent l'étude des séries chronologiques. En suite nous terminons avec les models de prévistion.

2.1. Définition d'une série chronologique

On appelle séerie chronologique (série temporelle, ou chronique) une suite d'observations numériques d'une grandeur effectuées à intervalles réguliers au cours du temps.

Définition : Une série chronologique ($\boldsymbol{Y_t}, \boldsymbol{t} \in \boldsymbol{T}$) est suite d'observation d'une variable y a differentes dates. Habituellement T est de nombrable, de sorte que $\boldsymbol{t}$ =1,2,….$\boldsymbol{T}$ **[17].**

Exemples:

- Lévolution du nombre de voyageurs utilisant le train.
- Nombre mensuel de vente de voitures neuves en France.
- Nombre annuel de naissance au Maroc.
- -Consommation d'électricité mensuelle.
- -Nombre d'attaques contre les réseaux informatiques.

2.2. Objectifs principaux

L'analyse d'une série chronologique permet d'examiner, de décrire et d'expliquer l'évolution d'un phénomène au fil du temps. Elle est également utile pour effectuer des contrôles, tels que la sécurité des réseaux ou le suivi d'un processus chimique. En général, l'étude d'une série chronologique peut présenter certaines difficultés.

Mais l'un des objectifs principaux de l'étude d'une série chronologique est la **prévision** qui consiste à prévoir les valeurs futures X_T (h) (h = 1, 2,3,...) de la série chronologique à partir de ses valeurs observées jusqu'au temps T : X_1, $X2$,..., XT. La prédiction de la série chronologique au temps $t + h$ est notée $\hat{X}_T$ (h) et, en général, est différente de la valeur réelle X_T+h que prend la série au temps $T + h$. Pour mesurer cette différence, on définira l'**erreur de prédiction** par la différence $\hat{X}_T$ (h) -X_T (h) en moyenne avec l'idée que plus h est grand, plus l'erreur est grande **[18].**

2.3. Domaine d'application

Les séries chronologiques ont un domaine d'application très large. Effectivement, nous avons la possibilité de classer tout phénomène évalué dans le temps comme un domaine d'application pour ces série Nous citons par exemple :

- Finance et économtrie : évolutions des indices boursiers, des prix, des donnees économiques des entreprises.
- Médecine : analyse d'électro-encéphalogrammes et d'électrocardiogrammes.
- Traitement du signal : signaux de communications, de radars, de sonars, analyse de la parole **[18].**

2.4. Les type des modèles d'une série chronologique

Un modèle est une image simplifiée de la réalité et qui peut résumer au mieux l'information. Mais il est possible qu'un modèle soit plus efficace qu'un autre pour décrire la réalité. Ici, une liste est présentée pour résumer et classer ces modèles.

2.4.1. Modèle additif

Nous considérons dans cette section une série $(X_t)_{t\in T}$ admettant une décomposition additive

$$X_{t=}z_t + s_t + \varepsilon_t, \quad t = 1 \dots T$$

Ou ***zt*** est la composante tendancielle, s_t la composante saisonnière et d'une perturbation aléatoire $\boldsymbol{\varepsilon_t}$qui

Représente les composantes (erreurs).

- ***zt*** exprime un mouvement à moyen terme de la série. Elle est le plus souvent modélisée par une fonction polynomiale du temps.
- s_t la composante saisonnière exprime un phénomène qui se reproduit de manière analogue sur chaque intervalle de temps successif.
- $\boldsymbol{\varepsilon_t}$ designe la composante aléatoire de la serie au temps de t.

2.4.2 Modèle multiplicatif

Soit une série $(X_t)_{t\in T}$ admettant une décomposition multiplicative :

$$X_t = Z_t(1 + S_t)(1 + \varepsilon_t), t = 1 \dots T$$

2.4.3 Modèle mixte

Il s'agit là de modèles ou addition et multiplication sont utilisées. On peut supposer par exemple que la composante saisonnière agit de façon multiplicative alors que les fluctuations irrégulières sont additives **[19].**

$$X_t = Z_t * S_t + \varepsilon_t \ , t = 1 \dots T$$

2.5.Choix du modèle

Avant de procéder à une modélisation et à une étude approfondie du modèle, il est essentiel de déterminer si nous sommes en présence d'une série où la variation saisonnière ***S*** s'ajoute simplement à la tendance ***Z***, ce qui est appelé le modèle additif.

La tendance ***Z*** est multipliée par la variation saisonnière ***S*** ; c'est le modèle multiplicatif.

Afin de faire cette distinction, on peut se baser sur une méthode graphique ou utiliser une méthode analytique.

- **Méthode de la bande**

On fait un graphique représentant la série chronologique (**Figure 2.1 et Figure 2.2**), puis on trace une droite passant respectivement par les minima et par les maxima de chaque saison. Si ces deux droites sont parallèles, nous sommes en présence d'un modèle additif. Dans le cas contraire, c'est un modèle multiplicatif.

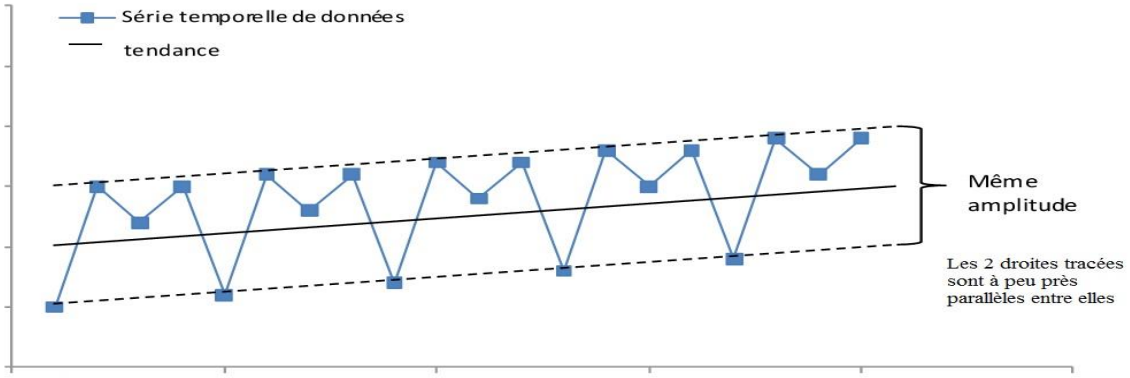


Figure 2.1 le modèle additif

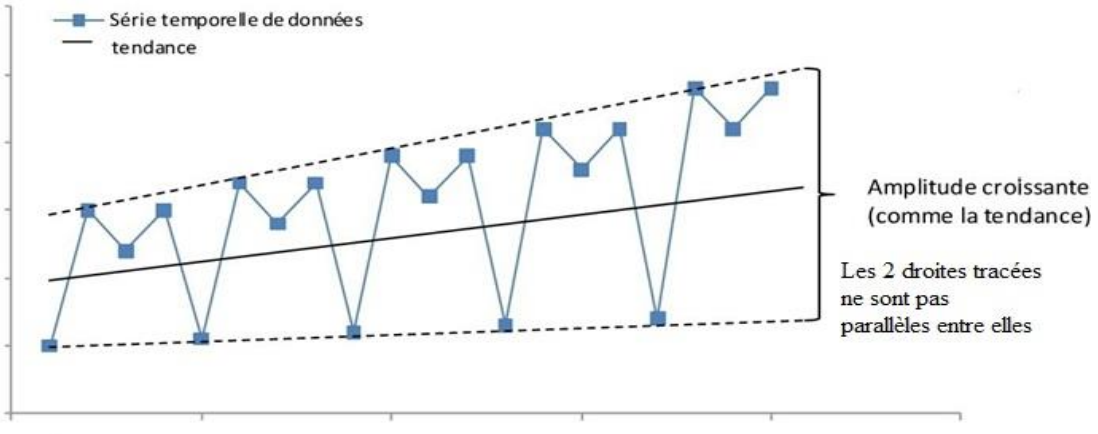


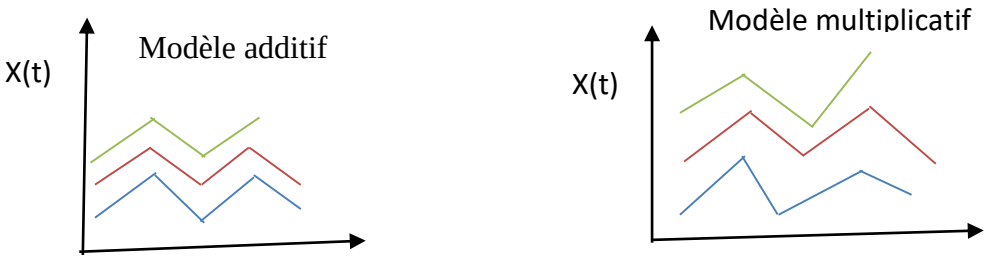
Figure 2.2 le modèle multiplicatif

Méthode du profil

Pour faire la détermination entre modèle additif et modèle multiplicatif graphiquement, on peut par exemple superposer les saisons représentées par des courbes de profil sur unmême graphique.

Si ces courbes sont parallèles, le modèle est additif, autrement le modèle est multiplicatif.

Exemple :



- **Méthode analytique**

On calcule les moyennes $\overline{x}$ et les écarts-types pour chacune des périodes considérées puis la droite des moindres carrés $\sigma = a\overline{x} + b$.

Si a est nul, c'est le modèle additif, si non c'est le modèle multiplicatif .

2.6 Les composantes fondamentales d'une série chronologique

On distingue en général qu'une série chronologique (Xt) est la résultante de trois composantes fondamentales :

- **tendance (trend)**

La tendance (ou trend) Z_t) représente l'évolution à long terme de la série étudiée et traduit le comportement "moyen" de la série.

