



التنبؤ لبيانات تلوث الهواء باستخدام الطريقة الهجينه MLR-RNN مع طريقة التراصf الزمني

خاتم وليد قادر  و اسامه بشير شكر 

قسم الاحصاء والمعلوماتية ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

الخلاصة

ان دراسة الجسيمات المعلقة ((PM10)) Particular matter والتken بها ضروري للتقليل والسيطرة على الأضرار البيئية وصحة الإنسان. هنالك العديد من مصادر التلوث او ما يسمى بالملوثات والتي ربما تؤثر على متغير PM10. كل هذه المتغيرات تصنف بياناتها كغير خطية. تم اخذ بيانات الدراسة من محطة مناخية في ماليزيا. تم استخدام نماذج الانحدار الخطى المتعدد (MLR) Multiple linear regression كطريقة إحصائية خطية للتنبؤ بمتغير PM10 من خلال تأثيره بمتغيرات الأرصاد الجوية المقابلة، لذلك فقد يعكس نتائج غير دقيقة عند استخدامه مع مجموعات البيانات غير الخطية. طريقة التراصf الزمني في أنماط مختلفة تم استخدامها لتحسين تلك النتائج وتحقيق التجانس ويتضمن مراصفة الموسام المتشابهة في السنوات المختلفة سوية لتكوين متغير جديد مختلف عن الاصلي. لتحسين نتائج التنبؤ تم اقتراح الشبكات العصبية المعاودة (RNN) Recurrent neural network لاستخدامه بعد التوفيق مع نموذج MLR. ضمن الطريقة الهجينه MLR-RNN. ان نتائج التنبؤ بشكل عام كانت الافضل باستخدام اسلوب التراصf الزمني. وكذلك عكست النتائج افضلية تنبؤات الطريقة الهجينه مقارنة مع نموذج MLR. وكاستنتاج في هذه الدراسة فمن الممكن استخدام الشبكات العصبية المعاودة وأسلوب التراصf الزمني كاساليب فعالة للحصول على افضل نتائج التنبؤ مع البيانات غير الخطية متعددة المتغيرات.

معلومات النشر

تاريخ المقالة:

تم استلامه في 7 اذار 2021

تم القبول في 17 نيسان 2021

متاح على الانترنت في 1 كانون الاول 2021

الكلمات الدالة:

الانحدار الخطى المتعدد (MLR) ،

الطريقه الزمنية (TS) ،

المواد الخاصة (PM10) ،

التنبؤ ،

تلوث الهواء ،

MLR-RNN الهجين

المراسلة:

خاتم وليد قادر

khetamalzubaidy86@gmail.com

DOI: [10.33899/IQJOSS.2021.169962](https://doi.org/10.33899/IQJOSS.2021.169962) , ©Authors, 2021, College of Computer Science and Mathematics, University of Mosul.

This is an open access article under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. المقدمة

في هذه الدراسة تم التطرق إلى دراسة التنبؤ بتلوث الهواء والتي تكمن اهميتها من خلال معرفة تأثيرها على الإنسان والحيوان والنبات وسائر الكائنات الحية على سطح الكرة الأرضية . تم استخدام بيانات متغير PM_{10} كمتغير معتمد لقياس تلوث الهواء والمتأثر بعده متغيراً تسمى تلوث الهواء والمتغيرات الجوية حيث ان هنالك علاقة وثيقة بين تركيز تلوث الهواء والمتغيرات التي تؤثر في حالة الطقس. تم استخدام نموذج الانحدار الخطى المتعدد MLR واستخدامه في التنبؤ بتلوث الهواء مع الأخذ في الاعتبار تأثيرات متغيرات الأرصاد الجوية المقابلة له. قام (2011) Vlachogianni *et al.* بدراسة لتطوير نماذج التنبؤ باستخدام الانحدار الخطى المتعدد ومقارنته للتنبؤات من نموذج MLR مع تلك الناتجة من استخدام الشبكة العصبية الاصطناعية ANN. ان نتائج تنبؤات الشبكات العصبية كانت أفضل من تلك المستحصلة من نموذج MLR. واستخدم (2011) Janssen *et al.* نماذج الانحدار الخطى المتعدد لمذكرة بيانات PM10 لمدينة باريس للمستويات اليومية. كما وقدر (2019) Ahmad *et al.* تركيز الجسيمات الدقيقة $PM_{2.5}$ باستخدام اسلوب هجين لانحدار الخطى والشبكة العصبية الاصطناعية وأظهرت النتائج أن كل من الانحدار الخطى وANN متقارن تماماً وقدران على تقدير تركيزات $PM_{2.5}$ مع دقة الشبكات العصبية الاصطناعية ANN. وفي مسار الدراسات السابقة ومن اجل تحسين دقة نتائج التنبؤ بالجسيمات المعلقة في الهواء PM_{10} والتي تعد بيانات بتغيرات

