

مقارنة بين نماذج السلاسل الزمنية الخطية وتقنيات تنقيب البيانات للتنبؤ بتلوث الهواء في مدينة بغداد

احمد طلال جبار^[2]

الجامعة التقنية الوسطى، الكلية التقنية الادارية/ بغداد، قسم تقنيات المعلوماتية، بغداد، العراق

ahmedtq15@gmail.comأ.م.د. نشأت جاسم محمد^[1]

الجامعة التقنية الوسطى، الكلية التقنية الادارية/ بغداد، قسم تقنيات المعلوماتية، بغداد، العراق

nashaat1974@gmail.com

المستخلص:

يعتبر الاهتمام بالبيئة احد الاولويات المهمة في البلدان المختلفة ، ويشكل التلوث البيئي اهم مصدر لتهديد البيئة حيث وصلت مستويات التلوث في جوانب البيئة المائية والجوية والارضية الى حدود خطيرة مما استدعى اهتمام الباحثين في مختلف العلوم للاهتمام بالابحاث التي من شأنها تقليل تلك المسببات ومراقبتها وتقليل نسبها ضمن الحدود المسموحة باستمرار. يعتبر تلوث الجو والهواء احد المخاطر الرئيسية في التلوث البيئي والتي لها تاثيرات مباشرة على حياة الانسان تمثلت بارتفاع درجة حرارة الارض وتاكل طبقة الاوزون بسبب انبعاث الغازات الخطير مباشرة الى الجو واهمهما غازي ثنائي أكسيد النتروجين (NO₂) وثنائي أكسيد الكبريت (SO₂).

في هذا البحث تم توظيف نماذج السلاسل الزمنية متمثلة بنماذج الانحدار الذاتي الاوساط المتحركة المتكاملة Auto Regressive Integrated Moving Average و أنموذج متجه الانحدار الذاتي Vector Auto Regressive وطرائق تنقيب البيانات متمثلة بخوارزميات الشبكات العصبية . اذ تضمن البحث الملوثين ثنائي أكسيد النتروجين (NO₂) وثنائي أكسيد الكبريت (SO₂) في مدينة بغداد للمدة (2015-2017) وبمعدل اسبوعي 157 مشاهدة.

وتوصل البحث الى تفوق خوارزميات الشبكات العصبية على نماذج متجه الانحدار الذاتي و الانحدار الذاتي الاوساط المتحركة المتكاملة. كذلك تفوق نماذج متجه الانحدار الذاتي على نماذج الانحدار الذاتي الاوساط المتحركة المتكاملة.

الكلمات المفتاحية: التنبؤ، تلوث الهواء، السلاسل الزمنية، نماذج الانحدار الذاتي الاوساط المتحركة المتكاملة، الشبكات العصبونية الاصطناعية، متجه الانحدار الذاتي.

A Comparison between Data Mining Techniques and Time Series Models for Forecasting Air Pollution in Baghdad

Dr. Nasshat Jasim Mohammed

Middle Technical University/ Technical College of Management / Iraq.

nashaat1974@gmail.com

Ahmed Tallal Jabbar

Middle Technical University/ Technical College of Management / Iraq.

ahmedtq15@gmail.com**Abstract:**

Concern for the environment is an important priority in different countries, Environmental pollution is the most important source of environmental threat. Pollution levels in the water, air and land environment have reached serious limits, requiring researchers in various sciences to take care of researches that reduce, monitor and reduce their causes within the limits allowed. Air pollution is one of the main threats to environmental pollution, which has a direct impact on human life. It is due to the increase in the temperature of the earth and the depletion of the ozone layer due to the dangerous emissions of gases directly into the atmosphere, mainly NO₂ and SO₂ .

This paper aims to compare the Box & Jenkins method that include Auto Regressive Integrated Moving Average models (ARIMA), Vector Auto regressive models (VAR) and data mining techniques that include Artificial Neural Network (ANN). The data gathered from Baghdad city for the period 2015-2017 with 157 observations.

This paper showed the superiority ANN models compared to VAR and ARIMA models, as well as VAR models compared to ARIMA models.

Keywords: Forecasting, Air Pollution, Artificial Neural Network, Auto regressive integrated moving average model, Vector Auto regressive Models.

1. المقدمة

" لقد اهتم الإنسان منذ أقدم العصور باستشراف المستقبل والتنبؤ بتغيراته وقيمه . ومع نمو الكثافة السكانية وزيادة تعقيدها، ظهرت العديد من المشاكل منها زيادة نسب الملوثات في الهواء التي مصدرها الغازات الملوثة ، الشوائب وعوادم السيارات ، المصانع والرياح وما الى ذلك ، ولذلك أصبحت عملية التنبؤ على المدى البعيد حاجة ملحة. ان العلاقة بين التلوث بالغازات الملوثة والعوامل الخارجية هي معقدة ولا خطية، مما يجعل من الصعب الحصول على تنبؤات دقيقة من خلال استخدام التقنيات التقليدية. ومن هنا ظهرت مشكلة البحث التي تتمثل باستخدام نماذج تقنيات تنقيب البيانات الخاصة بالتنبؤ والتحليل الطواهر البيئية ومقارنتها مع الطرائق التقليدية. اذ تم اختيار اخطر ملوثين ثنائيي أوكسيد النتروجين NO2 و وثنائيي أوكسيد الكبريت (SO2) ولمدة ثلاثة سنوات وبمعدل اسبوعي (157).

وكانت هنالك العديد من البحوث في هذا المجال ففي عام 2000 قام الباحثان [16] Jehng & Shang بتوقع جودة الهواء في تايوان باستخدام السلاسل الزمنية التقليدية والشبكات العصبية. حيث جمع بيانات الملوثين ثنائيي اوكسيد الكبريت والاوزون من محطتين رئيسيتين ومحطة صناعية، وكانت نتائج توقع الشبكة العصبية خطوة مرضية وجيدة ومتفوقة على النموذج التقليدي ARIMA . وفي عام 2008 قام الباحثون [14] Ming & Yafeng & Min بالتنبؤ بتركيزات ملوثات الهواء كل ساعة بالقرب من قوانغتشو، الصين في ثلاثة مواقع مختارة إذ إظهرت دراسة مقارنة أن نماذج الشبكات العصبية تتفوق على نماذج الانحدار الخطي المتعددة . وفي عام 2009 قام الباحثان Ujjwal & V. K. Jain [23] باستخدام نماذج ARIMA و ARMA بالتنبؤ بتلوث الهواء (NO2، NO ، CO ، O3) في مناطق تركيز حركة المرور في المناطق الحضرية نيودلهي، الهند . وفي عام 2013 قام الباحثون Snezhana & Atanas & Desislava & Doychin [13] بتطبيق تحليل العوامل ومنهجية Box-Jenkins لدراسة تركيزات ملوثات الهواء الأولية في بلغاريا خلال فترة سنة واحدة. وفي عام 2017 قام مجموعة من الباحثون [14] Suling & Xiuyuan & Haixia & Jianming , Yuanyuan & Jinxing بتطبيق نماذج هجينة للتنبؤ بتلوث الهواء في الصين ومقارنة نتائج هذه النماذج مع نماذج ARIMA ونموذج SAR، تظهر نتائج التنبؤات أن النموذجين الهجين المقترحين يتفوقان على ARIMA.

2. الاطار النظري:

1.2 السلاسل الزمنية ومكوناتها Time Series and components

تعرف السلسلة الزمنية رياضياً بأنها متتابعة من المتغيرات العشوائية ضمن فضاء من الاحتمالية متعددة المتغيرات وترمز له بالرمز (t) والذي يعود الى مجموعة من الفترات (T) وعادة مايرمز للسلسلة الزمنية $\{X(t); t \in T\}$ فإذا كان الزمن (t) يأخذ قيماً مستمرة (Continuous) فإن السلسلة الزمنية تسمى مستمرة في الزمن ويرمز لها $\{X(t); -\infty \leq t \leq \infty\}$ أما اذا كان الزمن (t) يأخذ قيماً متقطعة (Discrete) اي $\{t=0, \pm 1, \pm 2, \dots\}$ فإن السلسلة الزمنية متقطعة في الزمن ويرمز لها $\{X(t); t = 0, \pm 1, \pm 2, \dots\}$.