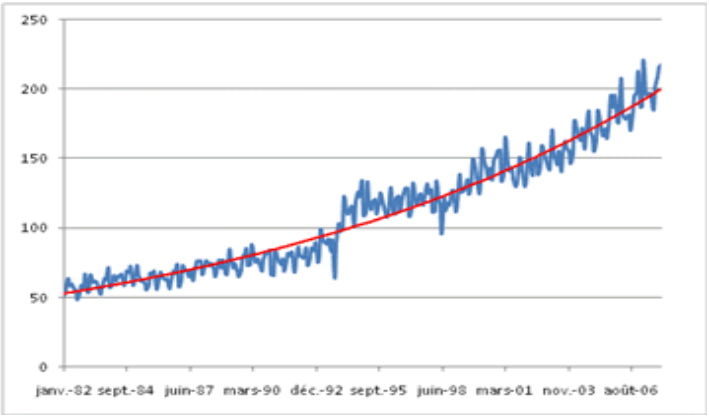


FIGURE 2.3 : Tendance série chronologique

La courbe ci-dessous montre l'évolution du frais aériens des (Aéroports de Paris) entre Janvier 82 et décembre 2007.

- **La composante saisonnière** (ou saisonnalité) (S_t) correspond à un phénomène qui se répète à intervalle de temps réguliers (périodiques i.e. $S_{(t+k\times p)} = S_t$). En général, c'est un phénomène saisonnier d'où le terme de variations saisonnières

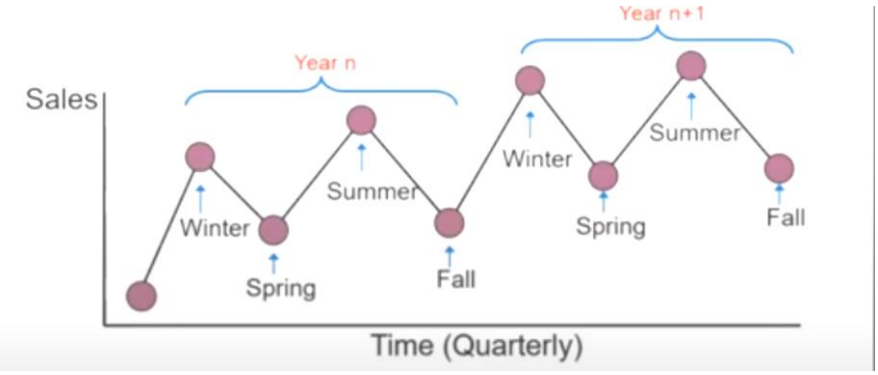


FIGURE 2.4 : Séries chronologies saisonnières [20]

- **La composante résiduelle** (**ou bruit ou résidu**) (εt) correspond à des fluctuations irrégulières, en général de faible intensité et de nature aléatoire, elle provient de circonstances imprévisibles : catastrophes naturelles, crise boursière, grèves …, On parle aussi d'aléas **[20]**

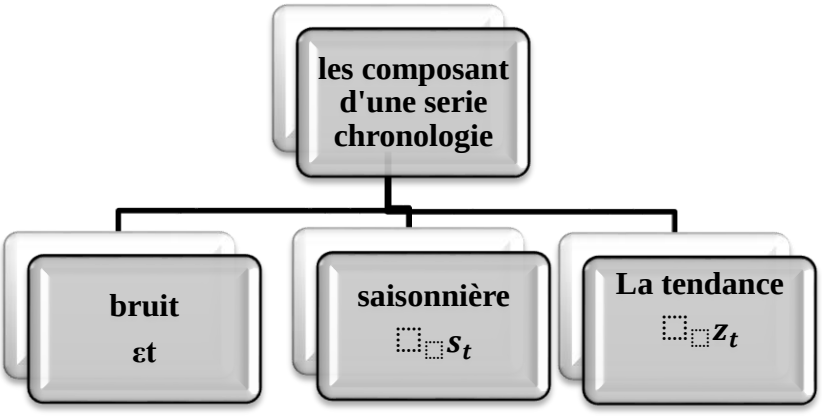


Figure 2.5 : Les Composantes d'une série chronologique

2.7 .Visualisation des séries chronologiques

La représentation visuelle des données revêt une grande importance dans l'analyse et l'exploration des séries chronologiques. On réalise cela en utilisant des graphiques qui représentent les valeurs observées sur l'axe des y en fonction d'un accroissement de temps sur l'axe des ***x*** (**voir la figure 2.6**). Ces graphiques mettent en évidence visuellement le comportement et les modèles des données**.**

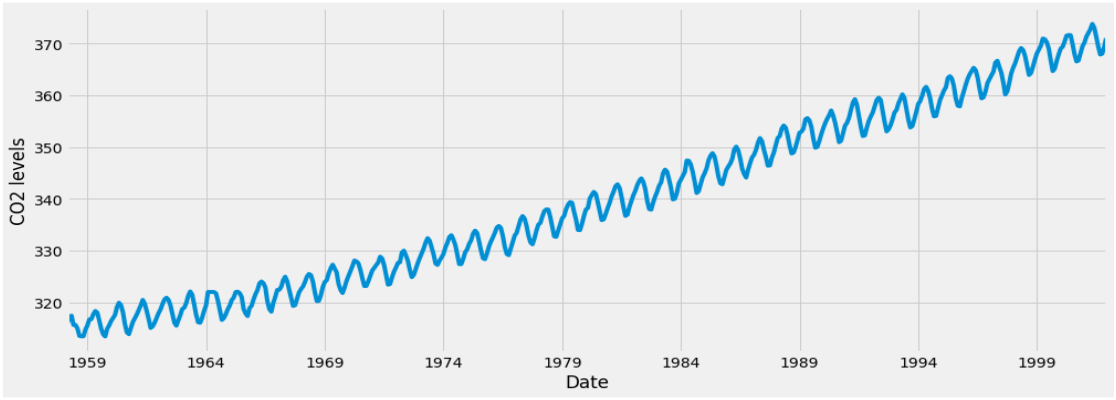


Figure 2.6 exemple de Graphe de d'une Serie chronologie de l'Atmosphère CO2 à partir d'échantillons d'air continus à l'observatoire de Mauna Loa, Hawaii, États-Unis

2.8 La stationnarité de la série chronologique

Nous aimerions mettre en évidence deux types de stationnarité.

- Stationnarité stricte : une série temporelle est dite strictement stationnaire si l'articulation la distribution de $\boldsymbol{X_{t_1}, X_{t_2}, X_{t_3}, \ldots, X_{t_{1n}}}$ est la même que la distribution conjointe de $\boldsymbol{X_{t_{1+k}}, X_{t_{1+k}}, X_{t_{2+k}}, \ldots X_{t_{n+k}}}$ pour tous les points de temps t_i et tous les décalages $\boldsymbol{k}$.

- **Définition** : Soit un processus temporel à valeurs réelles et en temps discret $X_1, X_2, \ldots \ldots \ldots \ldots X_t$, Il est dit stationnaire au sens faible (ou de second ordre) si
  - Moyenne arithmétique constante dans le temps

    $$\mathrm{E}(X_t) = \mu \text{ (ne depend pas de t) } \forall t = 1 \ldots T.$$

  - variance constant dans le temps

    $$\mathrm{Var}(X_t)) = \delta \neq \infty \; \forall \mathrm{i} = 1 \ldots \mathrm{t}$$

  - Aussi, la valeur de la covariance entre deux périodes de temps ne dépend que du décalage temporel entre les deux périodes et non du temps réel :

    $cov(X_t, X_{T-K}) = p_k$ (Ne depond pas de t) $\forall i = 1 \ldots t$ **[21].**

2.9 Série chronologique non stationnaire

Une série chronologique non stationnaire est l'opposé d'une série stationnaire, en ce sens que la moyenne des séries non stationnaire change continuellement avec le temps, soit par augmentation, soit par diminution, et qu'elle est également affectée par des variables saisonnières **[21].**

2.10 . Tests de stationnarité

2.10.1 Test de Dickey Fuller simple : sont les premiers à fournir un ensemble d'outils staistiques formels pour détecter la présence d une racine unitaire dans un processus autorégressif du premier ordre, ce test permet de tester l hypothèse : **[22].**

$$\begin{cases} H_0 : Le\ modeles\ est\ de\ racine\ unitaire \\ H_1 : Le\ model\ n'estpas\ de\ racine\ unitaire \end{cases}$$

Ce test est regroupé en 4 cas :

$$y_t = py_{t-1} + \varepsilon_t \quad H_0 : p = 1$$

$$y_t = \propto + py_{t-1} + \varepsilon_t \quad H_0 : \propto = 0 \text{ et } p = 1$$

$$y_t = \propto + py_{t-1} + \varepsilon_t \quad H_0 : \propto \neq 0 \text{ et } p = 1$$

$$y_t = \propto + \beta y_{t-1} + \varepsilon_t \quad H_0 : \propto = 0 \text{ et } \beta = 0 \text{ et } p = 1$$

Test de Dickey Fuller augmenté (**ADF)** :

Le test de Dickey-Fuller permet de savoir si une série est stationnaire ou non et permet aussi de déterminer la bonne manière de stationnariser la série.