غير خطية وبوجود مؤثرات جوية اخرى، تم في هذه الدراسة تقديم عدة طرائق مقتربة للتنبؤ اضافة لنموذج الانحدار الخطى المتعدد (MLR) كطريقة تقليدية احصائية فقد تم اقتراح احد الطرق الذكائى وهي الشبكات العصبية المعاودة (RNN) Recurrent Neural Network لتحسين نتائج التنبؤ من خلال تهجينها مع الطريقة التقليدية. تم التطبيق على بيانات سلسلة معلمات تلوث الهواء المتمثلة بالجسيمات المعلقة PM_{10} بواقع (1034) مشاهدة حيث تتكون هذه السلسلة من المتغير المعتمد PM_{10} ومتغيرات مستقلة وعددها (9) متغيرات. تم اختزال المتغيرات التفسيرية الى (4) متغيرات لان هذه المتغيرات الاربعة اعطت افضل نموذج انحدار وهذه المتغيرات هي CO, SO_2, NO, O_3 حيث CO هو احادي اوكسيد الكاربون (Carbon Monoxide) وان SO_2 هو ثاني اوكسيد الكبريت (Sulphur Dioxide) و NO هو احادي اوكسيد النيتروجين (Nitric oxide) و O_3 هو الاوزون (Ozoon) اما المتغيرات الاخرى فاعطت معلماتها قيمها مقدرة غير معنوية اعتمادا على قيمة P-value مما جعل من حذفها الحل الامثل للوصول الى النموذج الافضل. تم استخدام برنامج Minitab (مینیتاب) وبرنامج Excel للحصول على نموذج الانحدار الخطى المتعدد وتنبؤاته لمرحلتي التدريب والاختبار. ان البيانات التي تمت دراستها تم اخذها بحالتين الاولى تم اخذها كاملا اي بفترتها الكلية وتبدأ من (1/1/2013) وتنتهي الى (31/10/2015) وفي الحالة الثانية تم فيها استخدام اسلوب التراصف الزمني حيث تم تقسيم هذه البيانات الى اربعة مواسم لكي تصبح كل مجموعة من المجموعات الاربعة اكثر تجانسا وتلاؤما مع الطرق المقتربة. ستم عملية المراصفة الزمنية على شكل فصول موسمية حيث تحدد اربع مجاميع متراصفة زمنياً بالاعتماد على الفصول الموسمية في كل سنة ومتراصفتها زمنيا مع جميع الفصول المشابهة في السلسلة الزمنية ومن الممكن صياغة منهجة اسلوب التراصف الزمني المستخدم وكما هو مدرج ادناه: في الموسم الاول ورمزه (S_1) حيث سيكون عدد المشاهدات (239) مشاهدة وعدد الشهور (8) شهور) والشهر التي سوف يتم اخذها هي: (كانون الثاني، شباط، كانون الاول) ولثلاث سنوات (2013, 2014, 2015) وعلى التوالي. اما في الموسم الثاني ورمزه (S_2) سيكون عدد المشاهدات (276) مشاهدة وعدد الشهور (9) شهر) والشهر التي سوف يتم اخذها هي: (اذار، نيسان، ايار) ولثلاث سنوات (2013, 2014, 2015) وعلى التوالي. وفيما يخص الموسم الثالث ورمزه (S_3) حيث سيكون عدد المشاهدات (276) مشاهدة وعدد الشهور (9) شهر) والشهر التي سوف يتم اخذها هي: (حزيران، تموز، اب) ولثلاث سنوات (2015, 2014, 2013) وعلى التوالي. وكذلك بالنسبة للموسم الرابع ورمزه (S_4) سيكون عدد المشاهدات (243) مشاهدة وعدد الشهور (8) شهر) والشهر التي سوف يتم اخذها هي: (ايلول، تشرين الاول، تشرين الثاني) ولثلاث سنوات (2015, 2014, 2013) وعلى التوالي. ان الطرق المقتربة لتحسين التنبؤ بتلوث الهواء من خلال المتغير PM_{10} تم مقارنتها مع الانحدار الخطى المتعدد MLR ، اي مقارنة MLR-RNN مع نموذج MLR وستتم المقارنة في حالة البيانات الكلية والبيانات المتراصفة زمنيا.

