Revue Française d'Economie et de Gestion
ISSN : 2728- 0128
Volume 2 : Numéro 12

Réseau LSTM et méthodologie de Box et Jenkins pour la prévision des séries temporelles : essai sur l'indice MASI de la Bourse de CASABLANCA

LSTM network and Box and Jenkins methodology for time series forecasting: test on MASI index of CASABLANCA Stock Exchange

BOUDRI Imane
Doctorante
Ecole Nationale de Commerce et de Gestion
Université Sidi Mohammed Ben Abdellah-Maroc
Laboratoire de Recherche et d'Etudes en Management, Entreprenariat et Finance
Imane.boudri@usmba.ac.ma

EL BOUHADI Abdelhamid
Enseignant chercheur
Ecole Nationale de Commerce et de Gestion
Université Sidi Mohammed Ben Abdellah-Maroc
Laboratoire de Recherche et d'Etudes en Management, Entreprenariat et Finance
Maroc
El_bouhadihamid@yahoo.fr

Date de soumission : 27/10/2021

Date d'acceptation : 28/11/2021

Pour citer cet article :

BOUDRI.I & EL BOUHADI.A (2021) «Réseau LSTM et méthodologie de Box et Jenkins pour la prévision des séries temporelles : essai sur l'indice MASI de la Bourse de CASABLANCA», Revue Française d'Economie et de Gestion «Volume 2 : Numéro 12» pp : 13 - 36

Résumé

L'objectif de cet article est de réaliser des prédictions d'une série chronologique en se basant sur un échantillon composé de 127 observations journalières de rendements de l'indice MASI de la bourse de Casablanca durant le premier semestre de l'an 2020, et ce, en utilisant deux approches complétement différentes : une approche statistique (méthode de Box et Jenkins) représentée par un processus ARMA, et une approche connexionniste basée sur l'auto-apprentissage par un réseau LSTM (*Long Short Term Memory*) dans le but d'évaluer la qualité des prédictions en les comparant avec les valeurs réelles observées. Les résultats obtenus nous ont permis de conclure que le réseau LSTM est plus performant en matière de prédiction de valeurs futurs d'une série temporelle.

Mots clés : Série temporelle ; Prévisions ; Méthode de Box et Jenkins ; Réseau de neurones artificiels ; Réseau LSTM.

Abstract

The purpose of this paper is to make predictions of a time series based on a sample of 127 daily observations of returns of the MASI index of the Casablanca Stock Exchange during the first half of the year 2020, and this, using two completely different approaches: a statistical approach (Box and Jenkins method) represented by an ARMA process, and a connectionist approach based on self-learning by a LSTM (Long Short Term Memory) network in order to evaluate the quality of predictions by comparing them with the real observed values. The results obtained allowed us to conclude that the LSTM network is more efficient in predicting future values of a time series.

Keywords: Time series; Predictions; Box and Jenkins method; Artificial neural network; LSTM network.

Introduction

La prédiction des séries chronologiques a fait l'objet d'un nombre considérable d'études en raison des quantités innombrables de données temporelles et séquentielles produites quotidiennement par l'industrie de l'information et les différentes structures de recherche (Chen & Chen, 2015 ; Hu, 2017 ; Rubio et al., 2017 ; Ben Houad & Oubouali, 2018). Ce domaine a connu une effervescence spectaculaire avec l'explosion des données numériques, de la *Big Data* et notamment de l'intelligence artificielle depuis le début des années 1990 (Stern, 1996 ; Paquet, 1997 ; Dunis & Williams, 2002 ; Valipour et al., 2013 ; Xiao et al., 2014 ; Sezer et al., 2017).

La Bourse de Casablanca, alors qu'elle s'apprêtait a regagné sa place dans la catégorie « pays émergents » de l'indice MSCI (*Morgan Stanley Country Index*) après avoir clôturé une année 2019 avec une tendance haussière et des anticipations optimistes, quant à l'amélioration des résultats comptables annuels des sociétés cotées, ce qui auraient boosté les performances des indices boursiers MASI et MADEX. S'est trouvé au premier trimestre 2020 -selon un rapport publié par le Haut-Commissariat au Plan- face à l'une des plus grandes chutes jamais réalisées au cours des deux dernières décennies, suite aux incertitudes qu'a provoqué la crise sanitaire. A la fin du premier trimestre 2020, le MASI a baissé de 11,1% après avoir enregistré une hausse de 7,1% à la fin de l'année 2019. Ceci ne nous empêche pas de faire un essai de prévision sur le marché boursier marocain.

La problématique de ce papier s'articule autour de la question suivante : les modèles de réseaux de neurones artificiels sont-ils plus performants en matière de prédiction des séries temporelles ?

Dans le but d'apporter des éléments de réponse à notre problématique, nous allons procéder à une modélisation du rendement de l'indice MASI en se basant sur les données du premier semestre de l'an 2020. Le but étant de réaliser des prédictions en mobilisant deux approches de deux familles différentes : la première approche est basée sur une modélisation statistique linéaire et uni-variée alors que la deuxième approche repose sur l'apprentissage automatique par réseau de neurones artificiel.

Notre papier commence par un rappel du cadre théorique des deux approches. Par la suite il entame le volet empirique en traçant la méthodologie poursuivie dans l'application de la méthode de Box et Jenkins et la mise en place du réseau LSTM ainsi que les résultats y afférents, et finalement il formule une discussion autour de la qualité des prévisions obtenues.

1. Cadre théorique

1.1. Les caractéristiques stochastiques des séries temporelles

Une série temporelle (ou chronologique) est une suite d'observations (variables aléatoires) dans un intervalle de temps discret ou continu : $\{x_1, x_2, \dots, x_n\}$. Avant de procéder à la modélisation d'une série chronologique, il s'impose d'en étudier les caractéristiques stochastiques. Si ces caractéristiques s'avèrent constantes dans le temps, on parle d'une série stationnaire. Dans le cas opposé, la série est dite non stationnaire. De ce fait, un processus stochastique de bruit blanc[1] est un processus stationnaire.

La stationnarité représente la propriété la plus importante des séries chronologiques. Une série est dite stationnaire si elle est le résultat d'un processus d'observations dont les caractéristiques sont stables dans le temps. Ceci implique que la série ne comporte ni tendance ni saisonnalité.

L'étude de la stationnarité peut dans certains cas être effectuée à partir de l'analyse des fonctions d'autocorrélation simple et partielle ou par leur représentation graphique appelée « corrélogramme » qui permet de déceler des corrélations internes entre la série et elle-même.

Cependant, en cas de non-stationnarité, cette analyse ne permet pas de préciser de quel processus non stationnaire s'agit-il : de type TS (*Trend Stationary*) pour les processus caractérisés par une non-stationnarité de type déterministe ou DS (*Difference Stationary*) pour les processus caractérisés par une non-stationnarité de type stochastique.