Les hypothèses du test sont les suivantes

Cette procédure de test à des processus autorégressifs d'ordre **p**, il sagit alors des tests ***ADF*** : Ce test permet de tester **[23]**.

$$\begin{cases} H_0 : Le\ modeles\ est\ de\ racine\ unitaire \\ H_1 : Le\ model\ n'estpas\ de\ racine\ unitaire \end{cases}$$

Ce test peut être regroupé en 4 cas :

$$y_t = py_{t-1} + \sum_{i=1}^{p} \propto_i \Delta y_{t-i} + \varepsilon_t \qquad H_0 : p = 1$$

$$y_t = \propto + py_{t-1} + \sum_{i=1}^{p} \propto_i y_{t-i} + \varepsilon_t \qquad H_0 : \propto = 0 \text{ et } p = 1$$

$$y_t = \propto + py_{t-1} + \sum_{i=1}^{p} \propto_i y_{t-i} + \varepsilon_t \qquad H_0 : \propto \neq 0 \text{ et } p = 1$$

$$y_t = \propto + \beta t + py_{t-1} + \sum_{i=1}^{p} \propto_i y_{t-i} + \varepsilon_t \qquad H_0 : \propto = 0 \text{ et } \beta = 0 \text{ et } p = 1$$

2.6 Processus différentiel

Le processus différentiel est utilisé pour convertir des données non stationnaires en données stationnaires (exigence pour la mise en œuvre d'ARIMA). Notation différentielle **1**, définie comme suit :

$$\grave{X} = X_t - X_{t-1}$$

Pendant ce temps, si la première différentielle n'a pas rendu les données stationnaires, alors la notation différentielle **2** peut être utilisée, qui est définie comme suit : **[24]**

$$X'' = (\mathrm{X_t} - \mathrm{X_{t-1}}) - (\mathrm{X_{t-1}} - \mathrm{X_{t-2}})$$

2.7 La Fonction d'auto-covariance

Elle mesure la covariance entre deux valeurs séparées par un certain délai k (retard), elle fournit des informations sur la variabilité de la série et sur les liaisons temporelles qui existe entre les différentes composantes de la série Un processus est caractérisé, entre autres, par sa structure d'ordre deux.

On appelle fonction d'autocovariance la fonction **γ** définie de ***Z*** dans ***R*** par : **[25]**

$$\gamma_K = COV(X_{t,}X_{T+K}) \text{ .}\forall \mathrm{k, t} \in \mathrm{Z},$$

2.8 La fonction d'autocorrélation (ACF)

La fonction d'autocorrélation, notée FAC, est constituée par l'ensemble des autocorrélations de la série calculé pour des décalages d'ordre **k**,

$$p_k = corr(y_{t,}y_{t-k,})$$

On utilise l"autocorrélation pour caractériser les dépendances de linéaires dans des séries résiduelles (des séries temporelles corrigées de la tendance et la saisonalite). L'autocorrélation décrit la dépendance moyenne entre les valeurs d'une même série mais décalées d'un pas de temps k**.**
Elle se définit comme l'espérance mathématique du produit d'une série dont les éléments sont Séparés par un pas de temps ***k*** **[26].**

2.9 La fonction d'autocorrélation partielle (PACF)

La fonction d'autocorrélation partielle, notée FAP, est constituée par l'ensemble des autocorrélations partielles, le coefficient d'autocorrélation partielle mesurant la corrélation entre les variables (*Yt* et Yt-1), ignorant l'indépendance, est donc$\mathrm{X_t}$ considéré comme une constante,

$$\mathrm{X_{t=}x_t}, \qquad t = t+1, t+2, t+k-1$$

$$p_{kk} = corr\left(y_{t,}y_{t-k} \,|\mathrm{X_{t-1}} - \mathrm{X_{t-2}}, \ldots \mathrm{X_{t-k+1}}\right) \textbf{ [27]}$$

2.15. Modèles de prévision des séries chronologiques

Il existe plusieurs modèles linéaires couramment utilisés pour l'analyse des séries temporelles, tels que le modèle autorégressif (***AR***) et le modèle à moyenne mobile (***MA***). Ensuite, nous abordons le modèle ARIMA, qui combine ces deux modèles, ainsi que le modèle ARIMA. De plus, nous discutons d'une méthode de prévision des séries temporelles et détaillons les étapes de ce modèle**.**

2.15.1. Modèles d'autocorrélation d'un *AR(p)*

Modèles (***Xt***) est dit autorégressif d'ordre p noté ***AR***(p), si Xt est générée par une moyenne pondérée des observations passées jusqu'a la p-iéme période. Le processus est d´efini comme suit :

$$X_t = \sum_{i=1}^{p} \propto_t X_{t-i} + \varepsilon_t$$

Les $\propto_{\mathbf{t}}$ sont des paramétres réels à estimer généralement**.[28]**

2.15.2. Modèle moyenne mobile MA

Modèles ($\boldsymbol{X_t}$) de moyenne mobile (Moving Average) d'ordre q noté MA(q), chaque observation $x_t$est générée par une moyenne pondérée d'aléas jusqu' a la q-iéme période dans le passé.

$$X_t = \sum_{i=1}^{q} \theta_t \epsilon_{t-i} + \epsilon_t$$

2.15.3. Modèles ARMA

Les modéles ARMA (***p, q***) (Autoregresive Moving Average) combine la partie ***AR*** et la partie ***MA*** .en autre termes il contient des valeurs passees $z_{t-1,}z_{t-2,}$ … … … … … z_{t-p} et des erreurs passees $e_{t-1,}e_{t-2,}$ … … … … … t_{t-q}

Equation de ARMA :

$$X_t = \sum_{i=1}^{p} \propto_t X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_t \epsilon_{t-i} + \epsilon_t$$

2.15.4. Les modèles ARIMA (p, d, q)

Un modèle autorégressif de moyenne mobile intégrée d'ordre ***p, q*** et d de a étant donné la série temporelle yt, est un processus ARMA (***p, q***) stationnaire de sa différence dth, elle est dite intégrée en raison de la capacité de reconstruire la série par sommation ou intégrer les différences. Le modèle est noté ARIMA (***p, d, q***) et peut être écrit comme

$$X_t = \sum_{i=1}^{p+q} \propto_t X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_t \epsilon_{t-i} + \epsilon_t$$

Où $\propto_{\boldsymbol{t}}$ résulte de $\boldsymbol{\theta_t}$ en appliquant la différenciation sur la série **[29].**

2.16. La méthodologie du modèle ARIMA : comporte essentiellement quatre étapes :

2.16.1. Identification

La phase d'identification est la plus importante et la plus difficile : elle consiste à déterminer le modèle adéquat, c'est-à-dire les valeurs des paramètres $\boldsymbol{p}$, $\boldsymbol{q}$ du modèle ARMA. Elle est fondée sur l'étude des corrélogrammes simple et partiel.

Cette étape consite à obtenir la stationnarité des données en interprétant le graphique des autocovariances.

Un processus non-stationnaire a ses autocovariances qui décroissent lentement à l'inverse d'un processus stationnaire **[29].**

2.16.2. Estimation du modèle

L'estimation des paramètres d'un modèle $ARMA\ (\boldsymbol{p};\boldsymbol{q})$ lorsque les ordres p et q sont supposés connus peut se réaliser par déférentes méthodes dans le domaine temporel : Nous allons présenter ici brièvement la démarche de l'estimation par le maximum de vraisemblance.

L'étape 2 repose également sur l'interprétation de graphiques pour choisir l'ordre des parties AR et MA. On sait que les autocorrelations d'un processus $MA(q)$ deviennent nulles à partir de l'ordre$q\ +1$. Si le graphique des autocorrelations empiriques chute brusquement après $\boldsymbol{h\ =q}$, on pourra donc dire que l'on est en présence d'un $MA(q)$. Si l'on considère maintenant les autocorrélations totales d'un $AR(p)$, on sait qu'elles décroissent lentement dans le temps.

Mais il n'est guère possible de déduire une valeur de $\boldsymbol{p}$ `a partir de l'examen du correlogramme. On cherche donc une transformation du correlogramme qui soit plus interprétable. Il s'agit du graphique des autocorrelations partielles. Les autocorrelations partielles ont la propriété d'être nulles à partir de l'ordre $\boldsymbol{p+1}$ pour un processus$AR(p)$ **[29].**

2.16.2.1. Critères de choix des modèles

Souvent il n'est pas facile de déterminer un modèle unique. Le modèle qui est finalement choisi et celui qui minimiser l'un des critères à partir T observations.

- **L'erreur absolue moyenne (Mean Absolute Error)** : Également connu sous le nom d'EAM, il mesure la différence absolue moyenne entre les valeurs prédites, il fournit une interprétation directe de l'amplitude de laerreurs car il indique l'inexactitude moyenne à laquelle on peut s'attendre. L'EAM est également dépendant de l'échelle, mais contrairement à l'EMS et à l'EQM, elle est moins sensible aux valeurs aberrantes.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\varepsilon_t|$$

- **L'erreur quadratique moyenne (Mean Squared Error) :**

L'erreur quadratique moyenne ou ***MSE*** est une mesure qui calcule la moyenne au carré entre les valeurs vraies et prédites, l'opération quadratique place une valeur.

Pénalité sur les erreurs plus importantes et est plus indulgent envers les plus petites, ce qui la rend sensible aux valeurs aberrantes.

$$MSE = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2$$

- **La racine carrée de l'erreur quadratique moyenne (Root Mean Square Error)** : La racine de l'erreur quadratique moyenne ou **RMSE** comme son nom l'indique est la racine carrée de MSE, il pénalise toujours les erreurs plus importantes tout en fournissant des informations à la même échelle que la variable en question

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2}$$

- **Ecart absolu moyen en pourcentage (Mean Absolute PercentError) :**

L'erreur absolue moyenne en pourcentage ou **MAPE** mesure la différence absolue moyenne entre chaque valeur réelle et prévisionnelle, par rapport à la valeur réelle. **MAPE** est un puisqu'il fournit une mesure basée sur un pourcentage de l'erreur commise il convient de comparer la précision de plusieurs cas d'utilisation et ensembles de données.

$$MAPE = \frac{100}{T}\sum_{t=1}^{T} \left|\frac{\varepsilon_t}{X_t}\right|$$

- **Critères d'information**

Deux critères importants dans les applications des séries temporelles, sont

- **AIC (Akaike Information Critèrion) :**

$$AIC = 2\ln V + 2k$$

k : le nombre de paramètres.

2k : représente la pénalité.

V : est la vaisemblance.

Le modèle à retenir est celui qui montre l'AIC le plus faible, l'AIC utilise le principe du maximum de vraisemblance. Il pénalise les modèles comportant trop de variables, et évite le sur-apprentissage.

- **BIC (Bayesain Information Criterion) :**

$$BIC = 2\ln V + k\ln(n)$$

K : le nombre de paramètres libres du modèle.

n : le nombre de données.

Ln (**n**) : le terme de pénalité.