2. الطرق المستخدمة للتنبؤ

في هذا الجانب تم التطرق إلى طرق التنبؤ الشائعة الاستخدام بالبيانات متعددة المتغيرات ومن بينها نموذج الانحدار الخطى المتعدد MLR. ورغم ما يوفره نموذج الانحدار الخطى المتعدد من جودة في نمذجة البيانات الا انه قد يكون غير ملائم للتنبؤ بالبيانات غير الخطية على اعتبار ان نموذج MLR نموذج خطى مما يؤدي إلى ظهور بعض النتائج والتنبؤات بدقة قليلة أحياناً في حالة استخدام نموذج خطى مثل نموذج الانحدار الخطى المتعدد MLR . ولذلك يفضل استخدام طرق غير خطية مثل الشبكات العصبية المعاودة RNN مما يؤدي للحصول على نتائج اكثر دقة للتنبؤ والتحليل (Jahandideh *et al.*, 2009).

2.1 نموذج الانحدار الخطى المتعدد (MLR)

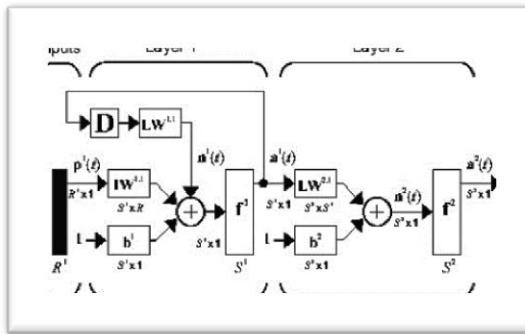
ان نموذج الانحدار الخطى المتعدد يعترف من الامثل للبيانات في التنبؤ خصوصا لبيانات السلسلة الزمنية عن طريق الاستخدام الامثل للبيانات في ايجاد علاقات سببية بين بيانات الدراسة ويعرف الانحدار الخطى المتعدد بشكل عام بأنه اسلوب رياضي لتوضيح العلاقة بين المتغير المعتمد Dependent Variable ومتغيرات أخرى تسمى المتغيرات التفسيرية Explanatory Variables وبهتم تحليلا الانحدار بوصف العلاقة بين المتغيرات على هيئة نموذج وقد يحتوي هذا النموذج على متغير تفسيري واحد فيسمى في هذه الحالة بنموذج الانحدار الخطى البسيط اما في حالة احتواء النموذج على عدة متغيرات تفسيرية عدة فإنه يسمى بنموذج الانحدار الخطى المتعدد (Honarasa *et al.*, 2015). كما ويعرف نموذج الانحدار الخطى المتعدد MLR بأنه عبارة عن انحدار للمتغير المعتمد y على العديد من المتغيرات التفسيرية $x_1, x_2, x_3, \dots, x_m$ لذا فهو يستخدم في التنبؤ. إذ يتم استخدام MLR لشرح العلاقة بين متغير معتمد ومتغيران تفسيريان او أكثر تتخذ المعادلة الخطية في الانحدار الخطى المتعدد الشكل التالي (Abrougui *et al.*, 2019)

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + ei \quad (1)$$

اذ أن y يرمز إلى المتغير المعتمد و α ترمز إلى القيمة الثابتة و x_1 يرمز إلى المتغير التفسيري الأول و x_2 يرمز إلى المتغير التفسيري الثاني و β_1 يرمز إلى المتغير التفسيري الاخير و β_2 يرمز إلى المعلمة الخاصة بالمتغير التفسيري الاول x_1 و β_2 إلى المعلمة الخاصة بالمتغير التفسيري الثاني x_2 و e^i يرمز إلى المعلمة الخاصة بالمتغير التفسيري الاخير x_m و x_m يرمز إلى الخطأ العشوائي.