Pour remédier à cette insuffisance, Dickey et Fuller (1976) proposent des tests de racine unitaire (*Unit Root Test*) DF qui permettent non seulement de mette à l'épreuve le caractère stationnaire ou non d'une série temporelle (avec un processus purement autorégressif d'ordre 1), mais aussi de déterminer de quel processus non stationnaire il s'agit. Avec comme hypothèse nulle (*H0 : $\phi = 1$*) : la présence de racine unitaire soit la non stationnarité stochastique de la série et ce, en estimant trois modèles :

Le premier modèle est un modèle autorégressif d'ordre 1 :

$$(1) : \quad x_t = \emptyset_1 x_{t-1} + \varepsilon_t$$

Le second modèle est un modèle autorégressif avec constante :

$$(2) : \quad x_t = \phi_1 x_{t-1} + \beta + \varepsilon_t$$

Le troisième et dernier modèle est un modèle autorégressif avec constante et tendance :

$$(3) : \quad x_t = \phi_1 x_{t-1} + b_t + c + \varepsilon_t$$

[1] Un processus de bruit blanc et un processus de variables aléatoires de moyenne nulle et de variance constante non autocorrélée.

Par ailleurs, le test de Dickey et Fuller a été étendu au test de Dickey et Fuller augmenté (ADF) qui permet de détecter la présence de racine unitaire dans des processus autorégressifs d'ordre *p*. Ce test se base sur l'estimation par les MCO (méthode des moindres carrées ordinaire) sous l'hypothèse alternative $H_1 : |\phi| < 1$, des trois modèles :

$$(4): \quad \Delta x_t = \rho x_{t-1} - \sum_{j=2}^{p} \phi_j \Delta x_{t-j+1} + \varepsilon_t$$

$$(5): \quad \Delta x_t = \rho x_{t-1} - \sum_{j=2}^{p} \phi_j \Delta x_{t-j+1} + c + \varepsilon_t$$

$$(6): \quad \Delta x_t = \rho x_{t-1} - \sum_{j=2}^{p} \phi_j \Delta x_{t-j+1} + c + b_t + \varepsilon_t$$

Phillips et Perron (1989) propose un test de racine unitaire avec une approche utilisant une composante déterministe non linéaire tout en maintenant une tendance polynomiale sous les hypothèses nulle et alternative.

En 1992, Kwiatkowski *et al*. développent un test de multiplicateur de Lagrange (LM) où l'hypothèse nulle est celle de la stationnarité de la série contrairement aux autres tests.

Toutefois, le test ADF reste le test de racine unitaire le plus utilisés par les académiciens et les chercheurs, et c'est le test que nous allons utiliser dans la vérification du caractère stationnaire de notre série des rendements.

1.2. Les modèles ARIMA

1.2.1. Processus autorégressif d'ordre *p*

L'histoire de l'analyse prédictive courte des séries temporelles en fonction de leurs valeurs passées remonte aux année 1920 avec l'apparition des premiers modèles univariés pour une modélisation linéaire. George Udny Yule en 1927 fut le premier statisticien à avoir proposé le processus autorégressif AR (AutoRégressif) dans la modélisation d'une série chronologique des nombres de taches solaires dans son article « *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers* ».

Dans un processus autorégressif d'ordre *p*, une observation à l'instant *t* peut être exprimée comme une moyenne pondérée des observations passées plus un bruit blanc gaussien :

$$(7): \quad \text{AR}(p): \; X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \mathcal{E}_t$$

Avec, ϕ_k (k = 1, 2, 3, …, p) sont les paramètres du modèle et ε_t est une erreur gaussienne.

1.2.2. Processus à moyenne mobile d'ordre *q*

Les modèles ARMA (*AutoRegressive Moving Average*) consistent à combiner un processus autorégressif (moyenne pondérée des valeurs passées) et un processus à moyenne mobile (moyenne pondérée des erreurs passées). Herman Word (1954) propose la modélisation des séries stationnaires en tendances et corrigées des fluctuations saisonnières par les processus

ARMA avec une estimation optimale des paramètres *p* et *q*. Un processus est dit ARMA(*p*,*q*) s'il vérifie l'équation (8) :

$$(8): \quad \text{ARMA}(p,q): X_t - \sum_{k=1}^{p} \phi_k X_{t-k} = \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Avec ϕ_k et θ_j les paramètres du modèle et ε_t les termes d'erreur.

1.2.3. Processus ARIMA et SARIMA

On parle d'un processus ARIMA (*AutoRegressive Integrated Moving Average*) dans le cas d'une série X_t non stationnaire, où il s'impose de la stationnariser par la méthode de passage aux différences selon un ordre d'intégration (ou de différentiation) noté *d*. La série différenciée est représentée par le modèle dit ARIMA (*p*,*d*,*q*).

Les modèles SARIMA (*Seasonnal AutoRegressive Integrated Moving Average*) permettent de modéliser les séries chronologiques qui présentent des variations périodiques en intégrant un ordre de différentiation pour les désaisonnaliser. Par la suite, l'estimation d'un modèle SARIMA se ramène à l'estimation d'un modèle ARMA sur la nouvelle série différenciée.

Box et Jenkins (1976) proposent à partir de l'ensemble de ces travaux une méthodologie pour modéliser et prédire les séries temporelles univariées, en étudiant leurs caractéristiques stochastiques et ainsi déterminer le modèle (de la famille des modèles ARIMA) le plus adéquat pour modéliser la série étudiée.

1.3. Approche par réseaux de neurones artificiels

1.3.1. Le principe d'apprentissage automatique

Le fonctionnement du cerveau humain et sa capacité à réaliser et à reconnaître des modèles complexes a pour longtemps motiver de nombreux travaux de recherche sur la simulation artificielle du système nerveux biologique en combinant des éléments de calcul relativement simples (neurones) pour produire un système fortement interconnecté, capable de calculer, d'apprendre, de se souvenir et d'optimiser et ce, en espérant l'émergence d'un phénomène complexe propre au cerveau humain tel que l'« intelligence » suite à un processus d'apprentissage.

Le réseau de neurones artificiels (*Artificial Neural Network*), est considéré comme un système informatique logiciel de la famille du *Deep Learning* dont le fonctionnement s'inspire de celui du cerveau humain. Théoriquement, un réseau neuronal est capable de produire une réponse appropriée à un problème donné (ou la meilleure réponse possible, lorsque plusieurs réponses sont possibles) même lorsque l'information est bruyante ou incomplète, ou lorsqu'il n'existe pas de protocole préétabli pour résoudre le problème (Borne P. et al., 2007).

C'est en utilisant des algorithmes qu'un *RNA* gère un flux d'informations entrant dans un neurone à partir d'une base de données pour en produire des résultats sous forme de sorties. Chaque neurone est connecté à un autre (ou plusieurs) neurone par une fonction d'activation[2], et chaque connexion est évaluée par un nombre réel, appelé poids, qui reflète le degré d'importance de la connexion donnée dans le réseau neuronal. Ce système est utilisé tant pour analyser des données que pour prédire le comportement futur d'une variable par le mécanisme d'apprentissage supervisé ou non-supervisé. Par analogie, les valeurs prédites sont modélisées de manière linéaire dans le cas d'un réseau simple dit *monocouche* ou encore *perceptron simple*, et de manière non-linéaire dans le cas d'un réseau *multicouches*.