Le BIC utilise le principe de la maximum vraisemblance. Il pénnalise les modèles comportant trop de variables, et évite le sur-apprentissage.

2.16.3. Validation du modèle

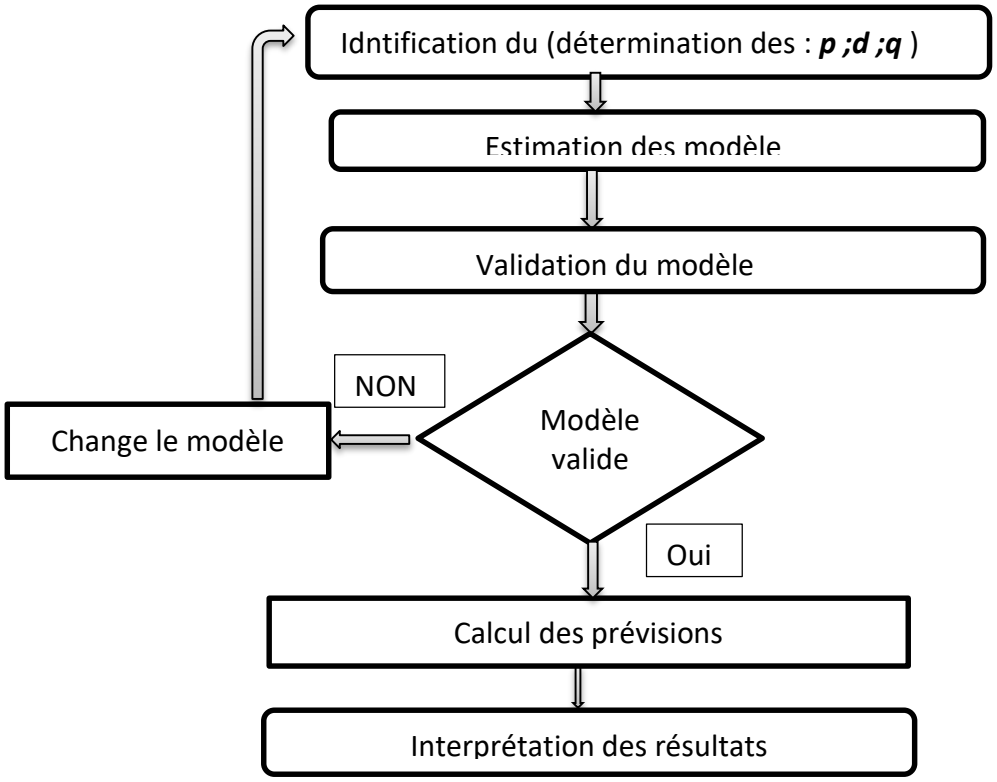
Il s'agit de vérifier notamment que les résidus du modèle $ARMA$ estimé, résidus notés $\widehat{\varepsilon_t}$, vérifient les propriétés requises pour que l'estimation soit valide, à savoir qu'ils suivent un processus BB, non autocorrélé et de méme variance, et qu'ils suivent une loi normale. Si ces hypothéses ne sont pas rejetées, on peut alors mener des tests sur les paramètres.

2.16.4. La prévision

Une fois qu'un modèle acceptable est trouvé pour la série chronologique à l'étude, les prévisions peuvent être calculées. On note $\hat{X}_{T+h}$ la prévision de X_{T+h} au temps $T+h$ où T est la taille de l'échantillon des observations X_T et h l'horizon de la prévision.

La prévision est calculée par la formule suivante : **[29]**

$$\hat{X}_{T+h} = E(X_{T+h}/X_T\ , X_{T-1}\ , \dots , X_1).$$



Figuire 2.7 : Les Etapes du modèl Arima

Conclusion

Nous avons exploré les concepts fondamentaux des séries chronologiques. ET les objectifs principaux de l'analyse des séries chronologiques, tels que la prévision, la détection de tendances et la modélisation de la saisonnalité.

Ce chapitre nous a fourni les bases nécessaires pour comprendre et analyser les séries chronologiques dans le contexte de notre mémoire sur la prévision de l'intensité des cyberattaques.

Chapitre 3 :

L’implemetation du système de prédiction

Introduction

Nous nous intéressons à la modélisation des séries chronologiques comprenant des accidents survenus au cours des années 2021, 2022 et 2023, recueillis à partir de la base de données Hackmageddon [29]. Nous utilisons le modèle ARIMA, qui est initialisé avec les données mensuelles. Ensuite, nous simulons l'opération en temps réel pour chaque jour : nous prévoyons le nombre d'attaques pour le lendemain, ajoutons le nombre enregistré d'attaques à l'ensemble de données et mettons à jour le modèle de prédiction.

Ce modèle de prévision utilise simplement la moyenne de la série comme valeur de prédiction. Bien que notre modèle soit plus robuste, ces résultats soutiennent l'idée qu'il existe une corrélation temporelle entre les cyberincidents, ce qui peut être utilisé pour prédire les tendances futures avec une certaine fiabilité. Le modèle ARIMA présente un léger retard, ce qui est attendu car ses prévisions sont basées sur des mesures récentes. Par conséquent, un mouvement drastique à la hausse ou à la baisse peut prendre quelques itérations pour se manifester réellement dans le modèle.

Une attention particulière a été accordée à la génération de modèles afin d'améliorer le temps de retard [30] et permettre des prévisions plus précises des pics de volume à court terme. Nous avons également effectué une analyse sur des sous-ensembles de l'ensemble de données, en nous intéressant à différents types d'attaques : déni de service (DOS), courrier électronique malveillant, URL malveillantes et attaques sur les services Internet en face-à-face (AOIFS).

3.1. Environnement d'éxecussion:

3.1.1. Google Colab :

Google Colab est un environnement de développement en ligne basé sur Jupyter Notebook, qui offre la possibilité d'écrire, d'exécuter et de partager du code Python. Il fournit un accès gratuit à des ressources de calcul puissantes, y compris des unités de traitement graphique (GPU) et des unités de traitement tensoriel (TPU) pour accélérer l'exécution des tâches d'apprentissage automatique et de calcul intensif [31].

3.1.2– Définition du langage Python en informatique :

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages.[32]

3.1.3 – Définition jupyter :

Jupyter se présente comme un outil extrêmement simple à mettre en œuvre qui vous permettra de transformer vos Jupyter Notebooks en applications web ou en Dashboard quasiment automatiquement.[33]

3.1.4. – Panda :

Pandas est un package Python open source qui est le plus largement utilisé pour la science et l'analyse des données [34].

3.1.5 – Numpy :

Le terme Numpy est en fait **l'abréviation de « Numerical Python »**. Il s'agit d'une bibliothèque Open Source en langage Python. On utilise cet outil pour la programmation scientifique en Python,

et notamment pour la programmation en Data Science, pour l'ingénierie, les mathématiques ou la science .[35]

3.1.6. – Scikit learn :

Scikit-learn est une bibliothèque en Python qui offre de nombreux algorithmes d'apprentissage supervisé et non supervisé. Elle repose sur des technologies que vous connaissez peut-être déjà, telles que NumPy, pandas et Matplotlib.

Les fonctionnalités fournies par scikit-learn comprennent :

- Régression, compris la régression linéaire et logistique.
- Classification, compris les voisins les plus proches (K-Nearest Neighbors).
- Sélection de modèles.
- Prétraitement, compris la normalisation Min-Max.

Scikit-learn est une puissante bibliothèque qui facilite l'implémentation de diverses techniques d'apprentissage automatique dans vos projets Python.[36]

3.2. Déscription de dataset :

Hackmageddon rassemble des données publiques sur les cyberattaques. Bien que des recherches et des investigations soient menées sur les attaques signalées, il est probable qu'il y ait des incidents non signalés ou signalés avec des informations incorrectes, comme la date de l'attaque. Hackmageddon ne prétend pas être exhaustif, mais fournit un échantillon d'attaques classées chronologiquement et provenant de sources publiques telles que des blogs et des sites d'information.

Hackmageddon est un site web qui enregistre les événements de cyberattaques depuis 2011 et dispose d'une base de données volumineuse. Le site publie une liste des attaques toutes les deux semaines, stockées dans un tableau. La base de données contient des informations telles que la date, l'auteur, la cible, la description de l'attaque, la classe cible et le pays concerné. Cependant, comme prévu, de nombreuses données manquaient. Étant donné que cette base de données concerne des cyberattaques où le motif ou l'identité de l'attaquant est parfois inconnu, cela était à prévoir.

Le cyberespionnage vise à obtenir un accès non autorisé à des informations confidentielles, généralement détenues par un gouvernement ou une autre organisation.

Dans la cyberguerre, l'objectif est de perturber les activités d'un État ou d'une organisation, en attaquant délibérément des systèmes d'information à des fins stratégiques ou militaires.

L'hacktivisme, quant à lui, est plus général. Il se définit comme la pratique consistant à obtenir un accès non autorisé à un système informatique et à mener diverses actions perturbatrices dans le but d'atteindre des objectifs politiques ou sociaux.

Notre ensemble de données de test comprend des incidents survenus en 2021, 2022 et 2023, avec 36 observations. Nous avons travaillé avec deux colonnes : la colonne temporelle, qui indique la date, et la colonne de la valeur à prévoir, c'est-à-dire les attaques. J'ai utilisé une méthode d'analyse automatisée pour examiner les catégories et les sous-classes du site web. Cependant, avant de pouvoir commencer à analyser les données, il a été nécessaire de les nettoyer. J'ai utilisé la bibliothèque Pandas de Python pour nettoyer le jeu de données.