2.2 الشبكات العصبية الاصطناعية (ANN) Artificial Neural Networks

تعد الشبكات العصبية الاصطناعية أحد أهم طرق الذكاء الاصطناعي والتي تستخدم في التنبؤ بالمتغير المعتمد للحصول على دقة اكبر، و تتمحور فكرتها حول محاكاة قدرة العقل البشري على التعرف على الأنماط و تمييز الأشياء باستخدام الحاسوب الآلي، والتي يتم فيها الاستفادة من الخبرات السابقة في سبيل الوصول إلى أفضل نتائج في المستقبل (Lin et al., 2020). تكون الشبكات العصبية الاصطناعية من مجموعة من الخوارزميات يتم من خلالها محاكاة الدماغ البشري المتتطور، و تصنف أدمغة إلكترونية قادرة على التعلم والتطور كما الدماغ البشري. والمميز في الشبكات العصبية الاصطناعية هو وجود طبقات عديدة تعمل على ما يسمى التعلم العميق، كل طبقة مختصة بعمل معين (Zhou et al., 2020). توجد عدة أنواع من الشبكات العصبية الشائعة الاستخدام واهم هذه الانواع والتي تؤدي غالباً إلى نتائج اكتر دقة هي الشبكات العصبية المعاودة (RNN) (Torkashvand et al., 2017). ان الشبكات العصبية المعاودة تستخدم مخرجات طبقة معينة وتعيها مرة أخرى للشبكة لكن كمدخلات ونتيجة لذلك يمكن أن يساعد ذلك في التنبؤ بالعديد من النتائج المحتملة خلال أي طبقة من طبقات الشبكة فتحتفظ كل طبقة بذاكرة من الخطوة السابقة فتتذكر النظام الترتيبات الخاطئة ويتعلم منها لتحسين تنبؤاته التالية ولذلك تسمى باسم الشبكات المعاودة لمعاودتها نفس الخطوات حتى الوصول إلى النتائج المطلوبة وبالتالي تستطيع شبكات RNN التعلم من كل خطوة للتنبؤ بالنتيجة في الخطوة التالية. تحتوي RNN على طبقة واحدة او اكتر وهذا بدوره يعالج غير خطية البيانات ويسهل نتائج التنبؤ وكذلك تحوي على Daley Layer وهذا يحسن كثيراً التعامل مع مشكلة عدم تجانس البيانات و عدم الخطية لانه يحتوي على ذاكرة اطول بقليل من خوارزمية الشبكة العصبية المغذية Feed-Forward والشكل التالي يمثل الشبكة وما تحتويه من ادخالات و اخراجات وطبقات Back Propagation .



الشكل (1) : يوضح الشبكة العصبية المعاودة RNN

في الشكل (1) فإن R هي الادخالات و $LW1$ هي وزن عشوائي للعصيوبن حيث يتم جمعها مع الجزء المتحيز $b1$ (التشويس الابيض) وناتجهما سيكونان الدالة $f1$. حيث ان اخراج الدالة $f1$ سيعود كادخال ثالث في الطبقة الاولى وقبلها سوف يمر على دالة التأخير (Delay) لتكون وزنا عشوائيا اخر وفي الطبقة الثانية فأن اخراج الدالة f يكون الوزن العشوائي للخلية العصبية LW مجموع مع b وبالتالي تخرج لنا مصفوفة احادية. تحتوي RNN في هذه الدراسة على طبقتين بالإضافة الى طبقة الادخال، الاولى تكون مخفية والثانية تكون طبقة الاصدار. حيث في طبقة الادخال سيكون هناك (R) من الادخالات وهذه الادخالات غالباً ماتكون توزن عشوائيا في كل طبقة مخفية وكذلك M من العصيوبنات. حيث يتم حساب العدد الامثل للعصيوبنات في الطبقة المخفية كما يلي:

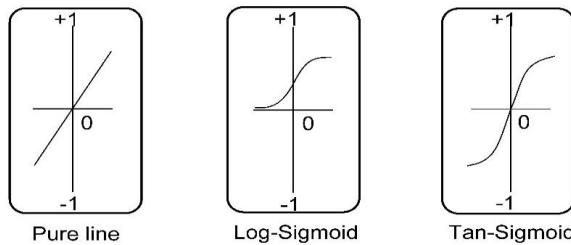
$$R^2+1 = \text{عدد العصيوبنات} \quad (2)$$

(Palit & Popovic, 2006; Sheela & Deepa, 2013)

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

$$net_j(t) = \sum_{i=1}^N w_{i,j} Z_j(t) + b_j \quad (3)$$

وان اكثر دوال التحويل استخداما في الطبقة المخفية وطبقة الارجح هي التحويل الزاوي (tan-sigmoid) والتحويل اللوغاريتمي (log-sigmoid) ودالة التحويل الخطية (linear). ان عملية اختيار الدالة للطبقتين امر مهم جدا يؤدي الى تحسين دقة النتائج بالاعتماد على طبيعة البيانات والدالة المختارة ومدى التجاوز بينهما، الشكل(2) يوضح الاختلافات بين دوال التحويل الممكن استعمالها للشبكة RNN.



الشكل(2) : انواع دوال التحويل في RNN

تستعمل دوال التحويل في الطبقة المخفية لتعكس نوعية العلاقة بين الادخالات والاخراجات في حين تستعمل دوال التحويل في طبقة الارجح لتعطي افضل وادق النتائج. والصيغ الرياضية للدوال الخطية واللوغاريتمية والزاوية هي كما يلي وعلى التوالي:

$$f(SUM) = SUM \quad (4)$$

$$f(SUM) = \frac{1}{1 + e^{-SUM}} \quad (5)$$

$$f(SUM) = \frac{2}{1 + e^{-2SUM}} - 1 \quad (6)$$

(Dawson & Wilby, 2001; Shrestha *et al.*, 2005; Yonaba *et al.*, 2010)

