المجلة العراقية للعلوم الإحصائية (29) 2019
عدد خاص بالمؤتمر الطلابي الاول
ص ص [37-46]

استخدام الأسلوب MLR–RNN للتكهن ببيانات التلوث الجوي

د.أسامة بشير شكر الحنون*
drosamahannon@gmail.com

مريم منيب محمد يحيى**
drosamahannon@gmail.com

المستخلص

ان نمذجة جودة الهواء اكتسبت أهمية كبيرة في تلوث الهواء الجوي بسبب الآثار السلبية على البيئة وصحة الإنسان. وفي هذه الدراسة تم التطرق الى ملوثات الهواء التي كان لها تأثير مباشر على بيانات PM_{10}. ان بيانات التلوث الجوي والارصاد الجوية عموماً تأخذ نمطاً غير خطياً حسب تجارب سبقت في هذا المجال مما يؤدي الى ظهور نتائج وتكهنات غير دقيقة في حالة استخدام نماذج خطية مثل نماذج الانحدار الخطي المتعدد. ان استخدام طرق غير خطية مثل الشبكات العصبية المعاودة (RNN) قد يؤدي الى تحسين ملحوظ في نتائج التكهن والتحليل. في هذه الدراسة تم اقتراح استخدام نموذج الانحدار الخطي المتعدد الأكثر شيوعا لدراسة البيانات متعددة المتغيرات كما وتم اقتراح استخدام الشبكات العصبية المعاودة لمعالجة مشكلة عدم خطية البيانات مما يؤدي الى نتائج اكثر دقة. وكذلك اقترح استخدام الطريقة الهجينة MLR-RNN. ومن خلال تقليل عدد المتغيرات التفسيرية فقد حسنت طريقة RNN أداء MLR ولذلك حسنت الطريقة الهجينة MLR-GA نتائج التحليل والتكهن بشكل افضل.

الكلمات المفتاحية: أسلوبMLR–RNN ،الانحدار، التلوث الجوي

Using the Hybrid MLR-RNN Approach for Air Pollution Forecasting

ABSTRACT

Air Quality Modeling gained great importance in atmospheric pollution because of its negative effects on the environment and human health. In our study, the relationship between (Particulate Matter PM_{10}) and other nine variables over three years is studied to applied the multiple linear regression models (MLR). The MLR model is the most common for studying like this multivariate case. The main problem for this type of data is the non linear style that has been referred by many researchers before. The recurrent neural network (RNN) is nonlinear method

*مدرس/ قسم الاحصاء والمعلوماتية / كلية علوم الحاسوب والرياضيات / جامعة الموصل
** طالبة بكلوريوس / قسم الاحصاء والمعلوماتية / كلية علوم الحاسوب والرياضيات / جامعة الموصل

تاريخ استلام البحث 2019/4/29 ـــــــــــــــــــــ تأريخ القبول 201 9/7/4

which can be used to solve the nonlinearity problem and resul better forecasting. The hybrid method MLR–RNN can be used also for the best results and lead to more accurate forecasting. The hybrid method MLR–RNN has improved the performance of MLR method separately.

1- المقدمة

يعرف نموذج MLR بانه احد الاساليب الاحصائية للتحقق من العلاقة بين المتغيرات، ويستخدم لتشكيل نموذج للعلاقة بين المتغير المعتمد والعديد من المتغيرات التفسيرية وفي الأرصاد الجوية، فعادة ما تكون بيانات الارصاد الجوية متعددة المتغيرات وكبيرة الحجم ان كانت يومية أو أسبوعية أو أنماط موسمية اخرى، او بيانات متعددة المحطات مما يؤدي الى صعوبة التكهن بالبيانات المستقبلية(Güler and Güneri İşçi, 2016). فضلا عن استخدام الأنماط الموسمية أو الأسبوعية ومواجهة الآثار السلبية على البيانات ونتائج التحليل.

واجه الكثير من الباحثين في الإحصاء صعوبة في استخدام نماذج الانحدار المتعدد، فتم اللجوء الى استخدام بعض الطرق للحد من عدد المتغيرات التفسيرية والوصول إلى أفضل الحلول، فاستخدموا عددا من الطرق منها التقليدية مثل تحليل المكونات الرئيسة وتحليل السلاسل الزمنية، واخرى حديثة مثل اسلوب الذكاء الاصطناعي كالشبكات العصبية المعاودة Reccurrent Neural Networks (RNN) وفي الواقع كان الأداء أفضل عند استخدام الأساليب الحديثة بالمقارنة مع استخدام الطرق التقليدية.