- Le tableau suivant présente un échantillon du jeu de données Hackmageddon **[28].**

ID	Date	Date Discovered	Author	Target	Attack	Target Clas	Attack Class
1	01/04/2023	During October	APT41 AKA	Taiwanese media	Malware	Information	Cyber
2	01/04/2023	During Q4 2022	Ursnif	Multiple organizations	Malware	Multiple	Cyber Crime
3	01/04/2023	During Q4 2022	Diceloader	Multiple organizations	Malware	Multiple	Cyber Crime
4	01/04/2023	Since at least	?	Multiple organizations in	Scam	Multiple	Cyber Crime
5	01/04/2023	-	9Near	Unnamed Thai	Unknown	Public admin	Hacktivism
6	01/04/2023	28/02/2023	D0nut Leaks	Montgomery General	Malware	Human health	Cyber Crime
7	02/04/2023	-	?	Alpi Aviation	Unknown	Manufacturin	Cyber Crime
8	03/04/2023	26/03/2023	?	Western Digital	Unknown	Manufacturin	Cyber Crime
9	03/04/2023	31/03/2023	?	Capita	Unknown	Professional,	Cyber Crime
10	03/04/2023	29/03/2023	North Korean state-	Multiple organizations in	Malware	Fintech	Cyber Crime
11	03/04/2023	Since October 2022	ALPHV AKA BlackCat	Multiple organizations	Malware	Multiple	Cyber Crime

- **Echantion de la serie chronologie**

Date	#Attaque
2021-01-01	194
2021-02-01	249
2021-03-01	279
2021-04-01	242
2021-05-01	177

3.3. Etude de la stationnarité

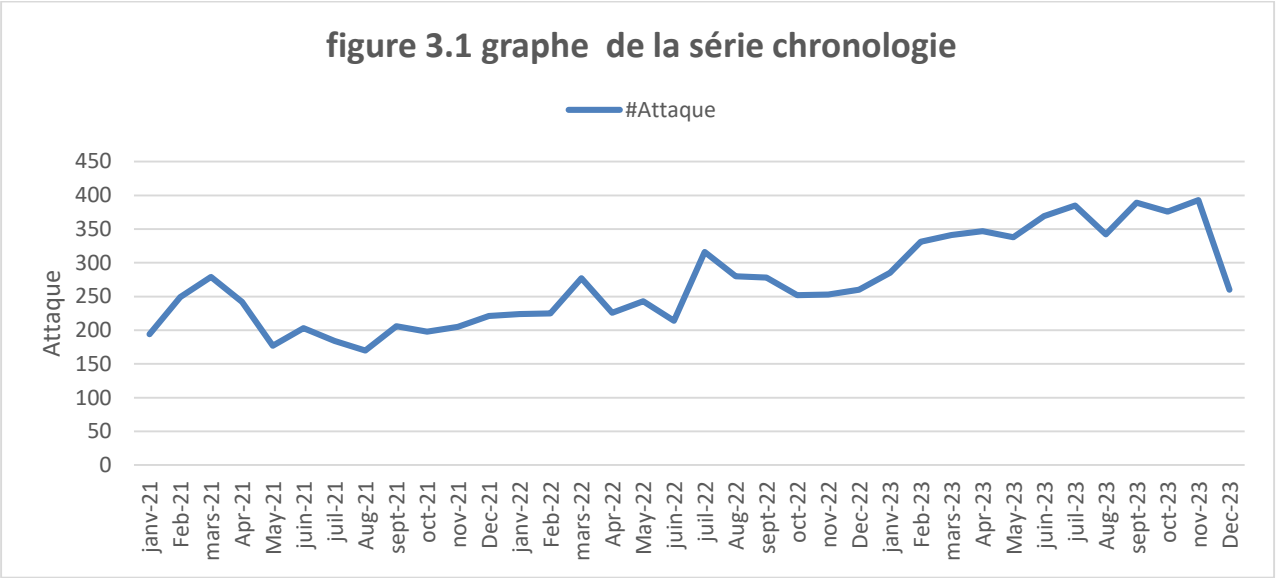
La stationarité : est un concept pour la série temporelle où les paramètres statistiques tels que la moyenne, la variance, etc sont tous constants au fil du temps.

Si la série de temps n'est pas stationnaire, alors les prédictions dévient des valeurs originales et augmentent l'erreur car nous ne connaissons pas les changements de ces paramètres statistiques car ils sont une fonction du temps.

3.4. Etude graphique de la série

La première étape de l'étude d'une série chronologique est la représentation graphique des données des ATTAQUES pendant les années 2021, 2022 et 2023. Cette visualisation donne des indications

très précieuses pour choisir un modèle. Le graphique présentant les "**Monthly Attacks**" (Attaques mensuelles) a été calculé à l'aide du logiciel **Python.**



Le graphique de la figure ***3-1*** de la série mensuelle sur les attaques montre une tendance aléatoire à long terme ainsi qu'une marche aléatoire reflétant la saisonnalité. Ce graphique montre une grande volatilité de L'attaque Mensuelle Sur Le Réseau. Ceci Est Confirmé La Serie Non Stationné.

3.5. TEST Écart-Type Et Moyen Mobile

L'écart type est calculé sur un ensemble de valeurs consécutives dans la série chronologique.
Cet ensemble se déplace le long de la série chronologique, montrant comment la variation et la dispersion changent au fil du temps.
Cette analyse aide à détecter les changements dans la dynamique des données ainsi que les fluctuations saisonnières ou cycliques

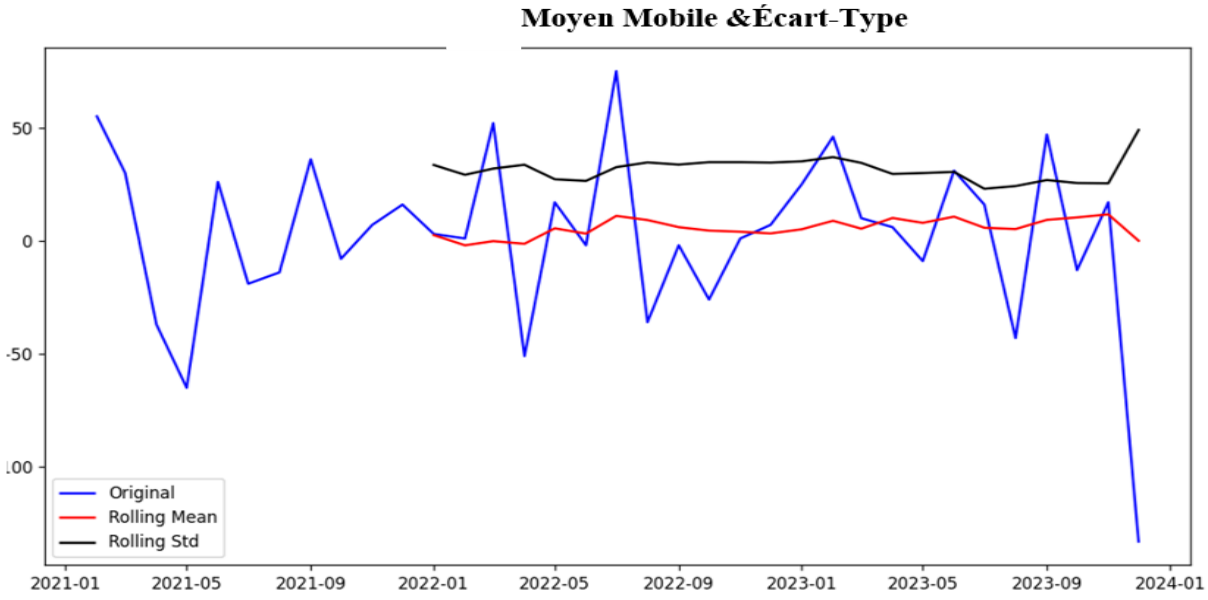


Figure 3.2.graphe d'Écart-Type Et Moyen Mobile

Ce graphique montre la moyenne mobile et l'écart-type mobile des données de la série chronologique.

En analysant la relation entre les données originales, la moyenne mobile et l'écart-type mobile, il est possible de mieux comprendre le comportement global et la dynamique de la série chronologique.

Résultats :

- Les données originales montrent des fluctuations importantes au cours de la période affichée.
- La moyenne mobile permet de voir la direction générale des données, lissant les fluctuations à court terme.
- L'écart-type mobile montre que le degré de variation des données a changé au fil du temps.

Pour verifier la stationnalité, On applique le test de Dickey-fuller (test racine unitaire) :

3.6. Test Augmented Dickey Fuller

Hypothèse nulle : Il suppose que la série temporelle est non stationnaire.

Hypothèse alternative : Si l'hypothèse nulle est rejetée, alors la série temporelle est stationnaire.

La sortie du test de Dickey-Fuller augmenté comprend :

- Valeur du test statistique
- Valeur de p
- Nombre de retards (#Lags)
- Nombre d'observations utilisées
- Valeurs critiques à 1%, 5% et 10%

Pour que l'hypothèse nulle soit rejetée et que l'on puisse accepter que la série temporelle est stationnaire, il y a 2 exigences :

1. Valeur critique (5%) > Valeur statistique du test
2. Valeur de p < 0,05

D'après le Tableau 3.1 de Dickey Fuller

Test Statistic	-1.36
p-value	**0.60**
#Lags Used	1.00
Number of Observations Used	34.00
Critical Value (1%)	-3.64
Critical Value (5%)	-2.95
Critical Value (10%)	-2.61
dtype: float64	

Le tableau « 1 » montre que :

La valeur du t-statistic calculé (**-1.36**) est supérieure à la valeur statistique de (**-2.95**).