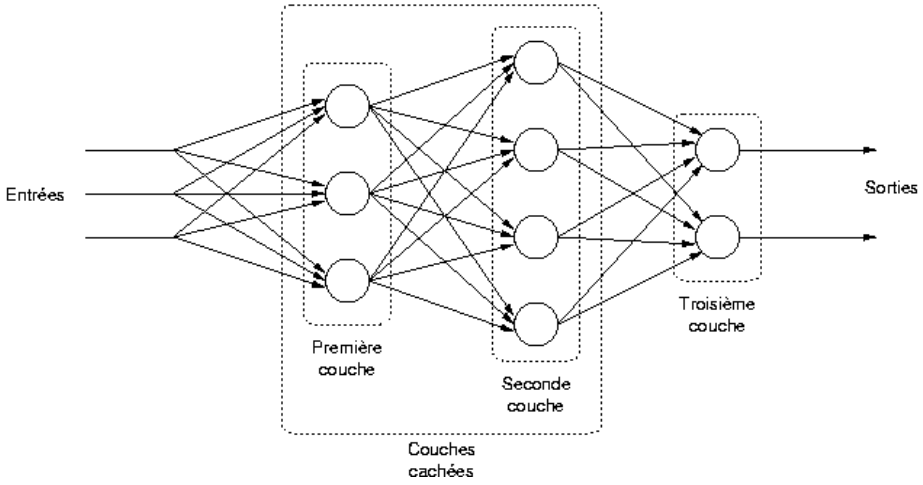
1.3.2. Le réseau multicouche

Dans le cas de la prédiction des séries temporelles, les réseaux de neurones multicouches (*MLP multilayer perceptron*) sont les mieux adaptés et les plus utilisés et sont devenus des modèles d'approximation universelle de classification et de prédiction, vu qu'ils permettent de réaliser une cartographie arbitraire d'un espace vectoriel sur un autre espace vectoriel.

Le réseau MLP est de type supervisé ou non, par correction des erreurs suivant la méthode de rétro-propagation algorithmique. Il contient une couche d'entrée, une ou plusieurs couche(s) cachée(s) à travers la(les)quelle(s) circulent l'information et finalement une couche de sortie, comme le montre la figure ci-dessous :

[2] Une fonction d'activation est une fonction mathématique qui opère la transformation des signaux d'entrée en les adaptant aux caractéristiques des sorties désirées. Elle permet donc de paramétrer l'influence d'un neurone sur un autre durant le processus de transfert d'information (Lallahem, 2003).

Figure 1: Architecture d'un réseau de neurones multicouches MLP



Source : Adapté de V. Gardeux, (2011)

Depuis la dernière décennie du 20ième siècle, L'utilisation des réseaux de neurones artificiels (*RNA*) réservée jusqu'à lors exclusivement aux sciences formelles, s'est étendu jusqu'au monde des sciences économiques en général et de la finance en particulier. Leur utilisation s'est avérée prétendument efficace lorsqu'il s'agit surtout de modéliser des systèmes complexes et difficiles à traiter par les méthodes statistiques classiques. On parle d'une approche connexionniste qui ne nécessite ni hypothèses préalables, ni relation spécifique entre variables dépendantes et variables indépendantes.

Nous avons pu constater que la littérature relative à ce champ de recherche est vaste et ne cesse de se développer et plusieurs revues scientifiques sont consacrées exclusivement à ce domaine de recherche.

B. Bower était parmi les premiers à avoir travaillé sur l'application des RNA dans le domaine de la finance. En 1988, il publie un papier sur le développement de réseaux neuronaux qui peuvent évaluer le risque des prêts hypothécaires et noter la qualité des obligations. En 1990, Hawley et al. proposent des applications potentielles des RNA dans le domaine de la finance, notamment, dans l'évaluation du risque de faillite des entreprises, l'identification des possibilités d'arbitrage sur les marchés de capitaux et l'analyse fondamentale et technique.

En ce qui concerne le comportement des prix sur les marchés financiers, un nombre considérable d'études y sont concentrées en mobilisant les méthodes par RNA au lieu des méthodes statistiques standards. Mentionnons l'étude de H. White 1989 qui a conçu un RNA

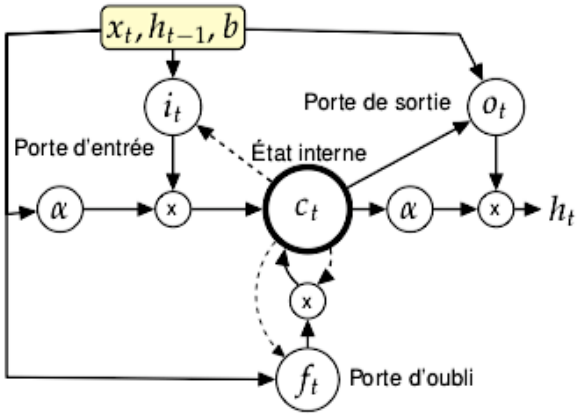
avec le rendement quotidien de l'action IBM sur une durée de 500 jours, l'objectif étant de prévoir les fluctuations futures du cours de l'action. Les résultats obtenus ont permis de mettre en évidence un comportement non aléatoire des rendements.

Plusieurs études avec des résultats similaires se sont succédés surtout lorsqu'il s'agissait de tester l'hypothèse de l'efficience des marchés (Weigend et al., 1992 ; Tsibouris & Zeidenberg, 1995 ; Kingdon, 1995 ; Chandra & Reeb, 1999) et plus récemment Culkin et al. en 2018 dans leur article « *Are Markets Truly Efficient? Experiments using Deep Learning for Market Movement Prediction* » où ils ont utilisé l'apprentissage profond pour anticiper l'évolution future de l'indice S&P 500 en utilisant les données quotidiennes de toutes les actions qui composent l'indice de la période allant de 1963 jusqu'à 2016. Leur étude a révélé que le comportement futur de l'indice S&P 500 peut être faiblement prédit en se basant uniquement sur les mouvements antérieurs des prix des actions sous-jacentes de l'indice. Cette difficulté de prévisibilité provient en grande partie des mouvements hasardeux et non pas de la non-stationnarité. Finalement, ils ont jugé qu'il s'agit du premier test d'efficience des marchés de forme faible, qui a utilisé un nombre important d'information et un réseau d'apprentissage approfondi entièrement connecté.

Toutefois, le principal problème rencontré par les chercheurs et les praticiens, était l'incapacité des neurones de se souvenir des informations précédentes qui ont circulé dans le réseau. Afin de pallier à ce problème qui présentait un frein surtout dans les prévisions des séries chronologiques et les données séquentielles, les réseaux de neurones récurrents (RNN) ont été conçus de telle façon que les couches cachées du système soient auto-connectées pour une durée relativement courte. Néanmoins Sepp Hochreiter et Jürgen Schmidhuber proposent en 1997 ce qu'on appelle le réseau LSTM (*Long Short Term Memory*) qui dispose d'un mécanisme dit de « gate » qui permet d'améliorer notablement la capacité de mémorisation du réseau.

La particularité d'un réseau LSTM réside dans sa cellule. L'état de la cellule ressemble à une bande transporteuse qui suit tout le circuit avec uniquement quelques interactions linéaires

Figure 2: Cellule d'un réseau LSTM



Source : Adapté de Mohamed Bouaziz (2017)

mineurs. Il est de ce fait, facile pour l'information de circuler tout au long du circuit sans modification. La cellule LSTM se caractérise par un nœud central qui constitue la mémoire interne de la cellule et 3 catégories de portes qui permettent de gérer l'information.