2.2 نماذج السلاسل الزمنية Time Series Models

1.2.2 نماذج الانحدار الذاتي المتكامل والمتوسطات المتحركة (ARIMA)(p,d,q)

Autoregressive Integrated Moving Average Models

"تتصف معظم السلاسل الزمنية في الواقع العملي بخاصية عدم الاستقرار Non Stationary مما يجعل تحليل هذه السلاسل الزمنية والتنبؤ بها أمراً في غاية الصعوبة لذا يجب تحويلها الى سلاسل مستقرة باستعمال التحويلات اللازمة". وبأستعمال تحويلية الفرق وعند ادخال معامل الفرق d لنموذج ARMA(p,q) تتم اضافة Integrated الى اسم الانموذج لتصبح انموذج انحدار ذاتي ووسط متحرك متكامل Autoregressive Integrated Moving Average ويرمز له ARIMA(p,d,q) :

والصيغة العامة لنماذج ARIMA(p,d,q) كالآتي [19, 28] :-

$$\nabla Z_t = \phi_1 \nabla Z_{t-1} + \phi_2 \nabla Z_{t-2} + \dots + \phi_p \nabla Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \dots - \theta_q \varepsilon_{t-q}$$

$$\nabla Z_t = \sum_{n=1}^p \phi_n \nabla Z_{t-n} + \varepsilon_t - \sum_{n=1}^q \theta_n \varepsilon_{t-n} + \varepsilon_t \quad \dots \dots \dots (1)$$

وتعتبر منهجية بوكس جينكس (Box – Jenkins Method) من الطرق الاحصائية المهمة لتحليل السلسلة الزمنية (المستقرة وغير مستقرة والموسمية وغير موسمية) هي الطريقة التي اقترحها الباحثان (Box & Jenkins) عام 1970 وتتمثل هذه الطريقة بعملية بناء أنموذج ARIMA(p,d,q) لتمثيل ظاهرة معينة والتنبؤ بها. وتتلخص عملية بناء الانموذج حسب طريقة بوكس جينكس بالمراحل الآتية [9, 10, 11] :-

المرحلة الاولى: التحقق من استقرارية السلسلة الزمنية Stationary Check

وذلك برسم مخططٍ للسلسلة الزمنية ومخطط لكل من دالة الارتباط الذاتي ACF ودالة الارتباط الجزئي PACF لفحص استقرارية السلسلة الزمنية والاتجاه العام للسلسلة والتغيرات الموسمية أو عن طريق الاختبار الاحصائي جذر الوحدة لديكي وفولر (Augmented Dickey-Fuller Unit Root Tests) فإذا كانت السلسلة غير مستقرة يطبق عليها تعديل الفروق للحصول على الاستقرارية [19].

أختبار ديكي – فولار الموسع Augmented Dickey-Fuller test(ADF)

يعدّ اختبار جذر الوحدة لديكي – فولار الموسع (ADF) (Dickey & Fuller, 1989) من بين الإختبارات القوية للكشف عن استقرار السلسلة الزمنية، في نماذج بوكس وجينكس في حالة تكون جذور الوحدة كثيرة الحدود خارج دائرة جذر الوحدة إذا السلسلة تكون مستقرة، لذا فإن اختبار ديكي وفولار لتحديد جذر الوحدة يعمل بفرضية العدم لجذر الوحدة باستخدام الانموذج الآتي [19] :

$$\nabla y_t = \alpha + \beta_1 t + \beta_2 t^2 + \gamma y_{t-1} + \phi_1 \nabla y_{t-1} + \dots + \phi_{p-1} \nabla y_{t-p+1} + \varepsilon_t \quad \dots \dots (2)$$

اذ ان: ∇y_t : عامل الفرق حيث $\nabla y_t = y_t - y_{t-1}$ ، α : ثابت الاتجاه العام للأنموذج trend

β_1 : معامل اتجاه الوقت time trend ، p : عدد القيم المزاحة للأنحدار الذاتي AR ، ε_t : سلسلة عشوائية بوسط ثابت

يطبق الاختبار الاحصائي على قيمة γ المقدرة وتحسب قيمة (τ) كالآتي :

$$\tau = \frac{\hat{\gamma}}{\sigma\gamma} \dots\dots(3)$$

وبمقارنة القيمة المحسوبة لإحصائية (τ) مع القيم الجدولية الخاصة باختبار ديكي- فولار فإذا كانت القيمة المطلقة لإحصائية (τ) المحسوبة أكبر من قيمة (τ) الحرجة فإننا نرفض الفرضية الصفرية وتكون السلسلة الزمنية مستقرة ، أما إذا كانت قيمة (τ) المحسوبة أقل من قيمة (τ) الحرجة فإننا لا نرفض الفرضية الصفرية وفي هذه الحالة تكون السلسلة الزمنية غير مستقرة ، لتحديد الدرجة التي تصل السلسلة عندها الى الاستقرار فإننا نعيد الاختبار بعد اخذ الفرق الاول فإذا استقرت السلسلة فهذا يعني ان البيانات متكاملة من الدرجة الاولى وهكذا الى (d) من الاختبارات ."

المرحلة الثانية : تشخيص وتعريف الانموذج الذي يلائم البيانات Identification & Diagnostic

" إن مرحلة تشخيص نماذج السلاسل الزمنية هي مرحلة مهمة من مراحل بناء الانموذج باستعمال طريقة بوكس – جينكس لكونها تمكننا من معرفة سلوك السلسلة الزمنية والتعرف على خصائصها وذلك باستعمال البيانات التاريخية للسلسلة الزمنية والكشف عن الكيفية التي تولد بها هذه السلسلة ،" فبعد تحقيق أستقرارية السلسلة الزمنية يجب تحديد أفضل أنموذج بيانات للسلسلة الزمنية وتحديد قيم (p,d,q) ، وتحدد درجة الانموذج من خلال استعمال دالتي الارتباط الذاتي والارتباط الذاتي الجزئي من خلال ملاحظة سلوك الرسم البياني وقيم الدالتي" والجدول أدناه يوضح سلوك الدالتين [10 ,11] :-

جدول رقم(1): يمثل السلوك النظري لدوال الارتباط الذاتي للنماذج المختلفة

الانموذج	دالة الارتباط الذاتي ACF	دالة الارتباط الذاتي الجزئي PACF
AR(p)	يتناقص اسياً	ينقطع بعد الازاحة p
SAR(P)	يتناقص اسياً	ينقطع بعد الازاحة P
MA(q)	ينقطع بعد الازاحة q	يتناقص اسياً
SMA(Q)	ينقطع بعد الازاحة Q	يتناقص اسياً
ARMA(p,q)	يتناقص اسياً	يتناقص اسياً
SARMA(P,Q)	يتناقص اسياً	يتناقص اسياً

ولغرض الحصول على أفضل انموذج مقدر يمثل البيانات يستعمل عدة معايير المقارنة وان الانموذج الافضل يظهر عند اقل المعايير :-

معايير معلومات اكاكي (AIC) Akaike s' information criterin

اقترحه العالم الياباني هيروتوغو اكاكي سنة 1973 ويستخدم كأداة في تشخيص واختيار الانموذج الامثل ويمكن حسابه وفق الصيغة الآتية [19] :-

$$AIC = 2K - 2Ln(L) \dots (4)$$

حيث ان k عدد المعالم في parameters الانموذج ، L : تمثل اعلى قيمة لدالة الاحتمال Likelihood للانموذج وبعد إجراء التجارب على عدد من النماذج يكون الانموذج الافضل الذي له اقل قيمة AIC والذي يمكن استخدامه .

المرحلة الثالثة تقدير معلمات الانموذج المشخص Parameters' Estimation

بعد مرحلة تشخيص الانموذج الملائم للسلسلة الزمنية نقوم بتقدير المعلمات لهذا الانموذج المشخص والتي تتمثل في انموذج $AR(p)$ في $(\varphi_i, i = 1, 2, \dots, p)$ و $MA(q)$ في $(\theta_i, i = 1, 2, \dots, q)$ في انموذج $ARMA(p, q)$. أذ توجد عدة طرق لتقدير هذه المعلمات وأهم هذه الطرق طريقة الاماكن الاعظم Maximum Likelihood method التي تم الاعتماد عليها في هذا البحث وذلك بجعل مجموع مربعات الخطأ أقل ما يمكن [11, 30].