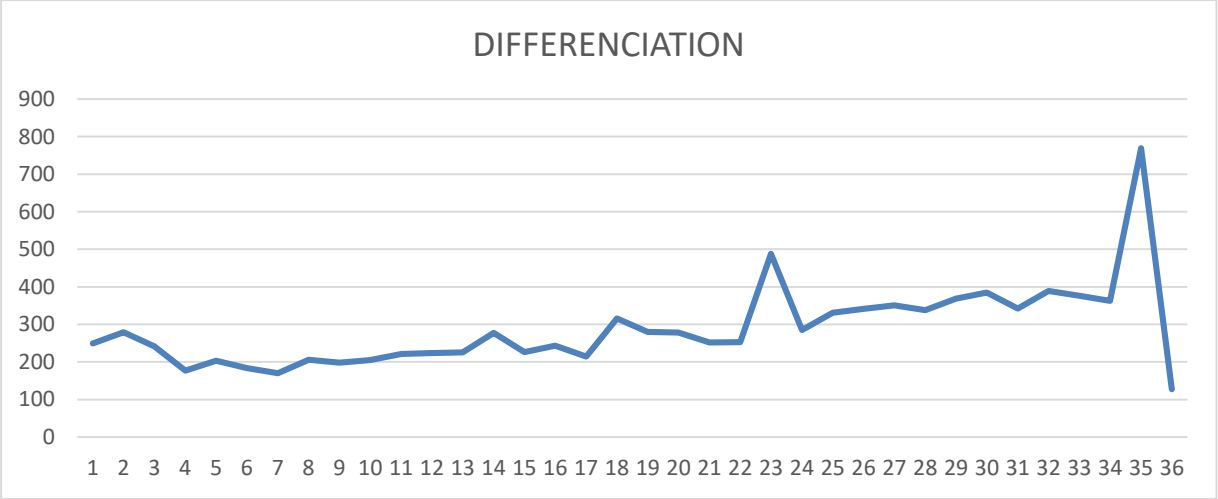
De plus, la valeur de p (**p = 0.60**) est supérieure à **0.05**. Cela signifie que notre série temporelle n'est pas stationnaire.

Par conséquent, nous devons appliquer une différenciation pour rendre la série stationnaire.

3.7. DIFFERENCIATION LA SERIE

Calculez les différences entre les valeurs consécutives de la série chronologique. Cela signifie que la différence entre la valeur de chaque point de la série et la valeur précédente sera calculée. L'objectif est de convertir la série d'origine en une nouvelle série qui contient les différences entre les valeurs au lieu des valeurs réelles. La nouvelle série s'appelle "data_diff" (figure 3.3).

Figure 3.3 Graphe De La Série Différencé.



Pour verifier la stationnalité, On applique le test de Dickey-fuller (test racine unitaire) :

D'après le Tableau 3.2.de Dickey Fuller

Results of Dickey-Fuller Test:	
Test Statistic	-3.882154
p-value	0.002170
#Lags Used	2.000000
Number of Observations Used	32.000000
Critical Value (1%)	-3.653520
Critical Value (5%)	-2.957219
Critical Value (10%)	-2.617588
dtype : float64	

Tableau 3.2. Au-dessus montre que

La valeur de t-statistic calculé (**-3.882154**) inferieur de Valeur statistique de (**-2.957219**).

Et (P= **0.002170** < 0.05 On peut dire que notre série est stationnaire.

3.8. MODELISATION PAR ARIMA Moyenne mobile intégrée auto-régressive

Le modèle ARIMA est une combinaison de **3** modèles :

AR (p) : Auto régressif

I (d) : Intégré

MA (q) : Moyenne mobile

(p, d, q) est connu sous le nom d'ordre du modèle ARIMA. Les valeurs de ces paramètres sont basées sur les modèles mentionnés ci-dessus.

p : Nombre de termes autorégressifs.

d : Nombre d'ordres de différenciation nécessaires pour rendre la série chronologique stationnaire.

q : Nombre d'erreurs de prévision décalées dans l'équation de prédiction.

Critères de sélection pour la commande du modèle ARIMA :

p : Valeur de décalage où le graphique d'autocorrélation partielle (PACF) se coupe ou tombe à **0** pour la 1ère instance.

d : Nombre de fois que la différenciation est effectuée pour rendre les séries temporelles stationnaires.

q : Valeur de décalage où le graphique d'autocorrélation (ACF) traverse l'intervalle de confiance supérieur pour la 1ère instance.

Le graphique ACF fournit la corrélation entre la série chronologie et ses décalages. Pour les séries chronologie ci-dessus, nous pouvons observer une corrélation positive à la baisse.

- Le graphique PACF fournit la corrélation entre la série chronologique et les décalages individuels. Ces coefficients de corrélation sont différents des corrélations mutuelles qui sont calculées en présence d'autres caractéristiques.

D'après le graphique PACF ci-dessus, le **1**er décalage est hors de l'intervalle de confiance et probablement le décalage le plus important. Cela dicte probablement le modèle pour le graphique ACF où le décalage suivant suit son décalage précédent.

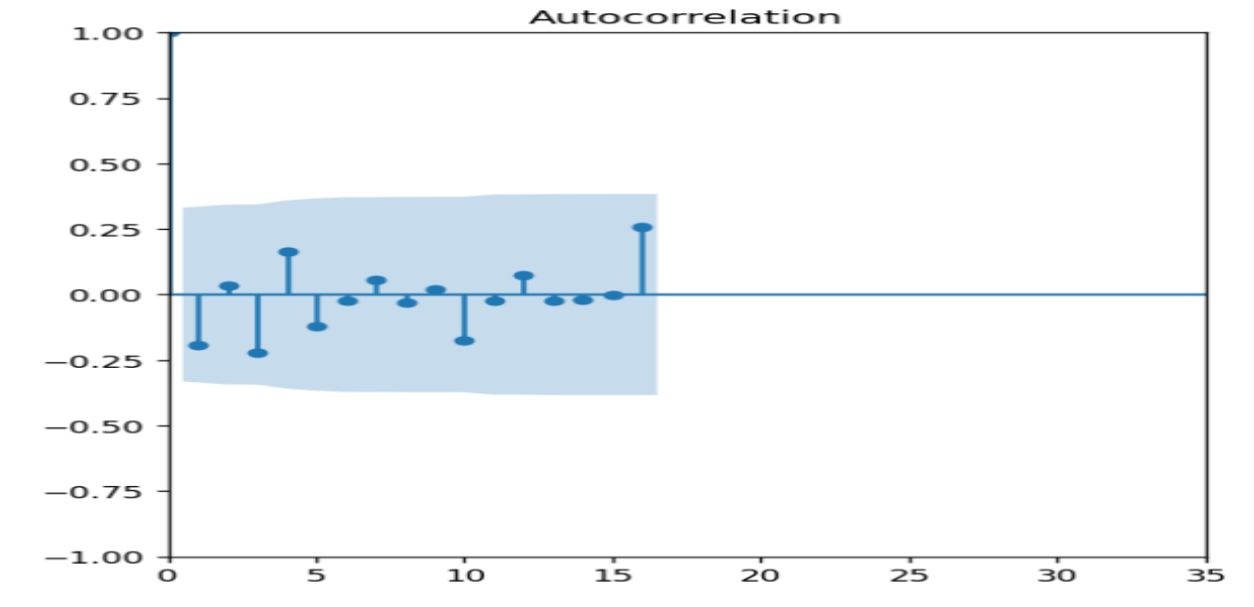


Figure 3.4 graphe de fonction d'autocorrelation

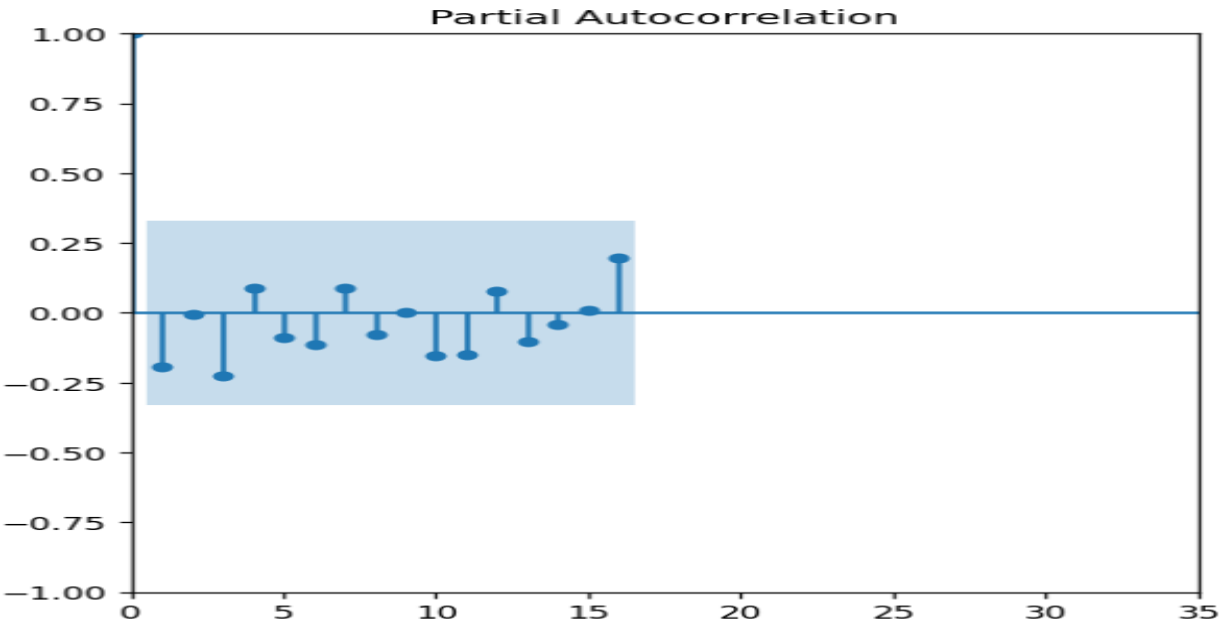


Figure 3.5 graphe de fonction d'autocorrelation partielle

3.9. Sélection des paramètres et chois du model ARIMA

Nous avons utilisé la bibliothèque **pmdarima** pour déterminer les parametre utilisées dans le modèle (par exemple, P, D et Q pour le modèle ARIMA) et la valeur standard utilisée pour choisir le meilleur modèle (comme AIC ou BIC). UN extrait des résultats du modèle ARIMA spécifié.