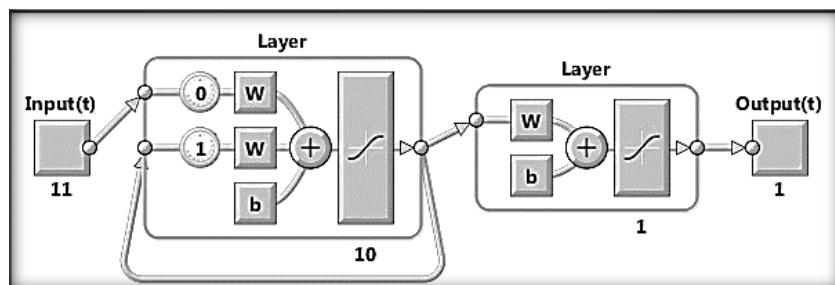
ان الاوزان العشوائية $w_{i,j}$ للادخالات يمكن كتابتها كمصفوفة وعلى النحو التالي

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,R} \\ w_{2,1} & w_{2,2} & \dots & w_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M,1} & w_{M,2} & \dots & w_{M,R} \end{bmatrix} \quad (7)$$

في حين تصاغ متغيرات الادخال كما يلي :

$$Z = [Z_{t1} \ Z_{t2} \ \dots \ Z_{tR}]' \quad (8)$$

حيث ان الهيكل العام للشبكة RNN يحتوي على طبقة واحدة مخفية واخرى للارجح، حيث الطبقة المخفية تحتوي على ثلاث مدخلات الوزن العشوائي والتشويس الابيض بالإضافة الى الوزن الناتج من الخطوة السابقة اما الطبقة الخارجية فمدخلاتها ستكون الوزن الناتج من الطبقة المخفية مع التشويش الابيض ويمكن تمثيله كما في الشكل (3) :



الشكل (3) : الهيكل العام للشبكة العصبية المعاودة

2.3 الطريقة الهجينه (MLR-RNN)

تتضمن هذه الطريقة تهجين الطريقتين التقليدية (MLR) والذكائية (RNN) من خلال الاستفادة من هيكلية المتغيرات التفسيرية في MLR لبناء الشبكة RNN وكما هو مدرج في الخطوات التالية:

- أ. يتم ضرب كل متغير تفسيري في قيمة المعلمة المناظرة له واعتماد المتغيرات الناتجة كمدخلات للشبكة العصبية المعاودة (RNN) وبناء طبقة الادخال .
- ب. تحديد عدد العصبونات المستخدمة في الطبقة المخفية.
- ج. بعد اعتماد هيكلية نموذج MLR لطبقة الادخال للشبكة العصبية حيث تم عمليي التدريب والاختبار للحصول على افضل التنبؤات وتسمى مخرجات شبكة RNN في هذه الحالة هذه بتنبؤات طريقة الهجينه للشبكات العصبية MLR-RNN، حيث كلما تكررت التدريبات وكانت اكثرا كلما كانت النتيجة ادق.
- د. يتم تدريب الشبكة على جميع دوال التحويل الخطية واللوغاريتمية والزاوية في كل الطبقتين المخفية والاخراج وبأخذ جميع الاحتمالات الممكنة .
- هـ.

3. الترافق الزمني (TS)

ان اسلوب الترافق الزمني هو وسيلة تحليلية تقوم بمراقبة البيانات زمنيا تبعا للتأثيرات الموسمية التي تظهر بشكل واضح كتأثيرات على سلوك السلسلة الزمنية وسلوك النتائج التنبؤية ويضمن دقة تقديرات معلمات الانحدار الخطي المتعدد وتجنب التحيز بسبب اتجاه التأثيرات الزمنية في السلسلة الزمنية، ويمكن تطبيق الترافق الزمني على السلسلات الزمنية المختلفة في حالة كانت تظم اتجاهات زمنية موسمية متكررة بنفس السياق والتأثير ويعمل على الوصول الى بيانات اكثرا تجانسا من البيانات الكلية وبالتالي الحصول على نتائج ادق (Malig *et al.*, 2015; Tobias *et al.*, 2014)

- أ. ويمكن ايجاز الخطوات التي يتم بها الترافق الزمني بالنقاط الآتية: رسم بيانات السلسلة الزمنية المحددة للبيانات.
- ب. تحديد الفترات الموسمية وفي نمط الموسم.
- ج. سحب البيانات في هذه الفترات من السلسلة ومراقبتها.