في دراسة بيانات الارصاد الجوية غالبا ما يستخدم اسلوب الانحدار الخطي المتعدد الذي يعد أحد النماذج الاحصائية الخطية و يؤدي ذلك الى تكهن غير دقيق وصعوبة في تحديد البيانات المدروسة في الماضي وفي حساب النماذج و دراستها وتحليلها وذلك للطبيعة غير الخطية لمثل هكذا نوع من البيانات.

الهدف من هذه الدراسة هو تحسين أداء الطرق التقليدية من خلال اقتراح MLR و RNN وطريقتهما الهجينة (MLR-RNN).

2-الطرق المقترحة.

يستعرض هذا القسم الانحدار الخطي المتعدد MLR ونماذجه، واستخدام الخوارزمية الجينية GA والتي هي احدى انواع الذكاء الاصطناعي وطريقة MLR-GA الهجينة للحصول على نتائج وتنبؤات دقيقة.

2-1: الانحدار الخطي المتعدد (Multiple Linear Regression MLR)

يعرف تحليل الانحدار بشكل عام بأنه أسلوب رياضي لتوضيح العلاقة بين المتغير المعتمد ومتغير أو متغيرات أخرى تسمى المتغيرات التفسيرية. يهتم تحليل الانحدار بوصف العلاقة بين المتغيرات على هيئة نموذج فقد يكون النموذج يحتوي على متغير توضيحي واحد فيسمى في هذه الحالة بنموذج الانحدار البسيط ,أما في حالة احتواء النموذج على متغيرات توضيحية عدة (اثنين أو أكثر) فأنه يسمى بنموذج الانحدار المتعدد(Nathans *et al.*, 2012). استخدم العديد من الباحثين الانحدار الخطي المتعدد ويمكن صياغة النموذج العام لمعادلة MLR على النحو الاتي:

$$Y = B_0 + B_1X_1 + B_2X_2 \ldots\ldots + B_iX_i + E \quad (1)$$

حيث Y_i هو المتغير المعتمد و $X_1.X_2.\ldots.X_i$ هي المتغيرات التفسيرية و $B_0.B_1.\ldots.B_i$ هي معاملات الثوابت او معلمات الانحدار E هو حد الخطأ لاقل تنبؤ رقم i حيث i=1,2…,n حيث ان عدد المشاهدات هو n يكون عدد المعادلات هو n ويمكن صياغة المعادلة (1) في صورة مصفوفة كما يلي:

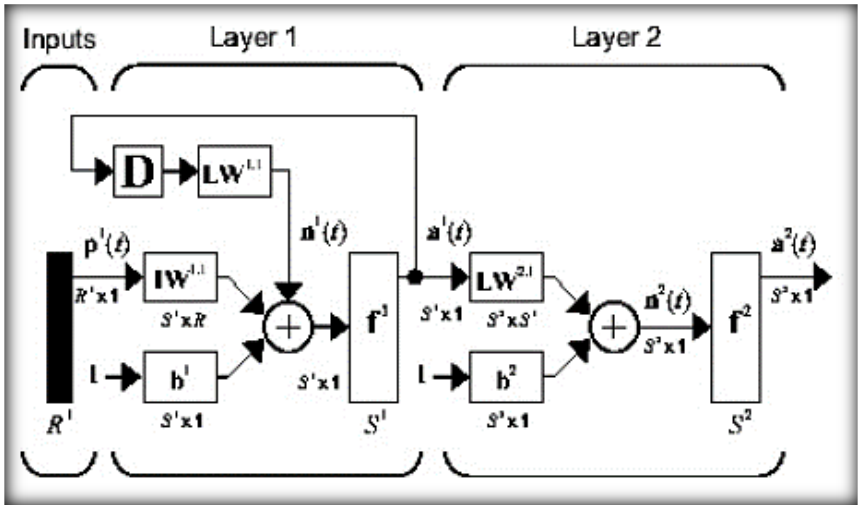
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ : \\ : \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1x_{11} & x_{12}....x_{1p} \\ .... & \\ 1x_{n1} & x_{n2}....x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ : \\ : \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ : \\ : \\ \varepsilon_i \end{bmatrix} \quad (2)$$