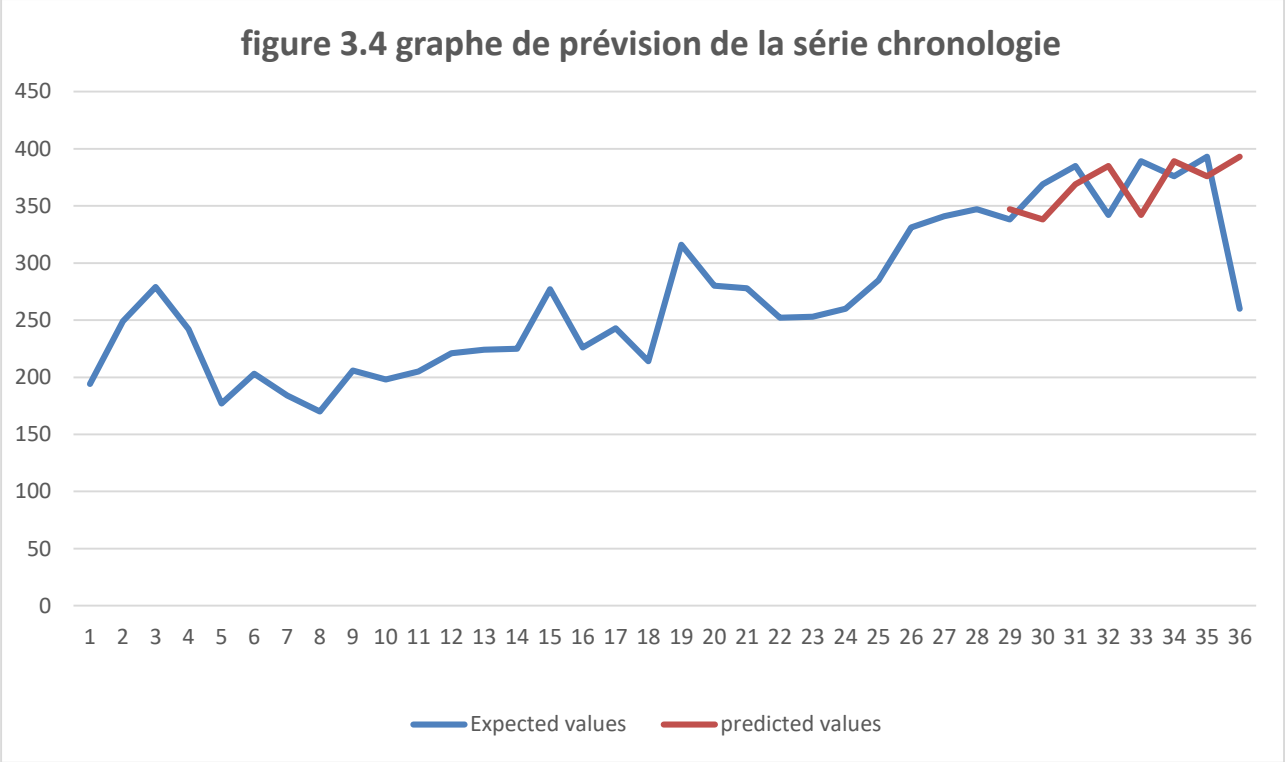
Table 3.3 Resultat de chois du model arima

```
  ARIMAX Results
==============================================================================
Dep. Variable :                        y   No. Observations :
28
Model :               ARIMAX (0, 1, 0)   Log Likelihood                -
132.321
Date :                Sun, 09 Jun 2024   AIC
266.642
Time :                        21:16:50   BIC
267.938
Sample :                    01-01-2021   HQIC
267.027
                          - 04-01-2023
Covariance Type :                   opg
==============================================================================
                 Coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--
sigma2      1057.4794    302.435      3.497      0.000     464.717
1650.241
==============================================================================
=======
Ljung-Box (L1) (Q) :                   0.84   Jarque-Bera (JB) :
0.05
Prob(Q) :                              0.36   Prob(JB) :
0.98
Heteroskedasticity (H) :               0.40   Skew :
-0.07
Prob(H) (two-sided) :                  0.19   Kurtosis :
2.85
==============================================================================
```

3.10. Prévision en échantillon

Le modèle prévoit les valeurs pour les points de données existants de la série temporelle. Il est similaire au format d'entraînement pour les problèmes de régression ou de classification.

- Nous divisons les données en ensembles de données d'entraînement et de test. Nous réservons 20 % pour l'ensemble de données de test et le reste pour le groupe de données d'entraînement.
- Pour cette prévision d'échantillon, nous utilisons la méthode de prévision par fenêtre glissante, c'est-à-dire que nous prévoyons une seule valeur à la fois et utilisons cette valeur prévue pour alimenter le modèle et prédire la prochaine valeur.



3.11. Mesures d'exactitude pour la prévision de la série temporelle

Ces métriques sont utilisées pour évaluer les performances du modèle ARIMA dans les prévisions, où le modèle vise à prédire les valeurs futures d'une série chronologique. Ces métriques peuvent être utilisées pour comparer les performances d'un modèle ARIMA avec celles d'autres modèles, ou pour évaluer les performances du même modèle ARIMA au fil du temps lorsqu'il est utilisé pour déterminer des valeurs futures.

	RMSE	**MAPE**
ARIMA MODEL	**54.20**	**0.123**

Table 3.4 les métriques de prévision de la série temporelle

Sur la base des résultats imprimés pour les mesures de précision, nous pouvons comprendre certaines informations sur les performances du modèle.

RMSE (Root Mean Square Deviation): La valeur mentionnée est **54,20.** Cette mesure mesure l'écart moyen entre les prédictions et les valeurs réelles, et plus la valeur est élevée, plus les différences entre les prédictions et les valeurs réelles sont importantes. Dans Ce cas, une valeur de **54, 20** indique qu'il existe des différences moyennes entre les valeurs prévues et les valeurs réelles.

MAPE (Mean Absolute Percentile Deviation): La valeur mentionnée est de **0,123** ou **12,3 %.** MAPE reflète la moyenne relative de l'écart entre les prédictions et les valeurs des verbes.

3.12. Desription De CODE MISE A JOUR ARIMA

```
# Forcasting future values
ARIMA_history_p = [x for x in train]
F2 = [ ]
for t in range (len (df1)) :
    model = ARIMA (ARIMA_history_p, order=auto_model.order)
    model_fit = model.fit ( )
    output = model_fit.predict (start=len(ARIMA_history_p), end=len(ARIMA_history_p), typ='levels')[0]
    ARIMA_history_p.append (output)
    f2.append(output)
  for i in range(len(f2)) :
    forecast.iloc[len(train) + i, forecast.columns.get_loc('ARIMA_Predict_Function')] = f2[i]
  forecast[['Attacks', 'ARIMA_Predict_Function']].plot(figsize=(12, 8))
```

- **Description**

Le code crée une liste appelée "ARIMA_history_p" et la remplit avec les valeurs de la liste d'entraînement. On suppose que la liste d'entraînement contient des données historiques d'une série chronologique qui seront utilisées pour entraîner le modèle ARIMA.

Ensuite, une liste vide appelée "f2" est créée pour stocker les valeurs prédites par le modèle ARIMA.

La boucle "for" itère sur la longueur de la série de données "df1", qui est supposée contenir les valeurs que vous souhaitez prédire.

À chaque itération, un modèle de la classe ARIMA est créé en utilisant la liste "ARIMA_history_p" et l'ordre du modèle "auto_model.order" (qui a été déterminé automatiquement à l'aide de pmdarima). "ARIMA_history_p" représente les données historiques collectées jusqu'à présent.

Le modèle est ensuite ajusté ("fit") en utilisant les données historiques "ARIMA_history_p".

Ensuite, la ligne suivante prédit une valeur future unique en utilisant le modèle entraîné : output = model_fit.predict(start=len(ARIMA_history_p), end=len(ARIMA_history_p), typ='levels')[0]

La méthode "predict" est utilisée pour effectuer une prévision sur une période spécifique, définie par "start" (début) et "end" (fin). Ici, nous prévoyons une seule valeur future. Le paramètre "typ='levels'" indique que les valeurs prédites seront dans le même format que les données d'origine.

La valeur prédite "output" est ensuite ajoutée à la liste "ARIMA_history_p" pour être utilisée dans les prévisions futures.

Enfin, la valeur prédite est également ajoutée à la liste "f2" pour enregistrer les prédictions.

En résumé, le code construit et ajuste un modèle ARIMA à l'aide des données historiques, puis utilise ce modèle pour prédire une valeur future à chaque itération de la boucle. Les valeurs prédites sont stockées dans les listes "ARIMA_history_p" et "f2" pour une utilisation ultérieure.