4. مقياس خطأ التنبؤ Forecasting Error Measurement

سيتم استخدام واقتراح العديد من الطرائق والاساليب، وللمقارنة بينها سيتم استخدام متوسط النسبة المئوية المطلقة للخطأ (MAPE) Mean Absolute Percentage Error . ويعرف الخطأ بأنه هو تقدير لفرق بين القيمة الحقيقية والقيمة المقدرة حيث كلما كان الخطأ قليل فستكون الدقة اكبر. ويحسب مقياس الخطأ MAPE على النحو التالي:

$$MAPE = \frac{1}{n} \left[\sum \left| \frac{e_i}{y_i} \right| \right] \times 100 \quad (9)$$

حيث e_i : تمثل خطأ التنبؤ، n : هي عدد المشاهدات، $i = 1, 2, 3, \dots, m$

y_i : هو السلسلة الحقيقة او الاصلية المستعملة كمتغير هدف. ويحسب خطأ التنبؤ كما يلي:

$$e_i = y_i - \hat{y}_i \quad (10)$$

5. النتائج والمناقشات

في هذه الدراسة تم استخدام طريقة احصائية تقليدية شائعة الاستخدام متمثلة بنموذج MLR للتنبؤ بالجسيمات المعلقة بالهواء المتمثل بمقاييس (PM₁₀). وكذلك تم اقتراح طريقة اخرى لتحسين التنبؤ بتلوث الهواء لمتغير PM₁₀ بعد تهجينها مع الطريقة التقليدية. وقد تمت مقارنتها مع الانحدار الخطي المتعدد MLR في حالة البيانات الكلية وفي حالة البيانات المتراصفة زمنيا ورغم ما يوفه الانحدار الخطي المتعدد من امكانية نمذجة البيانات لاستخدامه في التنبؤ بالمتغير المعتمد إلا أن بيانات التلوث الجوي والأرصاد الجوية تأخذ نمطاً غير خطياً مما يؤدي إلى ظهور بعض النتائج والتنبؤات بدقة قليلة أحياناً وقد حسنت الطريقة الهجينه المقترنة من نتائج التنبؤ من خلال ما تضمنته من حلول وامكانات للتعامل مع البيانات غير الخطية. (Jahandideh *et al.*, 2009). تم تقسيم البيانات الى

قسمين قسم تدريب وقسم اخر للاختبار وكما مدرج ادناه:

1. بيانات التدريب : ستدأ من الفترة (1/1/2013) الى (31/5/2015)
2. بيانات الاختبار: ستدأ من الفترة (1/6/2015) الى (31/10/2015) والتي تقارب نسبة 15% من البيانات وهو ضمن المعدل الذي تناولته معظم البحوث في مجال الدراسة.

وبهذا سيكون هناك (881) مشاهدة للتدريب و(153) مشاهدة للاختبار وسيتم استخدام طريقة الانحدار MLR حيث تم حذف المتغيرات غير المعنوية اعتماداً على قيمة P-value كما سنشاهده في النتائج أدناه وقد تم اختيار افضل معادلة نموذج انحدار خطى متعدد اعتماداً على افضل تفسير للعلاقة والتأثير ومعنوية المعلومات حيث ان افضل نموذج انحدار خطى متعدد في حالة البيانات الكلية كما مدرج أدناه:

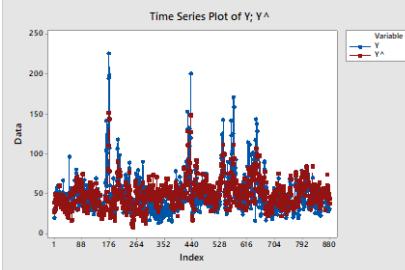
$$y = 79.01x_1 + 3501x_2 - 2029x_3 + 197.6x_4 \quad (11)$$

حيث يمكن تعريف متغيرات ومعلمات نموذج الانحدار الخطى المتعدد في (11) كما هو أدناه:
 y : هو متغير الاستجابة PM_{10} و x_1 : يمثل متغير احادي اكسيد الكاربون CO و x_2 : يمثل متغير احادي اكسيد الكبريت SO_2 و x_3 : يمثل متغير احادي النيتروجين NO و x_4 : يمثل متغير الاوزون O_3 . نلاحظ من خلال المعادلة (11) التي تمثل نموذج الانحدار الخطى المتعدد ان معلمات المتغيرات التفسيرية $\beta_1, \beta_2, \beta_3, \beta_4$ تساوى $(79.01), (179.60), (3501), (-2029)$ على التوالي والتي تمثل قيمتها قوة تأثير المتغير التفسيري على المتغير المعتمد بينما توضح الاشارة اتجاه ذلك التأثير.

ان جميع المعلمات المقدرة معنوية لأن قيم p -value هي اقل من مستوى المعنوية (0.05) مما يدل على معنوية المعلمات وكفاءة النموذج ولذلك فيعتبر النموذج الامثل. ان نتائج تنبؤات مرحلتي التدريب والاختبار من حيث دققها مقاسة من خلال قيمة MAPE كما في الجدول(2).