المرحلة الرابعة اختبار ملائمة الانموذج Diagnostic Checking

بعد ان تم التعرف على أنموذج مبدئي وأيجاد تقديرات معلمات الانموذج تأتي مرحلة مهمة هي مرحلة الاختبار اي اختبار مطابقة الانموذج ومدى كفاءته وملاءمة الانموذج ومدى صلاحيته لتمثيل بيانات السلسلة الزمنية وذلك بأستعمال اختبارات عدة منها اختبار طبيعية البواقي اي مدى تقارب البواقي المقدره الى التوزيع الطبيعي ، كذلك حساب دالة الارتباط الذاتي لسلسلة البواقي (residuals) فأذا كانت معاملات دالة الارتباط الذاتي للبواقي تقع ضمن مستوى ثقة معين فهذا يعني ان سلسلة البواقي عشوائية وبالتالي يمكننا القول ان النموذج المشخص ملائم ، كذلك يمكن استخدام اختبار بوكس وليونك Ljung-Box لمعرفة ملائمة الانموذج [28, 31].

في حال تطبيق هذه الاختبارات وكان الانموذج غير ملائم وسلسلة البواقي مترابطة فيجب الرجوع الى مرحلة التشخيص والانموذج الذي يلائم البيانات وبناء أنموذج اخر .

أختبار بوكس وليونك لسلسلة البواقي Test Ljung-Box For Residuals

طور الباحثان Box&Ljung في عام 1978 هذا الاختبار ويستخدم هذا الاختبار بعد مرحلة التشخيص والتقدير الانموذج المناسب للبيانات ، يحدد هذا الاختبار m من الارتباطات الذاتية للبواقي ، أذا كان الارتباط الذاتي قليل جدا نستنتج بأن الانموذج مناسب للبيانات ، وبشكل عام فإن اختبار Ljung-Box يكون وفق الفرضيات الآتية :-

H_0 : الانموذج يلائم البيانات

H_1 : الانموذج لا يلائم البيانات

يعرف الاختبار الاحصائي كالاتي :

$$Q = n(n + 2) \sum_{k=1}^m \frac{f^2}{n-k} \dots\dots\dots(5)$$

حيث f^2 تمثل قيم الارتباط الذاتي المقدر للسلسلة الزمنية عند الازاحة k و m تمثل عدد القيم المزاحة

الاختبار يرفض فرضية العدم اذا كان

$$Q > x_1^2 - \alpha, h$$

حيث ان $x_1^2 - \alpha$ هي قيم جدولية بتوزيع مربع كاي Chi-square بدرجة حرية ومستوى معنوية α

ولأن هذا الاختبار مطبق على سلسلة البواقي فإن درجة الحرية يجب حسابها على عدد معلمات الانموذج المقدر اي :

$$h = m - p - q , \text{ اذ ان } p, q \text{ تمثل درجة نماذج الانحدار الذاتي والمتوسط المتحرك .}$$

المرحلة الخامسة التنبؤ Forecasting

المرحلة الاخيرة من مراحل تحليل السلاسل الزمنية هي مرحلة التنبؤ ولا يمكن الانتقال الى هذه المرحلة الا بعد اتمام جميع الفحوصات الاحصائية الضرورية لتشخيص الانموذج الذي يتم اختياره كما مر ذكره في مراحل التحليل حيث ان هذه المرحلة هي مرحلة التطبيق العملي للانموذج المقترح من خلال الحصول على القيم التنبؤية المتوقعة للسلسلة وبعد تحديد درجات الانموذج التنبؤية المتوقعة للسلسلة وبعد تحديد درجات الانموذج (p,d,q) وتقدير الانموذج ومن ثم استخدامه للتنبؤ وذلك بأحلال القيم الحالية والماضية للمتغير التابع Z_t والبواقي α_t كقيمة تقديرية لحد الخطأ للحصول على القيمة المستقبلية الاولى Z_{t+1} وهو ما يسمى بالتنبؤ لفترة مستقبلية واحدة ، ويمكن الحصول على القيمة المستقبلية الثانية Z_{t+2} بأحلال القيمة المستقبلية الاولى Z_{t+1} التي تم النوصل اليها في الخطوة الاولى للتنبؤ في الطرف الايمن من المعادلة ، وهكذا حتى نصل الى الدالة المطلوبة مع افتراض ان حد الخطأ خارج العينة للدالة يساوي صفر أي ^[29].

معادلة التنبؤ للفترة الاولى هي :-

$$Z_{t+1} = \phi Z_t + \phi_1 Z_{t-1} + \alpha_t - \theta_1 \alpha_{t-1} \dots\dots\dots(6)$$

3.2 أنموذج متجه الانحدار الذاتي للسلاسل الزمنية المستقرة

(VAR) Stationary Vector Autoregressive Time Series Models

تعتبر نماذج متجه الانحدار الذاتي VAR من اكثر النماذج استخداماً في تحليل السلاسل الزمنية خصوصاً في المجالات الاقتصادية، وذلك بسبب سهولة تقدير معالمها عن طريق مقدرات لها صيغة صريحة ولا يتم حلها بالطرائق العددية وايضا ان صفات هذه النماذج تم دراستها بشكل مفصل من قبل العديد من الباحثين.

التعريف الرياضي لأنموذج:

يقال للسلسلة الزمنية بانها تتبع أنموذج متجه الانحدار الذاتي من الرتبة p، VAR(p)، اذا امكن التعبير عنها وفقا للانموذج التالي: [64]

$$z_t = \phi_0 + \sum_{i=1}^p \phi_i z_{t-i} + \alpha_t \dots\dots\dots(7)$$

اذ ان:

ϕ_0 : متجه من الثوابت ذي بُعد k.

ϕ_i : مصفوفات من الثوابت ذات بعد (k, k) لكل (i > 0) ، وتساوي مصفوفة اصفار عندما (i = 0)

α_t : هي سلسلة من المتجهات العشوائية المستقلة والمتماثلة التوزيع تتوزع بوسط حسابي يساوي متجه صفري ومصفوفة تباين Σ_a .

وباستخدام معامل الارتداد الخلفي فان الانموذج السابق سيكون:

$$\phi(B)z_t = \phi_0 + \alpha_t \dots\dots\dots(8)$$

اذ ان:

$$\phi(B) = I_k - \sum_{i=1}^p \phi_i B^i \text{ هي متعددة حدود من المصفوفات ذات الدرجة } p. \text{ [64]}$$

وعندما $p = 1$ ، فان الانموذج السابق يصبح:

$$Z_t = \phi_0 + \phi_1 Z_{t-1} + a_t \quad \dots\dots\dots(9)$$

وعلى فرض وجود مركبتان للسلسلة الزمنية ، اي (k = 2) ، فان الانموذج السابق يصبح:

$$\begin{bmatrix} Z_{1,t} \\ Z_{2,t} \end{bmatrix} = \begin{bmatrix} \phi_{1,0} \\ \phi_{2,0} \end{bmatrix} + \begin{bmatrix} \phi_{1,11} & \phi_{1,12} \\ \phi_{1,21} & \phi_{1,22} \end{bmatrix} \begin{bmatrix} Z_{1,t-1} \\ Z_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1,t} \\ a_{2,t} \end{bmatrix} \quad \dots\dots\dots(10)$$

وبشكل ابسط سيكون:

$$z_{1,t} = \phi_{1,0} + \phi_{1,11} z_{1,t-1} + \phi_{1,12} z_{2,t-1} + a_{1,t} \quad \dots\dots\dots(11)$$

$$z_{2,t} = \phi_{2,0} + \phi_{1,21} z_{1,t-1} + \phi_{1,22} z_{2,t-1} + a_{2,t} \quad \dots\dots\dots(12)$$

ان العنصر (1,2) في مصفوفة ϕ_1 ، والذي يمثل العنصر $\phi_{1,12}$ يقيس الاعتمادية الخطية ما بين السلسلة $Z_{1,t}$ على السلسلة $Z_{2,t-1}$ بوجود السلسلة $Z_{1,t-1}$. كما ان العنصر (2,1) في مصفوفة ϕ_1 ، والذي يمثل العنصر $\phi_{1,21}$ يقيس الاعتمادية الخطية ما بين السلسلة $Z_{2,t}$ على السلسلة $Z_{1,t-1}$ بوجود السلسلة $Z_{2,t-1}$. وهكذا لبقية العناصر.