اذ Y حجم (n×1) و المصفوفة X من درجة ((n×(p+1)) وحجم β هي (p×1) و درجة $\mathcal{E}$ هي (i×1). ويلاحظ أن العمود الاول في مصفوفة البيانات يحتوي على قيمة الواحد الصحيح عند كل المشاهدات من (1) إلى (n) وذلك لتقدير المعامل الثابت والعمود الثاني من المصفوفة يحتوي على قيم المتغير الاول (X_1)، وبذلك كل عمود يحتوي على قيم متغير تفسيري محدد، وباستخدام رموز المصفوفات يمكن اختصار كتابة نموذج الانحدار الخطي كما يلي:

$$Y = \beta X + \varepsilon \quad (3)$$

2-2: الشبكات العصبية المعاودة RNN:

Hu (1964) هو اول من استعمل الشبكات العصبية في التكهن بالاحوال الجوية وبعده تم استعمالها كثيراً في مجال التكهن. السبب الاساسي وراء اللجوء الى الشبكات العصبية هي طبيعة البيانات غير الخطية. تحوي RNN على طبقة واحدة او اكثر وهذا بدروه يعالج غير خطية البيانات ويحسن نتائج التكهن وكذلك تحوي على Daley Layer وهذا يحسن كثيراً التعامل مع مشكلة عدم تجانس البيانات والشكل التالي يمثل الشبكة وما تحويه من ادخالات واخراجات وطبقات.



الشكل (1) :هيكلية الشبكة العصبية المعاودة RNN

في الشكل السابق R هي الادخالات و lW هي وزن عشوائي للعصبون يتم جمعها مع b1 التشويش الابيض وناتجهما يكون الدالة f1. اخراج الدالة f1 يعود كادخال ثالث في الطبقة الاولى وقبلها يمر على دالة التاخير Delay ليكون وزنا عشوائيا اخر ، في الطبقة الثانية اخراج الدالة f يكون الوزن العشوائي للخلية العصبية LW مجموع مع b التشويش الابيض وبالتالي تخرج لنا مصفوفة احادية.

تحوي RNN في هذه الدراسة على طبقتين بالاضافة الى طبقة الادخال، الاولى تكون مخفية والثانية تكون طبقة الاخراج. في طبقة الادخال هناك R من الادخالات وهذه الادخالات غالبا ما توزن عشوائيا في كل طبقة مخفية وكذلك M من العصبونات.عادة افضل عدد للعصبونات في الطبقة المخفية هو R*2+1 كما ذكره (Sheela and Deepa, 2013) (Palit and Popovic, 2006)

كل متغير ادخال Z موزون عشوائيا. وان اوزان الادخالات ل N و M عصبون تجمع مع القيمة المتحيزة b بواسطة دالة التحويل. مجموع ادخالات المتغيرات في دالة التحويل F يمكن صياغتها كما ياتي

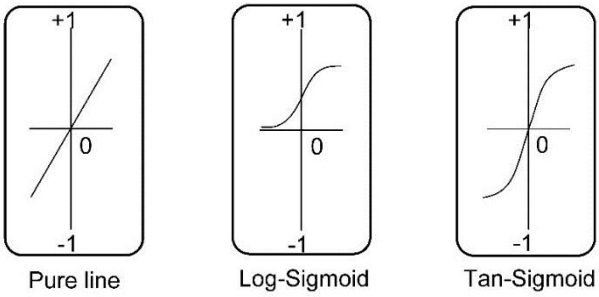
$$net_j(t) = \sum_{i=1}^{N} w_{i,j} Z_j(t) + b_j \qquad (4)$$

والدوال الاكثر استخداما في الطبقة المخفية وطبقة الاخراج هي كالتالي:

1. التحويل الزاوي (tan-sigmoid): الذي يولد اخراجات ضمن الفترة [−1،+1] للاحداثي الصادي.
2. التحويل اللوغارتمي(log-sigmoid): الذي يولد اخراجات ضمن الفترة [0,1]للاحداثي الصادي.

3. دالة التحويل الخطي (linear): الذي يولد اخراجات ضمن الفترة [−1،+1]للاحداثي الصادي.