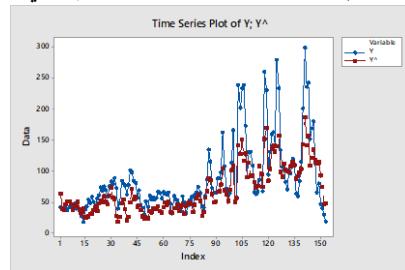
الجدول(2): قيم(MAPE) في حالة البيانات الكلية باستخدام نموذج MLR

| بيانات التدريب | بيانات الاختبار |
|----------------|-----------------|
| 27.223691 | 28.011285 |



الشكل (4): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التنبؤ) للمتغير

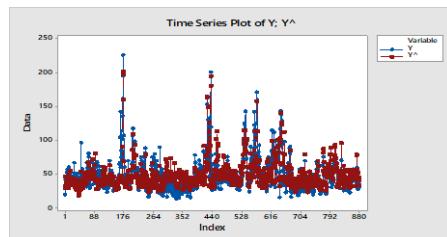
المعتمد(PM_{10}) خلال الفترة الكلية لبيانات التدريب في



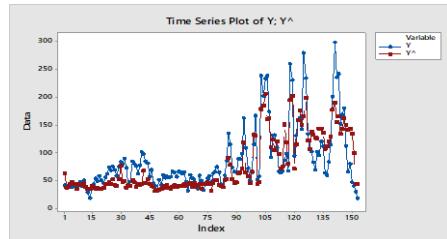
الشكل (5): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التنبؤ) للمتغير المعتمد(PM_{10}) خلال الفترة الكلية لبيانات الاختبار باستخدام MLR.

الجدول(3): قيم(MAPE) في حالة البيانات الكلية باستخدام MLR-RNN

| بيانات التدريب | بيانات الاختبار |
|----------------|-----------------|
| 22.0665 | 29.6825 |



الشكل(6): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التنبؤ) (PM_{10}) للفترة الكلية لبيانات التدريب باستخدام MLR-RNN .



الشكل(7): التطابق بين البيانات الأصلية مع البيانات التقديرية للمتغير المعتمد(PM_{10}) للفترة الكلية لبيانات الاختبار باستخدام MLR-RNN . حيث تم بناء افضل النماذج من خلال استخدام بيانات التدريب ثم التنبؤ بها اما بيانات الاختبار فتم التنبؤ بها باستخدام نفس النموذج الذي تم انشاؤه من بيانا التدريب، ولهذا السبب فأن دقة التنبؤ لبيانات التدريب علميا(tnbio داخل العينة In of sample forecasting) سيكون ادق من التنبؤ بالمشاهدات المستقبلية للسلسلة في حين لايمكن مقارنة النتائج للتنبؤ لبيانات التدريب والاختبار لاختلاف حجم العينة بينهما، ولنفس السبب فأن افضلية نتائج التنبؤ لبيانات التدريب باستخدام MLR-RNN مقارنة مع MLR وهي التي يمكن الحكم بها على الطريق الافضل وليس من خلال نتائج بيانات الاختبار. يلاحظ من الجدولين (2) و(3) والأشكال(4-7) ان هنالك تقارب وانسجام بين القيم الحقيقة والقيم المقدرة التنبؤية باستخدام MLR-RNN وبصورة اكثرا انسجاما مما كانت عليه في نموذج MLR وهذا يدل على ان الطريقة الهجينة RNN-MLR افضل من الطريقة التقليدية وهي نموذج MLR . بعد الحصول على التنبؤات لبيانات التدريب والاختبار لبيانات في الفترة الكلية فسيتم تقسيم البيانات الى اربعة مواسم من خلال استخدام الترافق الزمني وسيتم اتخاذ الاجراءات نفسها التي تم انجازها مع البيانات الكلية لكل موسم على حدة باستخدام الطرق نفسها MLR وMLR-RNN وكذلك سيم تم تقسيم البيانات الى مجموعتين بيانات تدريب وبيانات اختبار لكل موسم. بعد تقسيم البيانات الى بيانات تدريب وبيانات اختبار سيتم استخدام بيانات التدريب لايجاد افضل نموذج انحدار خطى متعدد لاربع مواسم S_1 S_2 و S_3 S_4 يعبر عنها وكما هو في المعادلات ادناه.