1.3.2 أستقرارية الأنموذج:

ان الشرط الاساسي والضروري لجعل السلسلة الزمنية متعددة المتغيرات التي تتبع أنموذج متجه الانحدار الذاتي مستقرة هو ان تكون القيمة المطلقة للقيم الذاتية للمصفوفة ϕ_i اقل من الواحد الصحيح ، علما ان القيم الذاتية لتلك المصفوفة هي الحل لمعادلة المحدد التالية:

$$|\lambda I_k - \phi_i| = 0 \quad \dots\dots\dots(13)$$

2.3.2 تقدير المعلمات: Parameter Estimation

لغرض تقدير معلمات أنموذج متجه الانحدار الذاتي VAR(P) ، توجد العديد من الطرائق "المربعات الصغرى ، العزوم، الامكان الاعظم والطرائق البيزية" ، وتعتبر طريقة الامكان الاعظم من اكفاء الطرائق في التقدير. [64] فاذا توفرت عينة من البيانات $\{Z_t \mid t = 1, 2, \dots, T\}$ من سلسلة زمنية متعددة المتغيرات تخضع لانموذج VAR(P) . فان الهدف الاساسي من عملية التقدير هو الحصول على القيم العددية لمصفوفة المعلمات $\{\phi_0, \phi_1, \dots, \phi_p\}$ و مصفوفة التباين المشترك Σ_a .

3.3.2 طريقة الامكان الاعظم: Maximum Likelihood Method

تعتمد هذه الطريقة على فرض ان متجه الاخطا العشوائية لانموذج VAR(P) يتوزع وفقاً للتوزيع الطبيعي متعدد المتغيرات، وعلى فرض $Z_{h:q}$ ان تمثل المشاهدات من الزمن h الى الزمن q . وبذلك فان دالة الامكان الشرطية للبيانات سنكون كالآتي:

$$\begin{aligned} L(Z_{(p+1):T} \mid Z_{1:p}, \beta, \Sigma_a) &= \prod_{t=p+1}^T P(Z_t \mid Z_{1:(t-1)}, \beta, \Sigma_a) \\ &\propto |\Sigma_a|^{-(T-p)/2} \exp \left[-\frac{1}{2} \sum_{t=p+1}^T \text{tr}(a_t' \Sigma_a^{-1} a_t) \right] \quad \dots\dots\dots(14) \end{aligned}$$

نلاحظ ان الدالة السابقة كانت بدلالة الاخطاء العشوائية التي تتوزع توزيع طبيعي متعدد المتغيرات ومعلمات الانموذج . ان مقدر الامكان الاعظم هو قيمة معلمة الانموذج التي تجعل دالة الامكان في نهايتها العظمى، وللوصول لهذا الهدف

سوف يتم اخذ اللوغاريتم الطبيعي لدالة الامكان ومن ثم اشتقاقها لمتجه المعلمات ومساواتها بالصفر للوصول الى حل الذي يمثل قيم تلك المعلمات، وكما يلي:

$$l(\beta, \Sigma_a) = c - \frac{T-p}{2} \log(\Sigma_a) - \frac{-1}{2} \sum_{t=p+1}^T tr(a_t' \Sigma_a^{-1} a_t) \quad \dots\dots\dots(15)$$

لغرض الحصول على مقدر متجه المعلمات فان الجزء الاخير فقط في المعادلة السابقة سوف يستخدم للتعظيم والذي يناظر تقليل مجموع المربعات للخطأ، وكما يلي:

$$l(\beta, \Sigma_a) = c - \frac{T-p}{2} \log(\Sigma_a) - \frac{1}{2} S(\beta) \quad \dots\dots\dots(16)$$

وباجراء الاشقاق للمعادلة السابقة ومساواتها بالمتجه الصفري واعادة الترتيب فان مقدر متجه المعلمات سيكون:

$$\hat{\beta} = (X'X)^{-1}(X'Z) = [\sum_{t=p+1}^T x_t x_t']^{-1} \sum_{t=p+1}^T x_t Z_t' \quad \dots\dots\dots(17)$$

اذ ان:

$$x_t = (1, z_{t-1}', \dots, z_{t-p}') \text{ متجه بيانات ذي بعد } (kp+1).$$

$$\beta' = [\phi_0, \phi_1, \dots, \phi_p] \text{ مصفوفة معلمات ذات بعد } (k, kp+1).$$

4.3.2 اختيار رتبة نموذج متجه الانحدار الذاتي (VAR(P) Order Selection)

لغرض اتباع منهجية بوكس - جينكس في تحليل السلاسل الزمنية ، فان اختيار رتبة الانحدار الذاتي تعتبر خطوة مهمة، هنالك العديد من الطرائق منها يعتمد على اختبار النسبة الاحتمالية المتسلسلة مثل معيار(اختبار نسبة الاحتمال المتسلسل) ومنها ما يعتمد على مقياس المعلومات منها(معيار اكيائي ، معيار بيز ، معيار اكيائي المطور). وكما يلي: [21]

1- معيار نسبة الاحتمال (اختبار نسبة الاحتمال المتسلسل) Sequential Likelihood Ratio Test

استخدم من قبل العالمان تايبو و بوكس عام 1981، الفكرة الاساسية له هي مقارنة الانموذج عند رتب متسلسلة $VAR(l), VAR(l-1)$ ، وتوظيف ذلك لاختبار الفرضية العدم التالية: $H_0: \phi_l = 0$ مقابل الفرضية البديلة $H_a: \phi_l \neq 0$

وبافتراض ان متجه الاخطا العشوائية يتوزع طبيعيا فان احصاءة الاختبار للفرضية السابقة سيكون:

$$M(l) = -(T - l - 1.5 - kl) \ln \left(\frac{|\hat{\Sigma}_{a,l}|}{|\hat{\Sigma}_{a,l-1}|} \right) \quad \dots\dots\dots(18)$$

والتي تحت فرضية العدم سوف تتوزع توزيع كاي سكوير بدرجة حرية (k^2) .

2- معيار المعلوماتية Information Criteria

اظهرت هذه المعايير كفاءة عملية واضحة في اختيار رتب نماذج السلاسل الزمنية ، وجميع هذه المعايير تتكون من جزئين الاول يقيس جودة مطابقة الانموذج اي الاحتمالية العظمى وفي حالة التوزيع الطبيعي فانها تمثل محدد مصفوفة التباين المشترك. والجزء الثاني يمثل مقدار الجزء المترتب على تعقيد الانموذج والذي يتناسب مع حجم البيانات وعدد

معلمات الانموذج. وعموماً وبافتراض ان متجه الاخطا العشوائية يتوزع طبيعياً فهناك ثلاثة معايير شائعة الاستخدام ضمن نماذج VAR(I)K، هي:

1. معيار اكيائي المعلوماتي AIC : Akaike information criterion

وتم اشتقاقه من قبل العالم Akaike سنة 1973 وياخذ الصيغة التالية:

$$AIC(l) = \ln|\hat{\Sigma}_{a,l}| + \frac{2}{T}lk^2 \quad \dots\dots\dots(19)$$

5.3.2 فحص الانموذج: Model Checking

ويسمى ايضا الفحص التشخيصي او تحليل البواقي ، ويستخدم للتأكد من كون البواقي المقدره تتصف بالاستقلالية وتخضع للتوزيع الطبيعي ، اي ان الانموذج المقدر استطاع الكشف عن الحركة الديناميكية للبيانات بشكل كبير. ويعتبر اختبار Multivariate Portmanteau Statistics احد الشائعة لانجاز تلك المهمة من خلال اختبار الفرضية الصفرية

$$H_0: R_1 = \dots = R_m = 0 \text{ مقابل الفرضية البديلة التالية: } H_a: R_j \neq 0, 1 \leq j \leq m$$

اذ ان R_l تمثل الدالة النظرية للارتباط المشترك لمتجه الخطأ عند الفرق (l). ان احصاء الاختبار المناسبة للفرضية السابقة ستكون:

$$Q_k(m) = T^2 \sum_{l=1}^m \frac{1}{T-l} tr(\hat{R}_l' \hat{R}_0^{-1} \hat{R}_l \hat{R}_0^{-1}) \quad \dots\dots\dots(20)$$

3. نماذج الشبكات العصبية الاصطناعية (ANN) Artificial Neural Networks Models

1.3 مفاهيم أساسية للشبكات العصبية الاصطناعية Basic Concepts of ANN

سميت الشبكات العصبية الاصطناعية بهذا الاسم كونها شبكة من وحدات الإتصالات الداخلية، أذ ان هذه الوحدات مستوحاة من دراسة انظمة الاعصاب الحيوية ويرمز لها اختصاراً ب ANN وهو اختصاراً ل Artificial Neural Network وهي احدى تطبيقات الذكاء الاصطناعي (AI) .