الشكل في الوارد لاحقا يوضح الفرق بين دوال التحويل الممكن استعمالها للشبكة RNN في MATLAB



الشكل (2) :انواع دوال التحويل في RNN

الدوال المستعملة في الطبقة المخفية تعكس نوعية العلاقة بين الادخالات والاخراجات في حين تبنى الدوال المستعملة في طبقة الاخراج لتعطي افضل وادق النتائج.

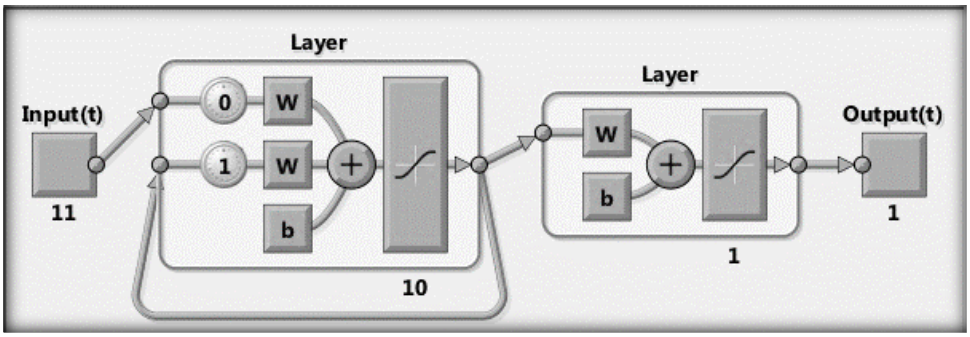
الاوزان العشوائية $w_{i,j}$ للادخالات يمكن كتابتها كمصفوفة وعلى النحو التالي

$$W=\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M,1} & w_{M,2} & \cdots & w_{M,R} \end{bmatrix} \quad (5)$$

في حين تصاغ معاملات الادخال كما يلي :

$$Z=\begin{bmatrix} Z_{t1} & Z_{t2} & \cdots & Z_{tR} \end{bmatrix}' \quad (6)$$

ان الهيكل العام للشبكة RNN يحوي على طبقة واحدة مخفية واخرى للاخرج ،المخفية تحوي على ثلاث مدخلات الوزن العشوئي والتشويش الابيض بالاضافة الوزن الناتج من الخطوة السابقة اما الخارجية فمدخلاتها الوزن الناتج من الطبقة المخفية مع التشويش الابيض كما في الشكل.



الشكل (3) : الهيكل العام للشبكة العصبية المعاودة RNN

3-2: طريقة (MLR-RNN) الهجينة

الطبيعة الخطية لنموذج MLR تجعل دراسة البيانات غير الخطية غير ممكنة وبنتائج غير دقيقة. في هذه الدراسة تم اقتراح الطريقة الهجينة MLR-RNN للتحليل والتكهن لبيانات خطية وغير خطية في آن واحد. حيث يتم الاعتماد على نموذج MLR كنموذج تقليدي احصائي في بناء هيكلية RNN وبالتالي الحصول على تكهنات ادق للبيانات غير الخطية قيد الدراسة.
المرحلة الاولى تتضمن بناء أفضل نموذج MLR للحصول على افضل هيكلية لادخالات RNN.
ان الجانب الايمن من النموذج MLR كما في المعادلة رقم (1) يمثل المتغيرات الخطية مع معلماتها وستستخدم كادخالات لشبكة RNN عدا متغير الخطأ. بالنتيجة فان ادخالات RNN ستكون هي نفسها ادخالات MLR من متغيرات مستقلة مضروبة بقيم معلماتها بطبيعتها الموجبة او السالبة. ان الهدف والذي من المرجو تطابقه او اقترابه من اخراج الشبكة العصبية للحصول على اقل خطأ للتكهن والذي سيمثل البيانات الاصلية. هذه الطريقة تدعى بالطريقة الهجينة MLR_RNN.
عدم خطية البيانات تضطر الدارسين والمهتمين في هذا المجال الى اللجوء الى دوال التحويل المذكورة آنفا واستخدام تركيبة تحمل كافة الاحتمالات الممكنة للطبقتين المخفية والاخراج.

3– النتائج والمناقشة.