Les réseaux LSTM sont désormais mobilisés dans plusieurs domaines surtout lorsqu'il s'agit de traiter des données temporelles ou séquentielles car ils permettent de mémoriser l'information dans un horizon long tout en évitant l'explosion ou la dissipation du gradient (Hochreiter et al., 2001).

2. Méthodologie et résultats

Dans le but d'analyser le comportement de l'indice MASI (Moroccan All Shares Index) durant le premier semestre de l'an 2020 et en absence de données à haute fréquence (*intraday*), nous allons travailler avec les données quotidiennes disponibles et téléchargeables directement sur le site officiel de la Bourse de Casablanca.

La série obtenue est une suite de variables aléatoires de 127 observations. Comme il s'agit d'évaluer l'évolution d'un phénomène au cours du temps, l'ordre est important. Nous notons $MASI_t$ la valeur observée de l'indice au moment t. Nous souhaitons estimer et ajuster un modèle économétrique qui représente et modélise au mieux les données disponibles pour en prévoir les valeurs futures.

L'observation du graphique fait ressortir trois phases d'évolution. La première phase allant du 1er janvier au 1er mars, est marqué par une résistance horizontale légèrement au-dessus de la valeur de 12 000 et en dessous de celle de 13 000. Cette phase ne dura pas longtemps car une cassure surgira le 02 mars (date qui coïncide avec l'annonce officielle du premier cas testé positif au Corona virus), le MASI chutera alors considérablement pendant 15 jours d'affilé marqués par une panique boursière et une baisse tendancielle avant que la valeur ne se stabilise en moyenne entre 10 500 et 9 000 en réaction de la décision de Bank Al-Maghrib de la baisse du taux directeur.

Toutefois, et vu que les chercheurs et les investisseurs s'intéressent plus au rendement d'un indice plutôt qu'à son évolution en niveau, nous allons travailler sur la série des rendements de l'indice MASI qu'on notera $RMASI_t$:

$$(9) : RMASI_t = \ln\left({MASI_t}/{MASI_{t-1}}\right)$$

Avec, $MASI_t$ valeur de l'indice à l'instant *t* et $MASI_{t-1}$ valeur de l'indice à l'instant *t-1*.

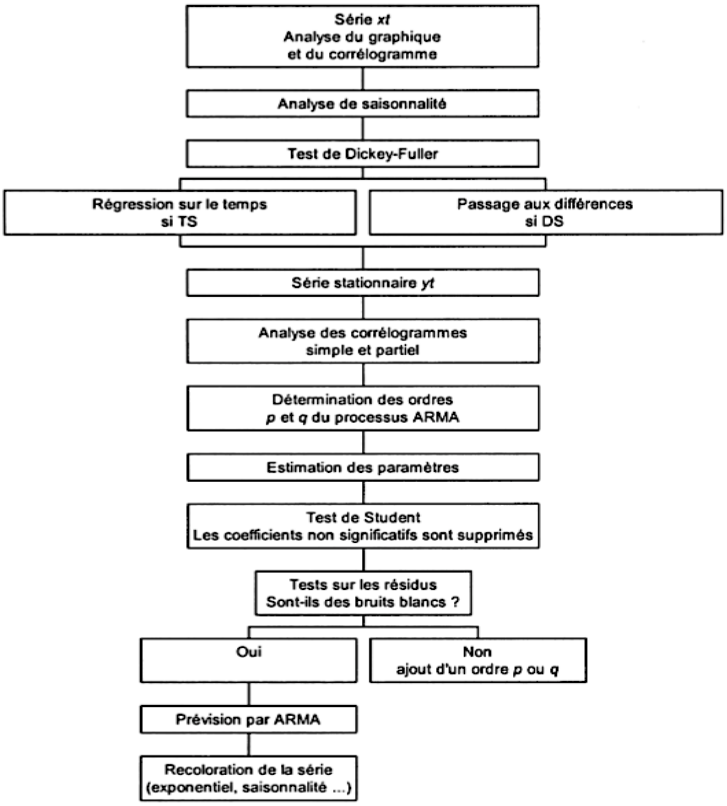
Figure 3: Graphique de l'évolution de l'indice MASI

Source :Bourse de Casablanca

2.1. Méthodologie de Box et Jenkins

Suivant la démarche de Box et Jenkins (1976), et afin d'aboutir à la phase de prévision, il faut suivre les étapes schématisées sur la figure ci-dessous :

Figure 4: Démarche de la méthodologie de Box et Jenkins



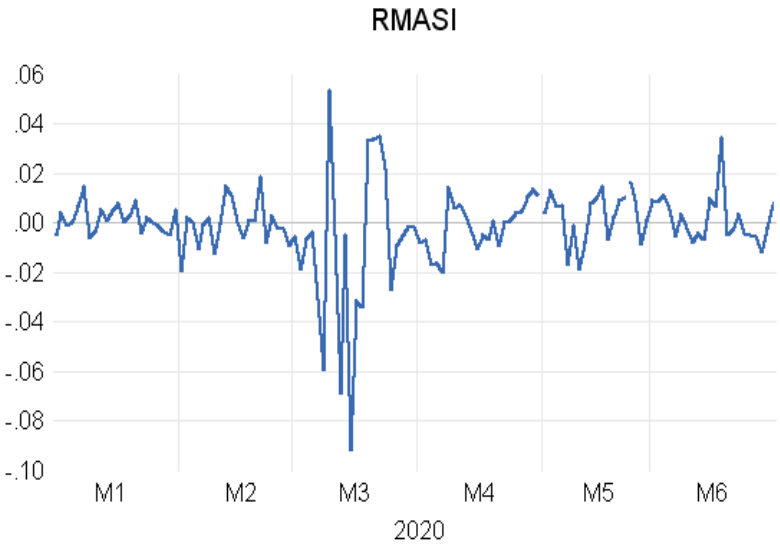
Source : Adapté de Régis Bourbonnais, (2015)

Les différents tests et calculs ont été réalisés sur le logiciel Eviews11.

2.1.1. Analyse du graphique et du corrélogramme

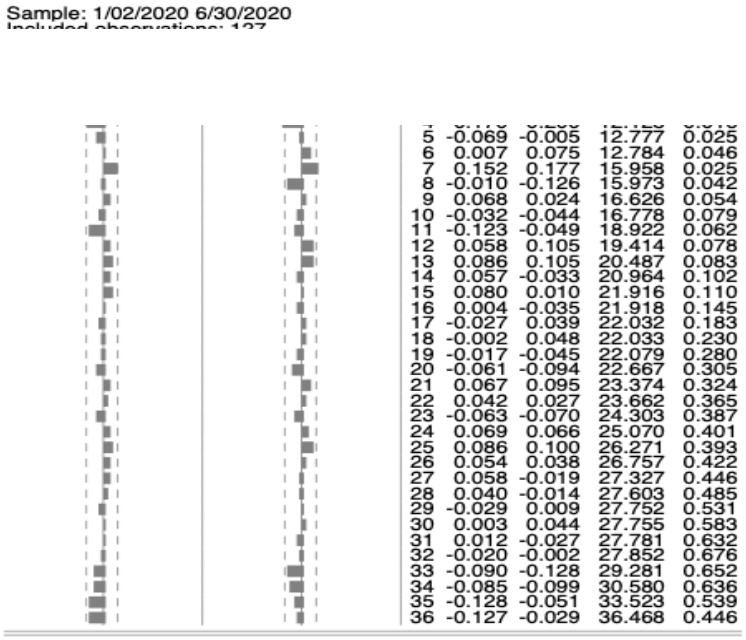
Le graphique ci-dessous retrace l'évolution de la série des rendements de l'indice MASI en unité de temps hebdomadaire ainsi que son corrélogramme :