أن فكرة عمل الشبكات العصبية الاصطناعية تتمثل بعملية محاكاة البيانات للوصول إلى نموذج لهذه البيانات لغرض التحليل او التصنيف او التنبؤ او اي معالجة اخرى دون الحاجة إلى نموذج مقترح لهذه البيانات ، لذلك فقد حازت الشبكات العصبية الاصطناعية أهتمام الكثير من الباحثين والعلماء أذ تمتلك مرونة كبيرة بالمقارنة مع الاساليب الرياضية المستعملة في عملية التعليم للانموذج البيانات وخرن المعلومات وبتها في الشبكة العصبية الاصطناعية^[2] .

2.3 مكونات الخلية العصبية الاصطناعية Components of Artificial Neural

تتكون الخلية العصبية الاصطناعية الواحدة ن اربعة أقسام وهي كالآتي^[2] :

- 1- قنوات الإدخال (synapse) : تستلم الخلية العصبية من هذه القنوات الإشارة القادمة من الخلايا المرتبطة بها وتسمى الاشارات الداخلة بالمداخلات ويرمز لها $(x_i, i=1,2,3,\dots,n)$.
- 2- دالة الجمع (Summation Function) : مهمة هذه الدالة الجمع او توحيد الاشارات الداخلة في إشارة واحدة
- 3- دالة التنشيط (Activation Function) : مهمة هذه الدالة نشر القيمة الداخلة إليها على وفق نوع الدالة المستعملة وفق مقياس حدود القيمة التي تخرج .
- 4- قناة الإخراج (Axone paths) : تقوم هذه القناة بأرسال إشارة الإخراج الى خلايا عصبية اخرى وتعد إشارة ادخال لتلك الخلايا وتسمى إشارة الإخراج بالمخرجات.

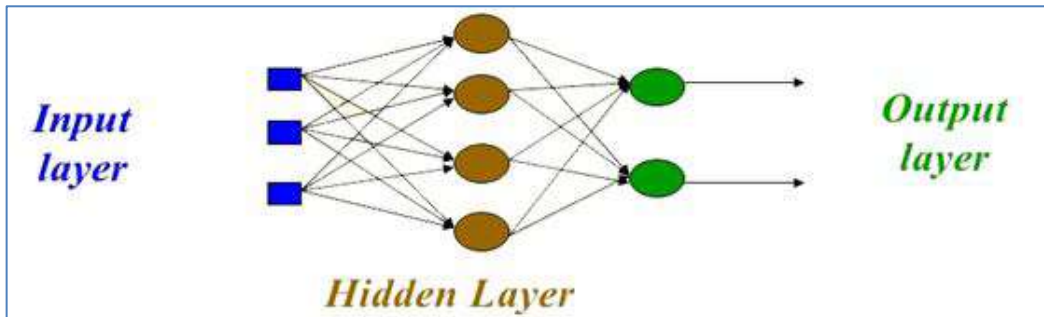
1.2.3 دوال التنشيط (التحويل) Activation Function

دوال التنشيط عبارة عن دوال لمعالجة البيانات الداخلة إلى الوحدات أو العقد العصبية، وهناك العديد من دوال التنشيط التي قدمت من قبل الباحثين والتي اختلفت تبعاً لأختلاف نوعية المخرج الذي يتم الحصول عليها واهداف الشبكة المراد تحقيقها . وأهم هذه الدوال المستخدمة في مجال التنبؤ الدالة اللوجستية ، الدالة ثنائية القطب .

3.3 الشبكة العصبية المتعددة الطبقات أمامية التغذية Multi-layer –perception Feed – Forward Neural Network (MLP-FFNN) :

وهي تسمح لأشارة بالانتقال الى الامام فقط من المدخلات الى المخرجات ومن ثم اضافة اكثر من طبقة مخفية (Hidden Layer) حيث لكل طبقة مخفية تحتوي على عدد من العقد المخفية (Hidden Nodes) تكون بين طبقة المدخلات وطبقة المخرجات، وسميت بهذا الاسم لانها لاتتصل بالمحيط الخارجي ومرتبطة فقط بالطبقة التي تسبقها والتي تليها بينما طبقة المدخلات وطبقة المخرجات تتصل بالمحيط الخارجي .

كذلك يمكن لهذه الشبكات التعامل مع مهام التعليم الغير خطي، والشكل (2-9) يوضح هيكل هذه الشبكة كما ان التصميم له دور مهم جداً في تحديد كفاءة ودقة عمليات التدريب حيث تتألف من ثلاث مستويات [44],[45] [48],[56] :



الشكل (2): هيكل الشبكة العصبية متعددة الطبقات أمامية التغذية FFNN

المستوى الاول : مستوى المدخلات Input Level

يحتوي على عدد من المدخلات المطلوب معالجتها ، عندما تكون وحدات الادخال تمثل أزاحات المتغير المعتمد (Lags) حسب مفهوم السلاسل الزمنية فإن أنموذج الشبكة العصبية سيكون مشابهاً الى أنموذج الانحدار الذاتي الغير خطي Non-Linear Autoregressive Model

المستوى الثاني: مستوى المخفي Hidden Level

وهو المستوى الثاني وقد يتكون من طبقة مخفية واحدة أو اكثر والتي تضم عدد من عقد مخفية ولكل عقدة لها وزن يربطها مع المستوى السابق الاول (المدخلات) ووزن يربطها مع المستوى اللاحق (المخرجات) .

المستوى الثالث: مستوى المخرجات Output Level

وهو المستوى الاخير من الشبكة العصبية ويتمثل بمخرجات الشبكة حيث تقارن مع القيمة الحقيقية لعمليات التدريب . Training

ان عملية التعليم للشبكات العصبية الاصطناعية ANN تكمن في ثلاث خطوات مستمرة :

- 1- حساب مخرجات مؤقتة .
- 2- مقارنة هذه المخرجات مع الاهداف المطلوبة .
- 3- ضبط الاوزان وتكرار العملية .

1.3.3 خوارزمية تعليم الشبكة The Learning Algorithm

تمثل الاوزان المعلومات الاولية او الابتدائية التي ستعلم او تتدرب بها الشبكة ،لذا لايد من تحديث الاوزان خلال مرحلة التدريب ، من اجل هذا التحديث نستعمل عدة خوارزميات مختلفة حسب نوع الشبكة ومن اهم هذه الخوارزميات خوارزمية الانتشار العكسي للخطأ Back propagation Algorithm التي اقترحت في عام 1960 ولكن قدرت اهميتها حتى عام 1970 والتي تستعمل في تدريب الشبكات العصبية كاملة الارتباط وذات التغذية الامامية وغير خطية ومتعددة الطبقات ، وفي هذه الشبكة يتم ايجاد الخطأ ما بين المخرج المستهدف والفعلي ويرجع بهذا الخطأ عكسياً من الطبقة الاخيرة الى الطبقات المخفية ثم أخيراً الى طبقة المدخلات وأثناء هذا الارتداد او التغذية العكسية يتم تغيير او تحديث الاوزان في اتجاه الذي يجعل الخطأ يتناقص باتجاه الصفر :

4.3 خوارزمية الانتشار العكسي للخطأ Error Back propagation Algorithm

أن تدريب الشبكة بأستعمال الانتشار العكسي يتضمن ثلاث مراحل هي [20],[21] [23],[22] :

- مرحلة الانتشار الامامي للخطأ Forward Stage
- مرحلة الانتشار العكسي للخطأ Back pass Stage
- مرحلة تحديث او توليف اوزان الشبكة Update Weights Stage

ويمكن أيجاز منهجية او خوارزمية عمل هذه الشبكة بالخطوات الاتية :

اولاً : التغذية الامامية Feed Forward :-

1- كل وحدة ادخال $(X_i, i = 1, \dots, n)$ تستلم القيمة المدخلة X_i وترسل هذه القيمة الى جميع الوحدات في الطبقة المخفية .