ان اولى الخطوات الرئيسة في عمل البحوث هي الحصول على البيانات التي تتناسب مع الجانب العملي من الدراسة من جانب ودقتها من جانب اخر وهي الركيزة الأساسية للوصول إلى النتائج التي يمكن الاعتماد عليها في أي عمل بحثي، ولذلك كانت دراسة بيانات الأرصاد الجوية في ماليزيا التي غطت ثلاث سنوات من 2013–2015. وتم اختيار البيانات الأكثر تأثيرا على بيانات (PM_{10} Particulate Matter) الذي يشير الى جسيمات المادة في الهواء هذه الجسيمات هي اشياء مثل الغبار العضوي والبكتريا المحمولة جوا وغبار البناء. في هذه الدراسة سيتم عرض نتائج التكهنات للبيانات الكلية وسيتم استخدام مقاييس خطأ التكهن MAPE للطرق المستخدمة في الدراسة MLR وMLR-RNN ومناقشتها. تم تنفيذ الجانب التطبيقي من الدراسة باستخدام البرامج Minitab, MATLAB, Excel لتنفيذ الاوامر التي تخدم هذه الدراسة. حيث صممت هذه الدراسة للتنبؤ بقيم PM_{10} في الغلاف الجوي وبعض ملوثات الهواء التي تؤثر في بيانات PM_{10} وهي عدد من العناصر العضوية وتكون صلبة او سائلة وتختلف في حجمها مثل CO, (O3, SO2, NO, NOX) وبعض المؤثرات منها درجة حرارة الهواء والرطوبة وسرعة الرياح بمقياسين مختلفين 10 m و xxm) حيث تغطي البيانات اليومية لمدة ثلاث سنوات من 1 كانون

الثاني 2013 الى 31 تشرين الاول 2015 في احدى الولايات الماليزية. وسيتم حساب نماذج MLR باستخدام برنامج Minitab لإيجاد نماذج الانحدار وبتطبيق المعادلة (1) تم الحصول على النموذج التالي للبيانات الكلية.

$$PM_{10} = -B0 + B1CO - B2O3 + B3SO2 + B4\,Nox - B5\,NO + B6\,AT - B7\,H + B8WS10m - B9WSxxm \quad (7)$$

حيث ان WS10m, WSxxm يمثلان سرع الرياح بمقياسين مختلفين وأن AT تمثل درجة الحرارة وكذلك تمثل H الرطوبة النسبية. ويمثل PM_{10} المتغير y. أما مجموعة المتغيرات CO، O3، SO2، NOX، NO درجة حرارة الهواء و الرطوبة وسرعة الرياح بقياسين مختلفين فتمثل المتغيرات التفسيرية $x_1, x_2 \ldots, x_p$ على التوالي، و كذلك $\beta_0 + \beta_1 + \ldots + \beta_p$ معلمات الانحدار للمتغيرات التفسيرية، و سيتم استخدام برنامج Minitab لايجاد نماذج الانحدار الخطي المتعدد، والذي يتضمن عددا كبيرا من البيانات والمتغيرات التي يصعب حلها دون استخدام الحاسوب.

وتم استخدام جزء من البيانات الكلية لجعلها بيانات تدريب وايجاد نماذج الانحدار لهذه البيانات ليتم ايجاد افضل النماذج والجزء الباقي للاختبار للتأكد من قوة التكهن بالنماذج. في هذه الدراسة تم اخذ بيانات التدريب بدء من 1 كانون الثاني 2013 إلى 30 نيسان 2015 والجزء المتبقي من البيانات المستخدمة للاختبار بدء من 1 ايار 2015 إلى 31 تشرين الاول ثم وجد

ان افضل نموذج انحدار خطي متعدد للبيانات لفترة التدريب بعد حذ المتغيرات غير المعنوية كان كما في النموذج التالي.