Figure 5: L'évolution de la série des rendements de l'indice MASI



Source : Eviews11

Figure 6: Les fonctions d'autocorrélation simple et partielle de la série



Source :Eviews11

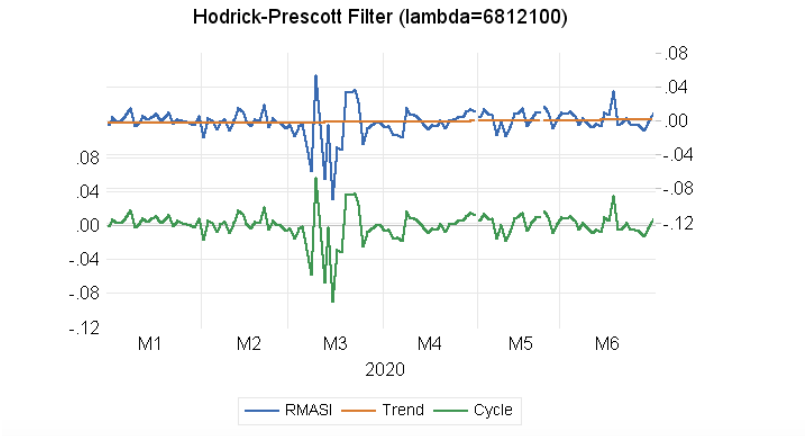
De prime abord, et d'après la lecture visuelle du graphique de l'évolution de la série des rendements, nous pouvons constater qu'elle est stationnaire en tendance et qu'elle fluctue autour de sa moyenne. Ce constat est confirmé par la représentation graphique des fonctions d'autocorrélation « corrélogramme », sur lequel nous remarquons que sur la quasi-totalité des observations, la probabilité de la statistique Q est largement > 5% et les termes du corrélogramme sont à l'intérieur de l'intervalle de confiance. Le processus est donc en principe stationnaire.

Néanmoins, l'observation et l'analyse du corrélogramme en elle seule ne suffit pas pour juger de la stationnarité de la série, il s'impose d'appliquer un test de racine unitaire « *Unit Root Test* » qui permettra de confirmer ou infirmer notre constat.

2.1.2. Analyse de saisonnalité

Afin de détecter l'éventuelle existence d'effets saisonniers sur notre série, nous allons procéder par la méthode graphique du filtre Hodrick-Prescott afin de dissocier la série brute des cycles conjoncturels (tendance et saisonnalité). Le graphique est fourni automatiquement sur Eviews :

Figure 7: Représentation graphique du filtre Hodrick-Prescott de la série des rendements



Source : Eviews11

L'analyse du graphique fait ressortir l'absence de perturbations qui se répètent dans le temps par intervalle régulier, donc notre série n'est pas affecté par une saisonnalité.

2.1.3. Étude de stationnarité

En vue de mettre à l'épreuve le caractère stationnaire ou non de notre série, nous allons appliquer les tests de Dickey-Fuller Augmenté.

Cependant, il faut tout d'abord préciser le nombre de retards *p* à introduire dans la régression, on parle alors de correction paramétrique de l'autocorrélation. Pour ce faire, on peut soit tester plusieurs valeurs sur les trois modèles du test et choisir le nombre de retards qui minimise les critères d'Akaike et Schwartz, soit utiliser la sélection automatique directement sur le logiciel Eviews.

Ainsi, en appliquant le modèle 3 avec constante et tendance, nous obtenons :

Figure 8: Estimation du modèle 3

Null Hypothesis: RMASI has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-9.015531	0.0000
Test critical values:	1% level	-3.482879	
	5% level	-2.884477	
	10% level	-2.579080	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(RMASI)
Method: Least Squares
Date: 09/27/20 Time: 19:47
Sample (adjusted): 1/03/2020 6/30/2020
Included observations: 126 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RMASI(-1)	-0.792913	0.087950	-9.015531	0.0000
C	-0.001079	0.001497	-0.720405	0.4726

R-squared	0.395946	Mean dependent var	0.000102
Adjusted R-squared	0.391075	S.D. dependent var	0.021454
S.E. of regression	0.016741	Akaike info criterion	-5.326148
Sum squared resid	0.034753	Schwarz criterion	-5.281128
Log likelihood	337.5474	Hannan-Quinn criter.	-5.307858
F-statistic	81.27980	Durbin-Watson stat	2.036680
Prob(F-statistic)	0.000000		

Source Eviews11

Nous remarquons que la valeur de t-Statistic obtenu par le test (0,68) est largement inférieur à sa valeur lue sur la table statistique au risque de 5% (3,14). Notre série n'est donc pas affecter par une tendance.

En suivant la stratégie du test ADF, nous passons à l'estimation du modèle 2 :

Figure 9: Estimation du modèle 2

Null Hypothesis: RMASI has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-9.022063	0.0000
Test critical values:	1% level	-4.032498	
	5% level	-3.445877	
	10% level	-3.147878	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(RMASI)
Method: Least Squares
Date: 09/27/20 Time: 19:13
Sample (adjusted): 1/03/2020 6/30/2020
Included observations: 126 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RMASI(-1)	-0.797150	0.088356	-9.022063	0.0000
C	-0.002876	0.003024	-0.951165	0.3434
@TREND("1/02/2020")	2.82E-05	4.12E-05	0.684695	0.4948

R-squared	0.398240	Mean dependent var	0.000102
Adjusted R-squared	0.388455	S.D. dependent var	0.021454
S.E. of regression	0.016777	Akaike info criterion	-5.314080
Sum squared resid	0.034621	Schwarz criterion	-5.246549
Log likelihood	337.7870	Hannan-Quinn criter.	-5.286644
F-statistic	40.70021	Durbin-Watson stat	2.035140
Prob(F-statistic)	0.000000		

Source : Eviews11

De même, la valeur de t-Statistic du modèle 2 (-0,72) est inférieure à la valeur tabulée (2,86). Donc notre série n'est pas affectée par une constante car elle est non significative. Il convient de passer à l'estimation du dernier modèle :

Figure 10: Estimation du modèle 1

Null Hypothesis: RMASI has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-9.004305	0.0000
Test critical values:	1% level	-2.583444	
	5% level	-1.943385	
	10% level	-1.615037	

Source : Eviews11

En comparant la valeur de t-Statistic avec sa valeur critique au seuil de 5%, on conclut au rejet de l'hypothèse nulle de racine unitaire. Notre série représente donc un processus stationnaire en niveau qui ne nécessite pas le passage aux différences premières pour pouvoir réaliser des prévisions. Nous pouvons passer directement à l'estimation des paramètres du modèle ARMA(p,q). Par analogie, si une série chronologique se montre avoir une racine unitaire, elle présente alors un modèle systématique imprévisible.