2- كل وحدة مخفية $(j, j = 1, \dots, p)$ يتم جمع القيمة المدخلة بعد ضربها في الاوزان كما في المعادلة الاتية :

$$Z - in_j = V_{0j} + \sum_{i=1}^n x_i v_{ij} \quad \dots\dots(21)$$

اذ ان: $Z - in_j$: مدخل الشبكة للوحدات المخفية Z_j ، V_{0j} : قيم التحيز Bias للوحدات المخفية Z_j

x : قيم السلسلة المدخلة $(X_i, i = 1, \dots, n)$ ، عدد قيم السلسلة المدخلة

v_{ij} : الوزن الخاص للوحدات المخفية Z_j في المتغير الداخل x_i ، ومن ثم تطبيق دالة التنشيط التابعة لحساب القيمة المخرجة :

$$Z_j = f(Z - in_j) \quad \dots\dots(22)$$

حيث ان $f(Z - in_j)$ تمثل القيمة المخرجة لدالة التنشيط الخاصة بالوحدات المخفية Z_j ثم ترسل هذه القيمة الى طبقة المخرجات

3- لكل وحدة أخراج $(y_k, k = 1, \dots, m)$ تقوم بجمع القيمة المدخلة بعد ضربها في الاوزان كما موضح في المعادلة الاتية :

$$y_in_k = w_{0k} + \sum_{j=1}^p z_j w_{jk} \quad \dots\dots(23)$$

ومن ثم نطبق دالة التنشيط التابعة لحساب القيمة المخرجة

$$y_k = f(y_in_k) \quad \dots\dots(24)$$

حيث ان $f(y_in_k)$ تمثل القيمة المخرجة لدالة التنشيط الخاصة بوحدة الاخراج y_k

ثانياً: التغذية العكسية للخطأ -: Backpropagaion of error

1- كل وحدة أخراج $(Y_k, K = 1, \dots, m)$ تستقبل الهدف المطلوب (Desired Target) تبعاً للمدخل المدرب،
 زحساب قيمة معلومات الخطأ له (Error Information Term) :

$$\delta_k = (t_k - y_k) f'(y_in_k) \quad \dots\dots(25)$$

ومن خلال هذه القيمة يتم حساب الاتي :

- القيمة التصحيحية للأوزان الخاصة بوحدات الاخراج (لتستخدم في تحديث الازان لاحقاً w_{jk}) :

$$\Delta w_{jk} = \alpha \delta_k z_j \quad \dots\dots(26)$$

اذ ان: α : تمثل نسبة التعليم وهو رقم ثابت تتراوح قيمته بين (0-1)

- قيمة تصحيح قيمة التحيز الخاصة بوحدات الاخراج (تستخدم لتحديث قيمة التحيز للوزن لاحقاً w_{0k}) :

$$\Delta w_{0k} = \alpha \delta_k \quad \dots\dots(27)$$

ومن ثم إرسال قيمة معلومات الخطأ δ_k الى الوحدات في الطبقة المخفية (الطبقة الأدنى) .

2- كل الوحدة المخفية $(z_j, j = 1, \dots, p)$ تقوم بجمع قيم المدخلات (القيم الموجودة من الوحدات في الطبقة الاعلى) :

$$\delta_in_j = \sum_{k=1}^m \delta_k w_{jk} \quad \dots\dots(28)$$

تضرب في مشتقة دالة التنشيط الخاصة بالوحدات المخفية لحساب قيمة معلومات الخطأ :

$$\delta_j = \delta_in_j f'(z_in_j) \quad \dots\dots(29)$$

حيث ان: δ_j جزء من تصحيح الخطأ الذي يرجع الى الخطأ في الوحدة المخفية z_j

$f'(z_in_j)$ مشتقة دالة التنشيط

ومن قيمة معلومات الخطأ المستخرجة يتم حساب الاتي :

- قيمة تصحيح الازان للوحدات المخفية (تستخدم لتحديث قيم الازان v_{ij} لاحقاً) :

$$\Delta v_{ij} = \alpha \delta_j x_i \quad \dots\dots(30)$$

- قيمة تصحيح قيم التحيز للوحدات المخفية (تستخدم لتحديث قيم التحيز للاوزان v_{0j} لاحقاً) :

$$\Delta v_{ij} = \alpha \delta_j \dots\dots(31)$$

ثالثاً : تحديث الاوزان وقيم التحيز Update Weights and Biases :-

- لكل وحدة أخراج $(y_k, k = 1, \dots, m)$ تحدث قيم الاوزان وقيم التحيز لها $(j=0, \dots, p)$

$$w_{jk}(\text{new}) = w_{jk}(\text{old}) + \Delta w_{jk} \dots\dots(32)$$

- لكل وحدة مخفية $(z_j, j = 1, \dots, p)$ تحدث قيم الاوزان وقيم التحيز لها $(i = 0, \dots, n)$

$$v_{ij}(\text{new}) = v_{ij}(\text{old}) + \Delta v_{ij} \dots\dots(33)$$

5.3 الشبكة العصبية Extreme learning machine Neural Network (ELM)

هي طريقة تعليم تم اكتشافها مؤخراً لتدريب الشبكات العصبية أحادية الطبقة أمامية التغذية المتقدمة اخترعها العالم Guang-Bin Huang ، تم تطوير (ELM) للشبكات العصبية أمامية التغذية وطبقة مخفية واحدة. (SLFNs) في خوارزمية ELM ، يتم تعيين الارتباطات بين طبقة الإدخال والخلايا العصبية المخفية بشكل عشوائي وتبقى كما هي دون تغيير أثناء عملية التعلم. عن طريق تقليل وظيفة التكلفة من خلال نظام خطي. تم تقليل العبء الحسابي لـ ELM بشكل كبير حيث أن التكلفة الوحيدة هي حل النظام الخطي. اجتذب التعقيد الحسابي المنخفض قدرأ كبيراً من الاهتمام من جانب مجتمع البحث ، وخاصة بالنسبة لتطبيقات البيانات عالية الأبعاد والكبيرة.

1.5.3 خوارزمية التعليم (ELM)

عند الحصول على بيانات التدريب $A = \{(x_i, t_i)\}_{i=1}^N$ ، يمكن التعبير عن وظيفة الإخراج للشبكة العصبية ذات الطبقة المخفية الواحدة أمامية التغذية (SLFN) single hidden layer feedforward neural network مع L من الخلايا العصبية المخفية كما في المعادلة الآتية :

$$f(x_i) = \sum_{j=1}^L \beta_j h_j(a_j, b_j, x_i) = h(x_i)\beta, \quad i = 1, \dots, N \quad \dots\dots(34)$$

حيث ان :

$$\beta = [\beta_1, \dots, \beta_L]^T \text{ هي مصفوفة وزن الاخراج}$$

$$h(x_i) = [h_1(a_1, b_1, x_i), \dots, h_j(a_j, b_j, x_i)]$$

التدريب على الشبكة هو العثور على معاملات شبكة مناسبة لتقليل وظيفة الخطأ $|H\beta - T|$

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix}, \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} \quad \dots\dots(35) \text{ عندما :}$$

هي مصفوفة إخراج الطبقة المخفية ومصفوفة الهدف لبيانات التدريب على التوالي . يوضح الشكل التالي الرسم التوضيحي لـ SLFN مع الخلايا العصبية المخفية.