$$PM_{10} = -198 + 110CO + 570\,NOX - 2062NO + 5.64AT + 4.05WS10m \quad (8)$$

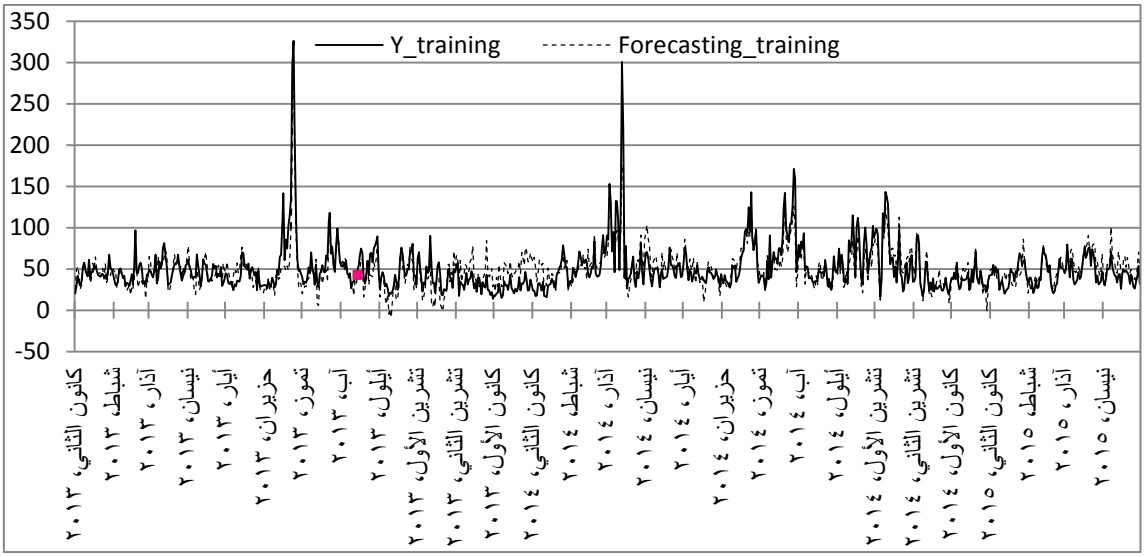
وبعد التكهن بالبيانات الخاصة بالمتغير المعتمد PM_{10} للبيانات الكلية الخاصة بالتدريب فمن الممكن اعتماد هيكلية نموذج الانحدار النهائي في المعادلة (8) كأفضل هيكلية ممكنة لإدخالات RNN.

الجدول التالي يعرض نتائج MAPE لطريقة MLR في الحالتين التدريب والاختبار.

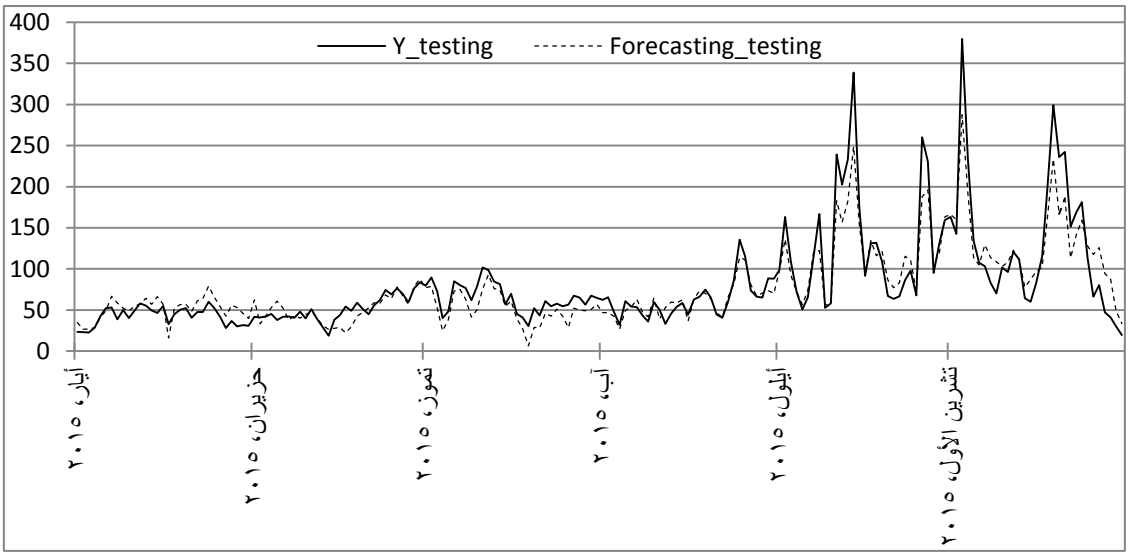
الجدول (1): قياس مفسرات الخطأ MAPE لطريقة MLR للتدريب والاختبار

MAPE	
Training	**Testing**
26.744	**20.755**

في الاشكال التالية سيتم توضيح الاتساق بين السلسلتين الاولى الاصلية PM_{10} و الثانية سلسلة التكهن لبيانات التدريب والاختبار للبيانات MLR ورسمها باستخدام برنامج Excel.