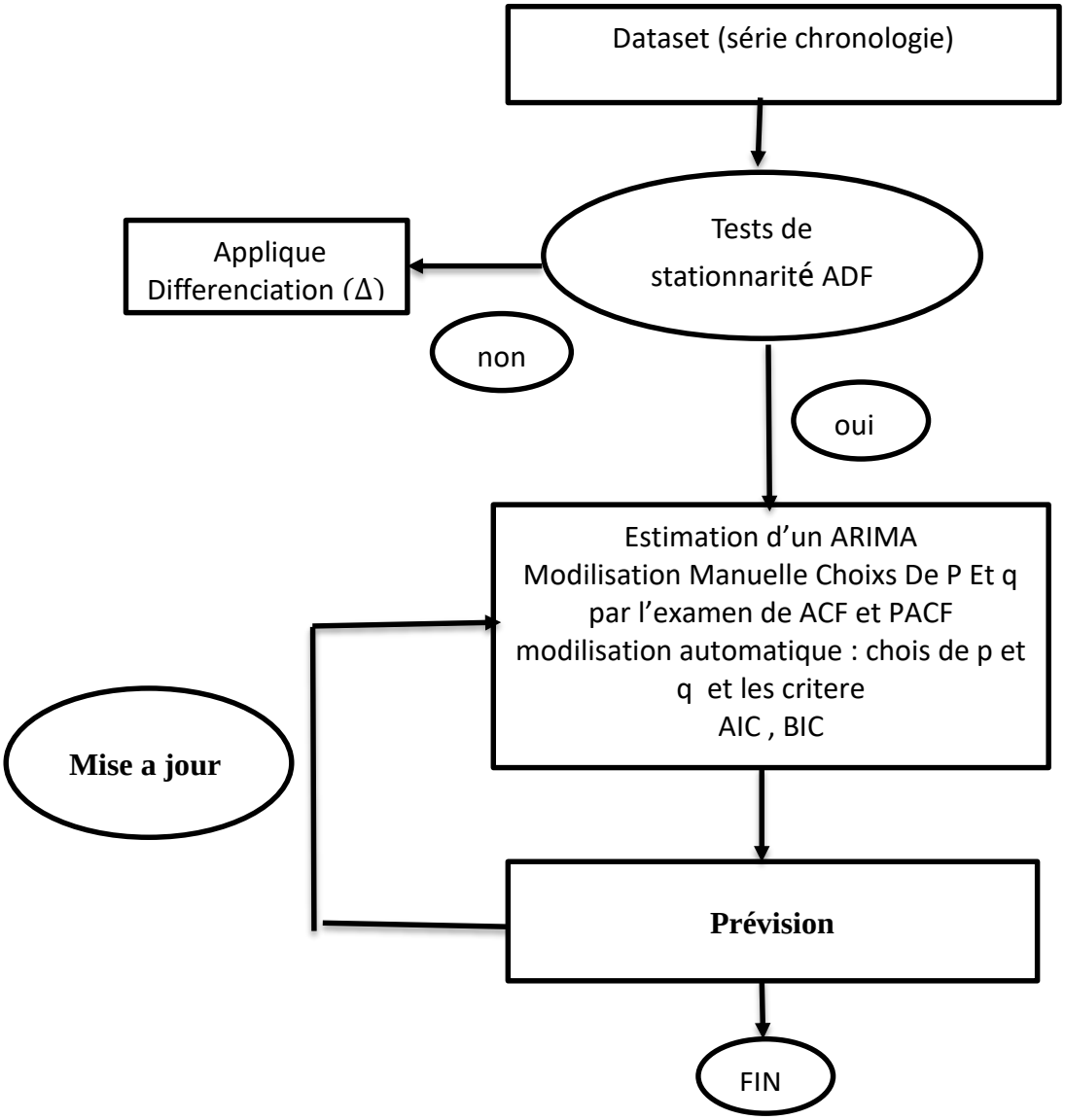


Figure 3-5 modèles de prévision

Conclusion

Nous avons présenté les résultats de la prévision de l'intensité des cyberattaques en utilisant le modèle ARIMA. Et analysé les performances de ce modèle à l'aide des mesures d'évaluation discutées précédemment. Les prévisions réalisées sur nos données de séries chronologiques ont été comparées aux valeurs réelles afin d'évaluer la précision du modèle.

De plus, nous avons développé un algorithme de mise à jour du modèle ARIMA ,Cet algorithme vise à ajuster dynamiquement les paramètres du modèle en fonction de l'évolution des données sur les cyberattaques.

Conclusion générale

Les résultats expérimentaux montrent que l'analyse de la fréquence des cyberincidents peut être utile pour prédire l'ampleur future de cyberaccidents. Notre système était capable de construire un modèle de précision basé uniquement sur le nombre d'incidents informatiques qui ont suivi l'intensité générale des outils informatiques au fil du temps. Traiter le nombre d'incidents cybernétiques par jour apparaît comme une chaîne de temps fortement liée, ce qui en fait un candidat prometteur pour l'analyse des prévisions.

En continuant à améliorer notre système, nous nous attendons à de meilleures prévisions pour les événements futurs, avec une précision plus fine. Alors que les données utilisées dans Hackmageddon ont montré des résultats prometteurs, l'existence d'un ensemble de données plus puissant que les cyberincidents pourrait mettre en lumière un citoyen plus fort ou plus faible dans le système. Le modèle ARIMA utilisé est une méthode de prévision standard, et la prévision peut également être utilisée comme une classification de type système combinant plusieurs techniques d'analyse statistique pour fournir une prévision unique de la probabilité d'un accident dans l'espace informatique.

Bibliographie

[1] **S. Yang, H. Du, J. Holsopple**, **and M. Sudit**. 2014. A.ack Projection. In Cyber Defense and Situational Awareness, A. Ko., C. Wang, and R. Erbacher (Eds.). Springer International Publishing, Cham, 239–261. DOI:h.p://dx.doi.org/10.1007/978-3-319-11391-3 12.

[2] **E. Gandotra, D. Bansal, and S. Sofat**. 2015. Computational Techniques for Predicting Cyber .reats. In Intelligent Computing, Communication and Devices: Proceedings of ICCD 2014, Volume 1, L. Jain, S. Patnaik, and N. Ichalkaranje (Eds.). Springer India, New Delhi, 247–253. DOI:h.p://dx.doi.org/10.1007/978-81-322-2012-1 26

[3]. **R. Yende**, « Cours de Sécurité Informatique & Crypto », support de cours, Congo- Kinshasa., vol. 139, 2018.

[4] 30 mars 2021 · Book PDF Available. Sécurité Informatique - Cours et TD. March 2021. Publisher: Université de Guelma. Authors: Mohamed Amine Ferrag. Université 8 mai 1945 …

[5] **I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury**, "The dual-tree complex wavelet transform,". IEEE signal processing magazine, vol. 22, no. 6, pp. 123–151,2005.

[6] **D. Burgermeister and J. Krier**, Les systèmes de détection d'intrusions." Article Disponible sur http://dbprog. Developpez. com, 2006.

[7] **Cédric Llorens.** Mesure de la sécurité "logique" d'un réseau d'un opérateur de télécommunications.domain_other. Télécom ParisTech, 2005. English. NNT : . pastel-00001492

[8] **Bernard Cousin**. « Sécurité des réseaux informatiques », Rennes .1 ere édition, 2005, 203 pages

[9]**Trend Micro Devices,** Inc. Sécurité informatique pour les nuls: edition PME-PMI. Wiley, 2010.

[10]**Pretavoine, Nicolas**. Gestion de l'environnement pour les PME-PMI. AFNOR, 2007.

[11] **Y.-W. Chen, J.-P. Sheu, Y.-C. Kuo, and N. Van Cuong**, "Design and Implementation of iot ddos attacks detection system based on machine learning," in 2020 European Conference on Networks and Communications (EuCNC), pp. 122–127,IEEE, 2020.

[12]**J-F. Carpentier,** « La sécurité informatique dans la petite entreprise : état de l'art et bonnes pratique », Edition ENI, vol. 265, 2009

[13]**J-F. Carpentier**, « La sécurité informatique dans la petite entreprise : état de l'art et bonnes pratique », Edition ENI, vol. 265, 2009

[14]**M. Belaoued**, « Approches Collectives et Coopératives en Sécurité des Systèmes Informatiques », Thèse de doctorat, Université 20 Août 1955. Skikda, 2016

[15]**G. Charpentier et al**, « VIRUS / ANTIVIRUS Nouvelles technologies Réseaux», support de cours, enseignant : Etienne DURIS, vol. 49, 2004

[16]**R. Rhouma,** "Audit et Sécurité Informatique."https://sites.google.com/site/rhoouma/teaching-at-esen/cryptographie-et-securitede-l-information.

[17]**Joshi, P., Massaron, L., and Hearty**, J. Python: Real World Machine Learning. PacktPublishing,

[18] **LOUDJANI Nawal** « Développement d'algorithmes pour l'analyse des séries temporelles des données de productiond'eau potable » Thèse Doctorat, Université Mohamed Khider – Biskra

[19] **Brockwell, P.J., Davies, R.A**. (2009). Times Series: Theory and Methods, 2nd edition, Springer.

[20] **Keogh, E., Chu, S., Hart, D., & Pazzani,** M. (2004). Segmenting time series : A survey and novel approach. In Data mining in time series databases (pp. 1-21).

[21] **D.C. Montgomery, C.L. Jennings, and M. Kulahci**. Introduction to Time Series Analysis and Forecasting. Wiley Series in Probability and Statistics. Wiley, 2015.

[22]**CHATFIELD,C**.(2001), « Time Series Forecasting »,Chapman&Hall.

[23] **MELARD, G.** (1990a). Méthodes de prévision à court terme, Editions de l'Université de Bruxelles, et Editions Ellipses, Paris.

[24]Yakubu et al. / International Journal of Global Operations Research, Vol. 3, No. 3, pp. 80-85, 2022

[25] **FARAWAY, J. CHATFIELD, C**. "Time series forecasting with neural networks: acomparative study using the airline data", Applied Statistics 47 (1998), pages : 231–250

[26] GOURIEROUX, C.(1997).ARCH models and financial applications.Springer series in statistics.New York, Springer-Verlag.

[27]GOURIEROUX, C. & MONFORT, A. (1990). Séries temporelles et modèles dynamiques, Economica, Paris

[28] **Ruey S. Tsay**. Analysis of financial time series. Wiley series in probability and statistics. Wiley-Interscience, 2010.

[29] **Box, G. E. P., Jenkins,** G. M. (1976). Time Series Analysis Fore- casting and Control. Revised Edition.

[30] **P. Passeri**. 2016. Hackmageddon Cyber Incident Data. (2016). h.p://www.hackmageddon.com/category/security/cyber-a.acks-statistics/.

[31] «https://datascientest.com/google-colab-tout-savoir/ Accède le 09/06/2024»

[32] "Python." https://www.journaldunet.fr/web-tech/dictionnaire-duwebmastering/1445304- python-définition-et-utilisation-de-ce-langage-informatique/.

[33] https://jupyter.org/ Accéde le 25/06/2024 REFERENCE BIBLIOGRAPHIE 61

[34] "Pandas - Python Data Analysis Library." https://pandas.pydata.org/

[35] «https://numpy.org/ Accède le 02/06/2024 »

[36]«https://www.inria.fr/fr/lancement-de-linitiative-scikit-learn?fbclid=IwAR1r89W0NsQHyu7BN31qRQJq5YEUS0iORwj37i51Zj0ds35stAwHCL-8N8c Accède le 02/06/2024. »