4. تقييم أداء طرائق التنبؤ Forecasting Evaluation

بعد اتمام عمليات التنبؤ تأتي مرحلة التقييم أي تقييم أداء طرائق التنبؤ لمعرفة أي النماذج هو الافضل لأستخدامه من خلال استخدام مقاييس التنبؤ، حيث ان خطأ التنبؤ هو مقدار الفرق بين المشاهدات للسلسلة الحقيقية والقيم التنبؤية للسلسلة الزمنية . مقاييس الخطأ التي استخدمت في هذا البحث [1],[2],[53] :-

1.4 متوسط مربع الخطأ (MSE) Mean Square Error

هو مجموع مربع الاخطاء (الفرق بين القيم الحقيقية والقيم التنبؤية للسلسلة) مقسم على عدد المشاهدات:-

$$MSE = \frac{\sum_{i=1}^n (Z - \hat{Z})^2}{n} \dots\dots\dots(36)$$

حيث ان \hat{Z} تمثل القيمة التنبؤية ، Z القيمة الحقيقية ، n عدد المشاهدات

5. الجانب العملي

تم الاعتماد على البيانات المأخوذة من وزارة البيئة ،اذ توجد هنالك اجهزة قياس خاصة لتلك البيانات في جانبي الرصافة والكرخ، ومن خلال تصدير البيانات الاسبوعية لثنائي اوكسيد النيتروجين NO2 و ثنائي أوكسيد الكبريت (SO2) الى المنصة التنفيذية لبرنامج R . اذ اظهرت نتائج التحليل الوصفي ان المتوسط تقريبا يساوي الوسيط اي ان البيانات تتوزع تقريبا طبيعياً ولكلا المتغيرين، وكما موضح في الجدول رقم(3).

جدول رقم(3): يوضح الاحصاءات الوصفية للبيانات

Variable Measure	NO ₂	SO ₂
Min	0.00800	0.00600
1st Qu	0.02600	0.02200
Median	0.03500	0.03300
Mean	0.03668	0.03552
3rd Qu	0.04400	0.04500
Max	0.08900	0.10100

ومن خلال تطبيق اختبار ديكي فولر الموسع Augmented Dickey-Fuller Test المستخدم لاكتشاف عدم الاستقرارية للسلسلة الزمنية، اذ كانت قيمة احصاء الاختبار معنوية اي اننا نقبل الفرضية البديلة التي تنص بان السلسلة الزمنية مستقرة ولا داعي لاخذ الفروق لجعلها مستقرة بالمتوسط ولكلا المتغيرين، وكما موضح في الجدول رقم (4)

جدول رقم (4): يوضح اختبار الاستقرار

Augmented Dickey-Fuller Test	
NO2	SO2
Dickey-Fuller = -8.7981, Lag order = 0, p-value = 0.01	Dickey-Fuller = -8.8957, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary	alternative hypothesis: stationary

ولاختبار استقرارية التباين تم تطبيق اختبار بوكس كوكس Box-Cox ، والذي يعتمد على تقدير المعلمة λ التي تمثل معلمة التحويل ، فإذا كانت تساوي صفر فيتم اخذ التحويل اللوغاريتمي وإذا كانت تساوي واحد فلاداعي للتحويل وغير ذلك يتم اعتماد المعادلة رقم () لاجراء التحويل المناسب على البيانات وجعل تباينها مستقر. ومن خلال تطبيق الخوارزمية الخاصة بالاختبار تبين ان قيمة λ المقدره تساوي 0.222 للمتغير NO2 و 0.2626263 للمتغير SO2 ، وطبقا لذلك سيتم اجراء التحويل المناسب.

من خلال مقارنة السلوك النظري لدالتي الارتباط الذاتي والارتباط الذاتي الجزئي مع نظيراتها المقدره وتطبيق معيار معلومات اكيباي AIC ومعيار معلومات اكيباي المطور AICc ومعيار بيز المعلوماتي BIC لعدد من الترتيب الممكنة فان الانموذج المرشح لتقدير بيانات المتغير NO2 المحولة ، سيكون انموذج الانحدار الذاتي من الرتبة الاولى $ARIMA(p = 1, d = 0, q = 0)$ و انموذج مختلط من الرتبة الثالثة للانحدار الذاتي والثالثة للاوساط المتحركة $ARIMA(p = 3, d = 0, q = 3)$ للمتغير NO2 المحولة.

وبالاعتماد على طريقة الامكان الاعظم لتقدير معاملات انموذج المتغير NO2 والبالغ عددها اربع معاملات (واحدة لانموذج الانحدار الذاتي وواحدة لانموذج الاوساط المتحركة والمتوسط العام وتباين حد الخطأ) وبالاعتماد على البيانات التي تم تحويلها بالخطوة السابقة ، اذ تعتمد هذه الطريقة على طرائق الحل العددي للحصول على قيم المقدرات ، وبعد ادراج فقط المقدرات ذات التأثير المعنوي كانت النتائج كما يلي:

$$x_t = -0.5458 + 0.7694x_{t-1} + \hat{w}_t - 0.5131\hat{w}_{t-1}$$

$$(0.0001) \quad (0.0433) \quad (0.00006)$$

إذ إن القيم بين الاقواس تمثل قيم الاحتمالية المرافقة للاختبار المعنوية للمعاملات (P_value) ومنه نلاحظ ان جميعها معنوية .

وان تقدير حد الخطأ هو $\hat{\sigma}^2 = 0.02868$ ، بدرجة حرية = 154

$$AIC=-2.513493 \quad , \quad AICc=-2.499078 \quad , \quad BIC= -3.455094$$

وبالاعتماد على طريقة الامكان الاعظم لتقدير معاملات انموذج المتغير SO2 والبالغ عددها ثمانية معاملات (ثلاثة لانموذج الانحدار الذاتي وثلاثة لانموذج الاوساط المتحركة والمتوسط العام وتباين حد الخطأ) وبالاعتماد على البيانات التي تم تحويلها بالخطوة السابقة ، اذ تعتمد هذه الطريقة على طرائق الحل العددي للحصول على قيم المقدرات ، وبعد ادراج فقط المقدرات ذات التأثير المعنوي كانت النتائج كما يلي:

$$x_t = -3.83663 + -0.9029x_{t-2} + \hat{w}_t + 1.0526\hat{w}_{t-2}$$

$$(0.0001) \quad (0.0433) \quad (0.00006)$$

اذ ان القيم بين الاقواس تمثل قيم الاحتمالية المرافقة للاختبار المعنوية للمعاملات (P_value) ومنه نلاحظ ان جميعها معنوية .

وان تقدير حد الخطا هو $\hat{\sigma}_w^2 = 0.0356$ ، بدرجة حرية = 150

AIC=-2.246184 , AICc=-2.227248 , BIC= -3.2099218

1.5 التنبؤ بأنموذج السلاسل الزمنية متعددة المتغيرات Multivariate Time Series Analysis model

يتم في من خلال توظيف الارتباط الموجود بين المتغيرين لتحسين دقة التنبؤات ومن خلال المخطط الزمني Time Plot للسلسلتين، تبين وجود نوع من التزامن او الاقتران بين السلسلتين وهذا يؤشر الى امكانية اجراء النمذجة الرياضية للسلسلتين في آن واحد.

وبالاعتماد على مصفوفة الارتباط المتقاطع بين السلسلتين ($z_{1,t} = NO_2, z_{2,t} = SO_2$) وبأخذ الفرق الزمني لغاية الفرق (12) ، نلاحظ من خلال دالة الارتباط المشترك وجود ارتباط معنوي عند الفرق الاول والثالث والرابع مما يؤشر وجود امكانية للاستفادة من تحليل السلسلتين في آن واحد وكما وموضح في اختبار مخطط اختبار المعنوية مقابل فرضية الضوضاء البيضاء White Noise .

من خلال تطبيق اربعة معايير لاختيار رتبة أنموذج متجه الانحدار الذاتي وهي (مقياس اكيكي المعلوماتي aic ، معيار bic، معيار HQ و معيار M(p) تبين ان اقل قيمة للمعايير الاربعة كانت عند الرتبة الاولى وباجراء اختبار للمعنوية كانت اعلى رتبة معنوية هي الرتبة الاولى.