الشكل (4) : اتساق السلسلتين الاصلية والتكهن لبيانات التدريب بطريقة MLR.



الشكل (5) : اتساق السلسلتين الاصلية والتكهن لبيانات الاختبار بطريقة MLR.

الشكلين (4) و (5) يعرضان اتساقا جيدا وتجانس كبير بين السلسلة الاصلية PM_{10} و سلسلة التكهن باستخدام طريقة MLR لكلا بيانات التدريب ةالاختبار على التوالي.

في هذه الدراسة تم استخدام RNN لتحسين أداء MLR وبأقل الاخطاء وذلك باستخدام برنامج MATLAB وإيجاد مفسرات الخطأ التي سيتم على اساسها المقارنة بين النتائج.

وبعد تطبيق طريقة RNN للوصول الى MLR-RNN الهجينة. عدم خطية البيانات تضطر الدارسين في هذا المجال الى اللجوء الى دوال تحويل مثل دالة التحويل واللوغارتيمي (log) في الطبقة المخفية لتنقية عدم الخطية وتصفيتها فيما يتم تحديد دالة التحويل الخطي (line) لاستخدامها في طبقة الاخراجات وهذا ضروري بعد ما تم معالجة البيانات من مشكلة عدم الخطية. وبافتراض دالة التحويل للطبقة الخارجية هي الدالة الخطية وحسب المعادلة

$f(SUM) = SUM$ والتي بتطبيقها ستكون الطريقة الهجينة جاهزة للتكهن بالبيانات وحسب المعادلة التالية:

$$net_j(t) = \sum_{i=1}^{N} w_{i,j} Z_j(t) + b_j \qquad (9)$$

عندما تمثل i عدد الادخالات وتمثل j,l الطبقة المخفية و b هي التشويش الابيض وتمثل f اخراج الدالة

$$y_j(t) = f(net_j(t)) \qquad (10)$$

$$net_j(t-1) = \sum_{i=1}^{N} w_{i,j} Z_j(t) + \sum_{i=1}^{m} y_l(t-1) u_{jl} + b_j(t) \qquad (11)$$

عندما u_{jl} هو الوزن الجديد المعاود (الناتج من التنفيذ السابق) و m هي عدد الحالات او العقد

$$net_k(t) = \sum_{j=1}^{m} w_{k,j} y_k(t-1) + b_k \qquad (12)$$

عندما g هي ناتج الدالة الخارجية و $w_{k,j}$هي الاوزان من الطبقة المخفية الى طبقة الخارجية.

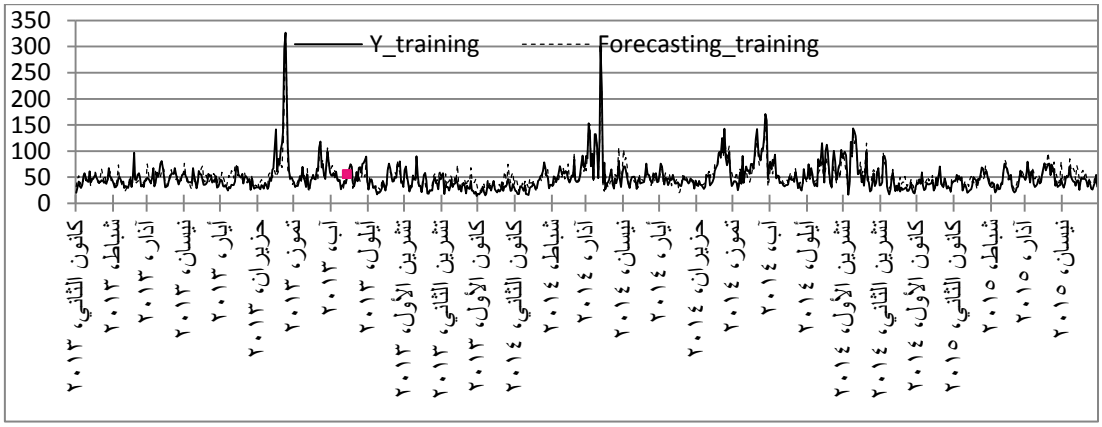
وتم حساب نتائج MAE والتي تم التوصل لها للبيانات كما ستعرض في الجدول التالي.