2.1.4. Estimation des paramètres du modèle ARMA

D'après l'examen de la forme des représentations graphique des fonctions d'autocorrélation simple et partielle plusieurs combinaisons sont possibles : AR(1), AR(4), ARMA(2,2), ARMA(1,2), ARMA(1,1), ARMA(1,3), ARMA(2,1), MA(2),… Cependant, le modèle retenu et celui qui minimise le critère d'information Schwartz. D'où nous retenons le modèle ARMA(1,2).

Figure 11: propriété du modèle

Dependent Variable: D(RMASI)
Method: ARMA Maximum Likelihood (BFGS)
Date: 09/28/20 Time: 19:21
Sample: 2 127
Included observations: 126
Convergence achieved after 11 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	-0.844497	0.030592	-27.60479	0.0000
MA(2)	-0.856262	0.039326	-21.77332	0.0000
SIGMASQ	0.000285	1.72E-05	16.58785	0.0000

Source : Eviews11

2.1.5. Validation du modèle

Selon le test de students, les coefficients sont significativement différents de 0 avec une probabilité critique inférieur à 0,05.

La validation du modèle passe par l'analyse des résidus. Ces derniers doivent obéir aux règles d'un processus de bruit blanc : absence d'autocorrélation entre les erreurs et homoscédasticité des résidus. Pour ce faire, nous allons analyser le corrélogramme du résidu et le corrélogramme du résidu au carré.

Figure 12: Corrélogramme du résidu et corrélogramme du résidu au carré

Sample (adjusted): 3 127
Q-statistic probabilities adjusted for 2 ARMA terms

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob*
		1	0.189	0.189	4.5630	
		2	0.111	0.078	6.1410	
		3	0.009	-0.026	6.1516	0.013
		4	-0.013	-0.020	6.1740	0.046
		5	-0.071	-0.066	6.8359	0.077
		6	0.070	0.101	7.4804	0.113
		7	-0.035	-0.055	7.6471	0.177
		8	-0.117	-0.126	9.5137	0.147
		9	0.013	0.068	9.5354	0.216
		10	-0.043	-0.038	9.7859	0.280
		11	-0.117	-0.109	11.695	0.231
		12	0.052	0.093	12.079	0.280
		13	0.082	0.074	13.026	0.292
		14	-0.019	-0.046	13.075	0.364
		15	0.092	0.072	14.286	0.354
		16	0.018	-0.023	14.335	0.425
		17	-0.046	-0.029	14.651	0.477
		18	0.023	0.027	14.728	0.545
		19	0.015	-0.026	14.764	0.612
		20	-0.077	-0.041	15.652	0.617
		21	0.091	0.120	16.907	0.596
		22	0.099	0.054	18.423	0.560
		23	-0.110	-0.135	20.301	0.502
		24	0.079	0.134	21.292	0.503
		25	0.049	0.018	21.672	0.540
		26	0.069	0.074	22.432	0.553
		27	0.002	-0.050	22.433	0.611
		28	0.022	-0.051	22.513	0.660
		29	-0.061	0.042	23.134	0.678
		30	0.020	0.025	23.204	0.723
		31	0.094	0.046	24.704	0.694
		32	-0.008	0.018	24.713	0.739
		33	-0.061	-0.069	25.361	0.751
		34	-0.070	-0.101	26.215	0.754
		35	-0.108	-0.023	28.271	0.702
		36	-0.138	-0.104	31.651	0.583

Sample (adjusted): 3 127
Included observations: 125 after adjustments

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob*
		1	0.018	0.018	0.0417	0.838
		2	-0.003	-0.003	0.0428	0.979
		3	-0.001	-0.001	0.0430	0.998
		4	-0.043	-0.043	0.2889	0.991
		5	0.035	0.036	0.4481	0.994
		6	0.014	0.013	0.4751	0.998
		7	0.003	0.003	0.4764	1.000
		8	-0.100	-0.103	1.8414	0.985
		9	-0.023	-0.016	1.9119	0.993
		10	-0.008	-0.008	1.9203	0.997
		11	0.038	0.039	2.1260	0.998
		12	-0.061	-0.073	2.6479	0.998
		13	-0.050	-0.043	3.0075	0.998
		14	0.072	0.078	3.7421	0.997
		15	-0.058	-0.056	4.2222	0.997
		16	0.099	0.085	5.6541	0.991
		17	-0.030	-0.042	5.7849	0.995
		18	-0.068	-0.058	6.4792	0.994
		19	-0.011	-0.009	6.4983	0.996
		20	-0.069	-0.073	7.2184	0.996
		21	0.095	0.084	8.5938	0.992
		22	0.030	0.029	8.7281	0.995
		23	-0.049	-0.054	9.1047	0.996
		24	-0.010	0.007	9.1208	0.997
		25	0.059	0.057	9.6702	0.997
		26	-0.064	-0.071	10.324	0.997
		27	0.086	0.076	11.525	0.996
		28	-0.021	-0.042	11.596	0.997
		29	-0.139	-0.107	14.796	0.987
		30	-0.020	-0.038	14.863	0.990
		31	-0.077	-0.069	15.858	0.989
		32	0.008	-0.011	15.869	0.992
		33	0.060	0.072	16.491	0.993
		34	0.015	0.037	16.529	0.995
		35	-0.011	-0.028	16.551	0.997
		36	-0.062	-0.057	17.237	0.997

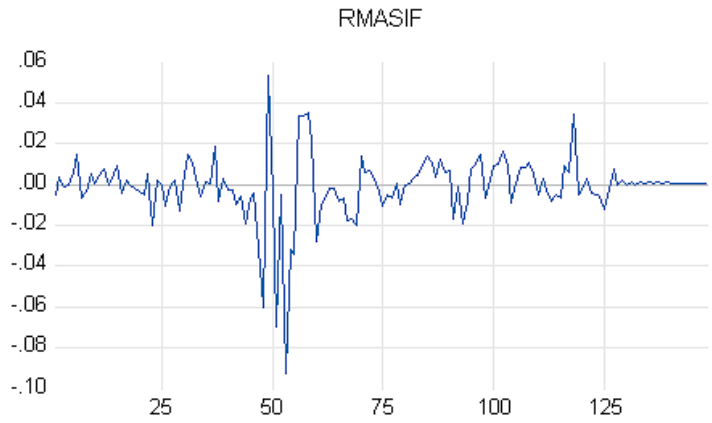
Source : Eviews11

Le corrélogramme du résidu montre qu'il s'agit bien d'un processus sans mémoire vu que les probalibités de la statistique Q sont supérieurs à 0,05, donc il n'existe pas d'autocorrélation entre les erreurs. Ainsi, le test ARCH d'hétéroscédasticité (corrélogramme du résidu au carré), indique que les résidus sont homoscédastiques car les termes de la fonction ne sont pas significativement différents de 0 (ils sont tous situés à l'intérieur de l'intervalle de confiance). Les résidus se comportent alors comme un bruit blanc gaussien selon la probabilité critique de la statistique de Jarque-Bera (0,29). Nous acceptons l'hypothèse H0 de normalité des résidus.
La série des rendements de l'indice MASI peut être modélisée par un processus ARMA (1,2) sans constante.

2.1.6. Prévisions

Les prévisions effectuées sur la série étudiée concernent le mois de juillet 2020. Ces résultats vont être comparés par la suite avec ceux obtenus en utilisant l'approche par apprentissage automatique et les observations réelles journalières téléchargées sur le site de la bourse de Casablanca.