ويعد معرفة رتبة متجه الانحدار الذاتي سوف يتم الاستعانة بطريقة الامكان الاعظم لتقدير معاملات الانموذج وهي متجه الثوابت ومتجه مصفوفة المعلمات ومصفوفة التباين المشترك لحد الخطاء مع الاخطاء المعيارية لتلك المقدرات ، اذ سيكون الانموذج كما يلي:

$$Z_t = \begin{bmatrix} 0.0241205 \\ 0.01746689 \end{bmatrix} + \begin{bmatrix} 0.320 & 0.022 \\ 0.361 & 0.137 \end{bmatrix} Z_{t-1} + a_t$$

كما ان مصفوفة التباين المشترك لمتجه الانحدار الذاتي حد الخطاء ستكون:

$$\hat{\Sigma}_a = \begin{bmatrix} 0.0001790282 & 0.0001522548 \\ 0.0001522548 & 0.0002782629 \end{bmatrix}$$

كما ان محدد مصفوفة التباين المشترك لمتجه الانحدار الذاتي كان: $2.663535e-08$

2.5 التنبؤ بأنموذج الشبكات العصبية الاصطناعية Artificial Neural Networks Models

NO2

سيتم توظيف اسلوب الشبكات العصبونية التي تعتمد خاصية الانتشار الامامي feed-forward، حيث استخدام خوارزمية multilayer perceptrons يتم ابتداء تحويل البيانات الى قيم محصورة بين الصفر والواحد الصحيح ومن ثم افتراض عدة فرق لمداخلات الشبكة لكي يتم اختيار افضلها، وبافتراض عدة طبقات مخفية اقل من 10 طبقات فان افضل معمارية كانت طبقة مخفية واحدة تتكون من خمس عقد علما ان الاوزان الأولية تم اختيارها عشوائيا.

ويعد اجراء عملية التدريب فان متوسط مربعات الخطاء كان (MSE= 0.0001488946)، وباستخدام خوارزمية extreme learning machines (ELM) فان ان متوسط مربعات الخطاء كان (MSE=0.000200202) ، اما بالنسبة للملوث SO2 باستخدام طريقة MLP ، وبعد اجراء عملية التدريب فان

متوسط مربعات الخطاء كان ($MSE= 0.0002777279$)، وباستخدام خوارزمية extreme learning machines فان علما ان متوسط مربعات الخطاء كان ($MSE= 0.0003217693$)
 استخدمت مقاييس الخطأ الاتية التي تم التطرق اليها في الجانب النظري لتقييم اداء طرق التنبؤ المستخدمة في هذا البحث :

جدول (5): يوضح نتائج معايير الخطأ للنماذج التي طبقت بهذا البحث

MODEL	MSR		RMSR	
	SO2	NO2	SO2	NO2
ARIMA	0.0356	0.02868	0.596657	0.169352
VAR	0.000278262	0.00017902	0.000278262	0.0133798
ANN(ELM)	0.000321769	0.000200202	0.017937921	0.0141493
ANN(MLP)	0.000277727	0.000148894	0.016665143	0.0122022375

6. مناقشة النتائج:

بيّض من معايير الخطأ في الجدول اعلاه أن تقنيات تنقيب البيانات التي تخص نماذج الشبكات العصبية المستخدمة بهذا البحث تفوقت على النماذج الاحصائية المستخدمة ، وذلك لأن الشبكة تحدث أوزانها في كل مرة يتم تدريب الشبكة حتى تتعلم وتتكيف بشكل يساعد على أيجاد القيم المقدرة بشكل دقيق بينما في النماذج الاحصائية في مرحلة التقدير تتم بتقدير المعلمة مرة واحدة على جميع السلسلة ، هذا بالإضافة على قدرة الشبكات العصبية على التكيف مع سلوك السلسلة بشكل اكثر مرونة من الانماذج الاحصائية التي تفترض شكل مسبق قد يكون في اغلب الاحيان بخلاف واقع السلسلة وبالتالي فان علم تنقيب البيانات يذهب باتجاه ان البيانات هي التي تفقد التحليل وليس العكس.

7. الاستنتاجات:

1. اظهرت نتائج التحليل بوجود ارتباط متقاطع معنوي بين السلسلتين وبالتالي ظهرت الحاجة الى توظيف نماذج السلاسل الزمنية متعددة المتغيرات التي تأخذ بنظر الاعتبار المعلومات المشتركة بين السلسلتين في النمذجة الرياضية والتي تنعكس على دقة التنبؤات .
2. اظهرت نتائج التحليل ان افضل هيكلية للشبكة (الطبقات ، نسبة التعلم ، الاوزان ، التكرار) كانت عدد الطبقات = 5 .
3. اظهرت نتائج التحليل بتفوق طرق تنقيب البيانات (ANN(MLP – ELM) على الطرق الاحصائية ARIMA و VAR
4. اظهرت نتائج التحليل بتفوق طريقة ANN(MLP) على طريقة ANN(ELM) .
5. اظهرت نتائج التحليل تفوق طريقة VAR على طريقة ARIMA .
6. ان نتائج قيم متوسط الاخطاء اظهرت تقارب طريقتي ANN مع طريقة VAR .

المصادر

1. شومان، عبد اللطيف حسن (2008). مقدمة في الاحصاء التطبيقي ، الطبعة الاولى – دائرة المكتبة الوطنية، المملكة الاردنية الهاشمية ، رقم الايداع 2008/1/7 ، أخرج وتدقيق دار الجنان .
2. عاشور، مروان عبد الحميد (2018) . الشبكات العصبية الاصطناعية وطرائق تنبؤ السلاسل الزمنية . الطبعة الاولى – دار الكتب والوثائق ، بغداد ، رقم الايداع 2018/248 (439) ص.
- [2]. **R. TSAY**. "Analysis of Financial Time Series Financial Econometrics" .John Wiley&Sons, Inc, USA .2005.
- [3] Wei,W.S., "Time series Analysis Univariate and Multivariate Methodes " , Addison-Wesly publishing company INC, USA.2006 .
- [4]. Box, G. E. P, Jenkins, G. M., and Reinsel, G. C. , "Time Series Analysis: Forecasting and Control" ,4rd Ed., Holden-Day, San Francisco.2008 .
- [5]. Brockwell, P. J and Davis, R. A. "Introduction to Time Series and Forecasting" , Springer-Verlag , New York.1996..

- [6]. Singer, Michael Bliss and Thomas Dunne. "Identifying eroding and depositional reaches of valley by analysis of suspended sediment transport in the Sacramento River, California." Water Resources Research vol.37.issue.12 .2001.
- [7] Moyo, Pilate, and James MW Brownjohn."Application of Box-Jenkins models for assessing the effect of unusual events recorded by structural health monitoring systems." Structural Health Monitoring 1.2 .2002.
- [8]. Montgomery,Douglas C., Cheryl L. Jennings, and Murat Kulahci. "Introduction to time series analysis and forecasting". John Wiley& Sons, 2015 .
- [9]. Box,George EP, et al. "Time series analysis: forecasting and control". John Wiley & Sons, 2015 .
- [10]. Box, George EP, and Gwilym M. Jenkins. "Time series analysis: forecasting and control" , revised ed. Holden-Day, 1976 .
- [11]. Shumway, Robert H., and David S. Stoffer. "Time series analysis and its applications: with R examples" . Springer Science & Business Media,2010 .
- [12]. Edlund, Per-Olov. "Identification of the multi-input box-Jenkins transfer function model." Journal of Forecasting 3.3 .1984.
- [13]. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques, Second Edistion Elsevier, 2011 .
- [14]. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques, Third Edistion Elsevier, 2014 .
- [15]. Akintola, K. G., B. K. Alese, and A. F. Thompson. "Time Series Forecasting with Neural Network: A Case Study of Stock Prices of Intercontinental Bank Nigeria." IJRRAS Dec.2011 .
- [16]. Al Aboodi, Ali H. "Prediction of Tigris River Discharge In BAGHDAD City Using Artificial Neural Networks ." Kufa journal of Engineering 5.2 .2014.