الجدول (2): قياس مفسرات الخطأ MAPE لطريقة MLR-RNN للتدريب والاختبار

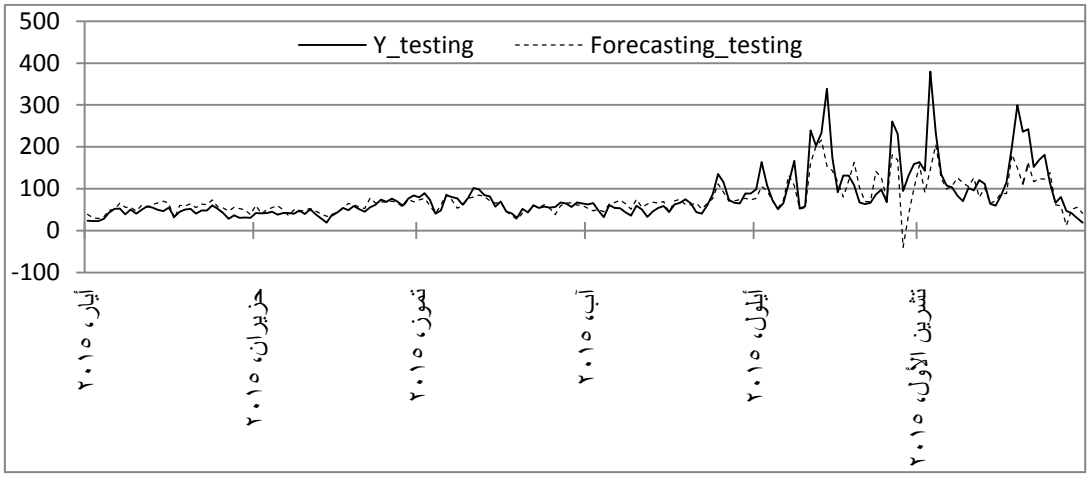
MAPE	
Training	**Training**
20.112	**19.508**

الجدولين (1) و (2) تم ملاحظة أن هناك تحسنا في النتائج عند تطبيق الاسلوب الهجين MLR-RNN بتكهنات افضل من سابقتها في MLR.

وباستخدام البيانات الناتجة عن MLR-GA سيتم رسم الاتساق للبيانات الاصلية مع بيانات التكهن الناتجة لبيانات التدريب والاختبار.



الشكل (6) : اتساق السلسلتين الاصلية والتكهن لبيانات التدريب بطريقة MLR-RNN.



الشكل (7) : اتساق السلسلتين الاصلية والتكهن لبيانات الاختبار بطريقة MLR_RNN.

الشكلين (6) و (7) يعرضان اتساقا جيدا وتجانس كبير بين السلسلة الاصلية PM_{10} و سلسلة التكهن باستخدام طريقة MLR-RNN لكلا بيانات التدريب والاختبار على التوالي وهي افضل مما كانت عليه في طريقة MLR لوحدها.

4–الاستنتاجات Conclusion

1. يمكن استخدام نموذج MLR لنمذجة بيانات متعددة المتغيرات باستخدام نموذج إحصائي خطي.

2. اثبت استخدام RNN من خلال النموذج الهجين MLR-RNN تحسنا ونتائج أدق عند ادخاله في نماذج التكهن الهجينة مقارنة مع استخدام الطرق التقليدية لوحدها وذلك انه يعالج مشكلة اللاخطية.

المصادر

Güler, N. and Güneri İşçi, Ö. (2016). The Regional Prediction Model of Pm10 Concentrations for Turkey. *Atmospheric Research*, 180, 64-77.

Hu, M. J.-C. (1964). *Application of the Adaline System to Weather Forecasting.* Master Thesis, Department of Electrical Engineering, Stanford University. Stanford, CA.

Nathans, L. L., Oswald, F. L. and Nimon, K. (2012). Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. *Practical assessment, research & evaluation*, 17(9).

Palit, A. K. and Popovic, D. (2006). *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications*: Springer.

Sheela, K. G. and Deepa, S. N. (2013). Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Mathematical Problems in Engineering*, 2013.