Figure 13: Prévisions de la série des rendements



Source : Eviews11

2.2. Auto-apprentissage par réseau LSTM

Pour prédire les valeurs futures de notre série, nous formerons un réseau LSTM par régression de séquences séquentielles, où les réponses sont les séquences de training (apprentissage) avec des valeurs décalées d'une unité de temps (un jour). Autrement dit, pour chaque valeur journalière d'entrée, le réseau LSTM apprend à prédire la valeur du jour suivant.

Les différents tests mobilisés sont réalisés sur MATLAB R2019b et les données d'entrée sont remodelées sous forme de vecteurs en ligne.

2.2.1. Préparation des données

L'évaluation de la performance des prévisions par les réseaux de neurones récurrents nécessite la séparation des données en deux sous-ensembles : sous-ensemble de training et sous-ensemble de test. Dans notre cas, les données quotidiennes du premier semestre 2020 feront l'objet des données d'apprentissage et les données du mois de juillet celles du test.

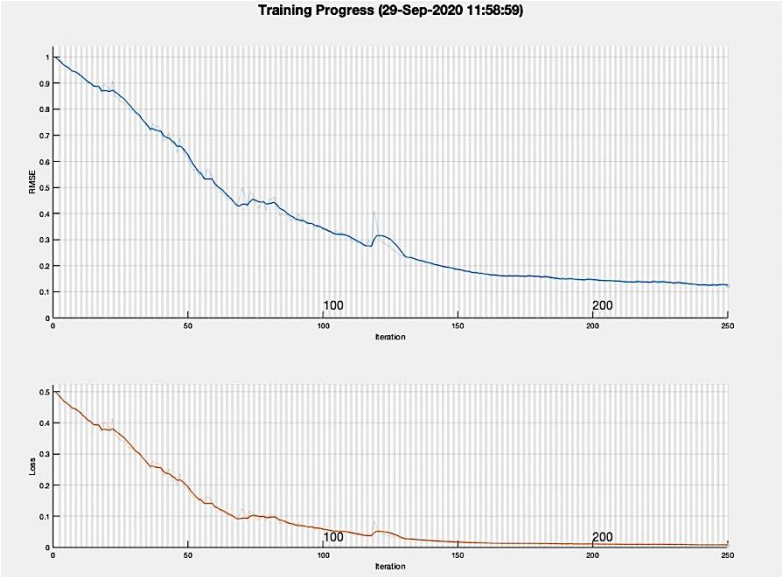
Ainsi, pour un meilleur ajustement, nous avons normalisé les données du traning de telle façon qu'elles aient une moyenne et une variance nulle. De même, au moment de la prédiction il faut normaliser les données du test.

La prédiction par réseau LSTM est basée sur le principe de régression. En effet, une valeur prédire à l'instant *t* n'est autre que le résultat d'apprentissage par l'ensemble des valeurs qui la précèdent jusqu'au *t-1*.

2.2.2. L'architecture du réseau

Le réseau conçu est constitué de 200 couches cachées, 250 cycles d'apprentissage et un seuil de gradient de 1(pour éviter son explosion ou dissipation). Le taux d'apprentissage initial est de 0,005 qu'on multipliera par un facteur de 0,2 après le 125ème cycle.

Figure 14: Le processus d'apprentissage



Source : MATLAB R2019b

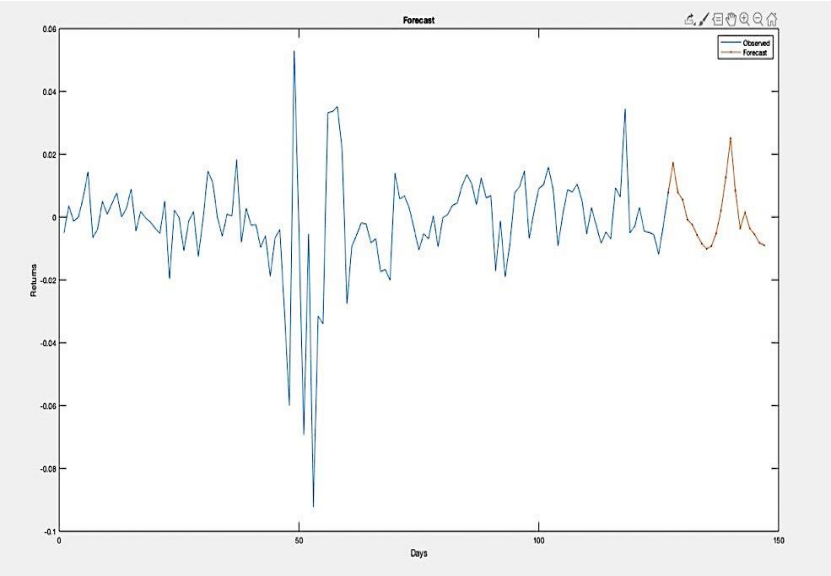
Nous constatons que la valeur du critère de l'erreur quadratique moyenne (la racine carrée de la moyenne des carrées des erreurs entre les valeurs prédites et les valeurs observées) notée RMSE (root-mean-square error) diminue en fonction de l'avancement des cycles d'apprentissage.

2.2.3. Prévisions

Pour prévoir les valeurs du mois de juillet, il faut mettre à jour l'état du réseau après chaque prédiction. Pour chaque valeur prédite, nous utilisons la prédiction précédente comme entrée dans la fonction.

Après la normalisation des données du test, nous obtenons les résultats suivants :

Figure 15: Graphique des prévisions



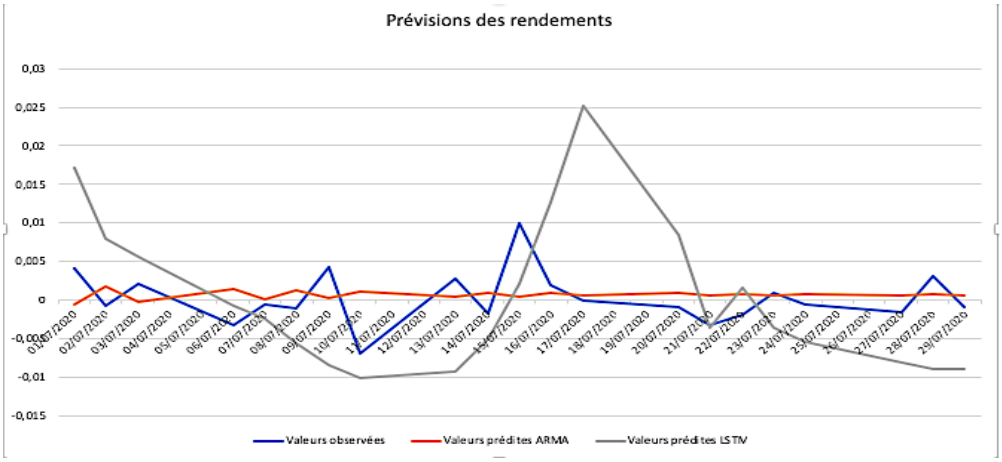
Source : MATLAB R2019b

3. Discussion

La prévision des indices boursiers a pour longtemps suscité l'intérêt de plusieurs chercheurs. Cependant l'objectif de ce papier et non uniquement de faire des prédictions mais surtout de comparer les résultats obtenus par une méthode statistique systématique telle que celle de Box et Jenkins, et une méthode connexionniste basée sur l'auto-apprentissage telle que le réseau LSTM.

Le graphique ci-dessous schématise les valeurs (rendements de l'indice MASI) prédites par chacune des deux méthodes en les confrontant avec les valeurs réelles observées :

Figure 16: Graphique des rendements du mois de juillet



Source : MATLAB R2019b

D'après le graphique, nous remarquons que le réseau LSTM a pu suivre l'évolution de l'indice en découvrant après chaque cycle d'apprentissage la relation entre les valeurs d'entrée et celles de sortie contrairement au modèle ARIMA où les valeurs prédites sont relativement stables.
Bien évidemment l'examen du graphique en lui seul ne permet pas de juger de la pertinence de chaque méthode. Le RMSE reste un indicateur largement accepté quant à l'appréciation du pouvoir prédictif d'un modèle (Bennett et al., 2013). Il est obtenu par la formule :

$$(10) : \qquad RMSE = \sqrt{\frac{1}{n}\sum_{t-1}^{n}(\varepsilon_t)^2}$$

Où, ε_t est l'écart entre la valeur prédite et la valeur observée à l'instant *t*.
L'application de la formule sur les valeurs prédites par le modèle ARMA (1,2) nous fournit un RMSE de 0,0171. Pour le réseau LSTM, la valeur de l'indicateur est obtenue automatiquement sur MATLAB.

	ARMA (1,2)	Réseau LSTM
RMSE	0,0171	0,0104

En se basant sur les valeurs du RMSE nous pouvons conclure que la qualité des prédictions du réseau LSTM est meilleure que celle du modèle ARMA.
Nos résultats sont similaires à ceux obtenus dans d'autres études réalisées dans le cadre du rapprochement entre les algorithmes d'apprentissage automatique et les différents modèles économétriques. R. Tir (2004), calcule les prévisions des deux indices boursiers de la place de Tunis en comparant un modèle ARIMA et un réseau neuronal multicouche et conclue que l'utilisation des réseaux neuronaux pour prévoir le comportement futur des indices boursiers conduit à des résultats satisfaisants. Ben Houad et Oubouali (2018), effectuent des prévisions de la liquidité des actions cotées à la place casablancaise en utilisant tout d'abord la modélisation ARIMA et ensuite le réseau de neurone autorégressif non linéaire avec entrées exogènes NARX. En se basant sur l'erreur moyenne quadratique et l'erreur absolue moyenne, ils affirment que la modélisation NARX est plus performante en termes de la qualité des prévisions obtenus.

Conclusion

Le marché financier est un marché complexe et dynamique caractérisé par une forte volatilité, d'où la difficulté d'y réaliser des prévisions. Les implications financières et stratégiques d'une

prédiction précise des mouvements sur les marchés financiers ont motivé les chercheurs et les praticiens à déployer une panoplie de méthodes de modélisation purement statistiques. Or, les techniques économétriques en matière de prédiction des séries temporelles ne peuvent pas battre de manière significative la marche aléatoire la plus simple (Xiao et al., 2014), ce qui a encouragé les chercheurs à développer des modèles de prévision plus performants. Ainsi, les modèles mobilisant l'intelligence artificielle ont permis d'obtenir de bons résultats par rapport aux modèles statistiques classiques (Valipour et al., 2013).

La présente étude est selon notre connaissance, la première tentative de prédiction du rendement de l'indice MASI de la Bourse de Casablanca en mobilisant à la fois une approche économétrique traditionnelle et une approche connexionniste. Cependant, bien que le réseau LSTM s'est avéré plus performant en termes de prévisions sur les séries temporelles, l'approche présente néanmoins, l'inconvénient majeur de la nécessité d'un grand nombre de données pour obtenir de meilleures prévisions. Le nombre de données d'entrée que nous avons mobilisé est jugé modeste par rapport à la capacité du réseau. Dans ce sens, des résultats meilleurs peuvent être obtenus en utilisant des méthodes plus récentes et plus sophistiqués basées sur des algorithmes d'apprentissage par renforcement de type Q-Learning qui mobilisent des données limitées pour obtenir des récompenses maximales. De même, l'utilisation de données à haute fréquence au lieu de données journalières dans la prédiction des séries chronologiques sur les marchés boursiers pourrait aboutir à des résultats considérables.

La rigueur de la méthode de prévision n'intéresse pas uniquement les investisseurs dans la mesure où elle constitue un outil de prise de décision, mais aussi les chercheurs en finance autour du débat de l'efficience informationnelle des marchés qui stipule qu'il est quasiment impossible de faire des prévisions sur un marché efficient.

BIBLIOGRAPHIE

Ben Houad, M., Oubouali, Y. (2018). Prévisions de la liquidité des actions cotées à la bourse des valeurs de Casablanca : comparaison entre la modélisation ARIMA et les réseaux de neurones NARX. La Revue Gestion et Organisation 10 (2), 83-99.

Chen, M. Y., Chen, B. T. (2015). A hybrid fuzzy time series model based on granular computing for stock price forecasting. Journal of Information Sciences 294, 227-241.

Borne, P., Benrejeb, M., Haggège, J., (2007). Les réseaux de neurones : présentation et applications, Paris, Technip.

Bourbonnais, R., (2015). Économétrie, Paris, Dunod.

Hawley, D.D., Johnson, J.D., Raina, D., (1990). Artificial Neural Systems: A New Tool for Financial Decision-Making. Financial Analysts Journal, 46, 63–72.

Hu, Y.-C. (2017). Electricity consumption prediction using a neural-network-based grey forecasting approach. Journal of the Operational Research Society, 68(10), 1259-1264.

Pavlidis N. G., Tasoulis D. K., Vrahatis M. N., (2002), Financial forecasting through unsupervised clustering and evolutionary trained neural networks. Working paper, Department of Mathematics, University of Patras, Greece.

Rubio, A., Bermúdez, J. D., & Vercher, E. (2017). Improving stock index forecasts by using a new weighted fuzzy-trend time series method. Journal of Expert Systems With Applications, 76, 12–20.

Sezer, O.B., Ozbayoglu, A.M., Dogdu, E., (2017). An Artificial Neural Network-based Stock Trading System Using Technical Analysis and Big Data Framework. Presented at the SouthEast Conference, ACM, Kennesaw GA USA, 223–226.

Stern, H.S., (1996). Neural Networks in Applied Statistics. Journal of Technometrics, 38, 205–214.

Svozil, D., Kvasnicka, V., Pospichal, J., (1997). Introduction to multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems, 39, 43–62.

Thiria, S., Lechevallier, Y., Gascuel, O., (1997). Statistique et méthodes neuronales, Sciences sup, Paris, Dunod.

Valipour, M., Banihabib, M. E., Behbahani, S. M. R., (2013). Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. Journal of Hydrology 476, 433–441.

Wise, J., (1956). Stationarity Conditions for Stochastic Processes of the Autoregressive and Moving-Average Type. Biometrika, 43, 215–219.

Xiao, Y., Xiao, J., Liu, J., Wang, S., (2014). A multiscale modeling approach incorporating ARIMA and ANNS for financial market volatility forecasting. Journal of Systems Science and Complexity, 27(1), 225–236.