$$y = 16.46x_1 + 4357x_2 + 210x_3 + 954x_4 \quad (12)$$

$$y = 81.56x_1 + 1887x_2 - 2194x_3 + 308.3x_4 \quad (13)$$

$$y = 122.76x_1 - 3117x_2 - 2179x_3 + 358x_4 \quad (14)$$

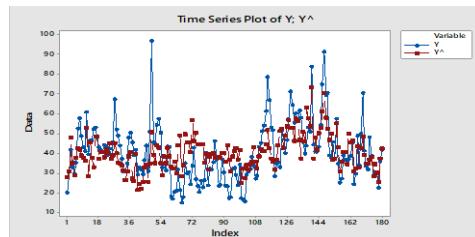
$$y = 82.02x_1 + 6149x_2 - 2387x_3 - 120x_4 \quad (15)$$

ولغرض اجراء المقارنة مع البيانات الكلية فسيتم القبول بالنموذج(12-15) اعلاه على الرغم من عدم معنوية بعض معلماته لاثبات هيكلاية موحدة للمتغيرات كوجه التشابه ليصح اجراء المقارنات على التوالي حتى في حال كانت بعض المعلمات غير معنوية وذلك لانه تم استخدام نموذج الانحدار الخطى المتعدد بنفس المتغيرات لبيانات الكلية . والجدول (4) سيوضح قيم مقياس الخطأ.

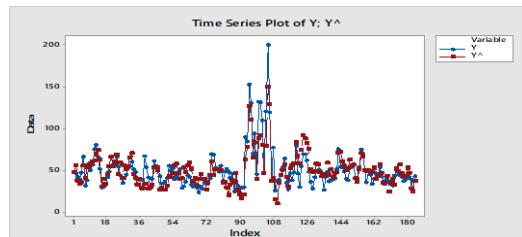
الجدول(4): قيم (MAPE) في حالة استخدام نموذج MLR .

| | بيانات التدريب | بيانات الاختبار |
|-------|----------------|-----------------|
| S_1 | 25.317 | 25.027 |
| S_2 | 20.739 | 22.057 |
| S_3 | 24.823 | 19.028 |
| S_4 | 31.627 | 29.683 |

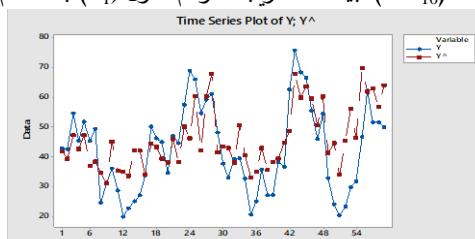
الأشكال (8-15) ادناه توضح الانسجام بين البيانات الأصلية مع البيانات التقديرية (سلسلة التنبؤ) للمتغير المعتمد(PM_{10}) لبيانات التدريب والاختبار لبيانات المترافقه زمنياً باستخدام نموذج (MLR) .



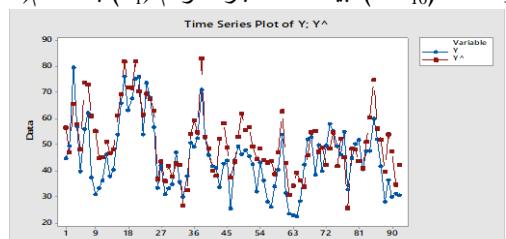
الشكل (8): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات التدريب للموسم الأول (S_1) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (1) باستخدام (MLR)



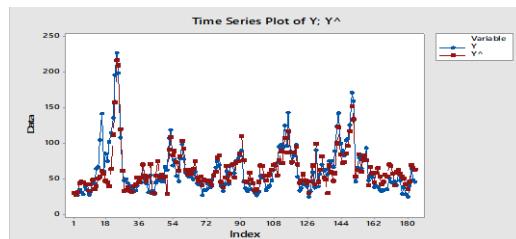
الشكل (9): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات التدريب للموسم الأول (S_1) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (1) باستخدام (MLR)



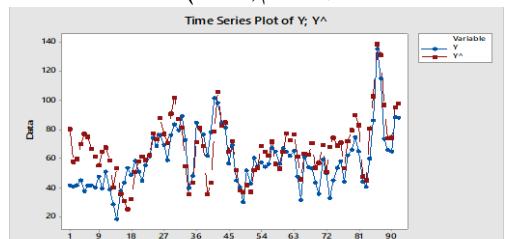
الشكل (10): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات التدريب للموسم (2) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (2) باستخدام (MLR)



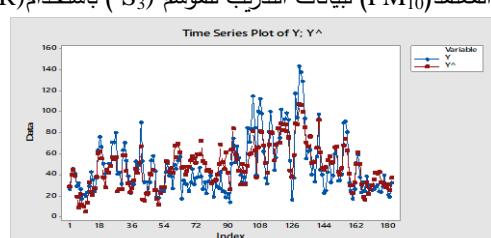
الشكل (11): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات الاختبار للموسم (2) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (2) باستخدام (MLR)



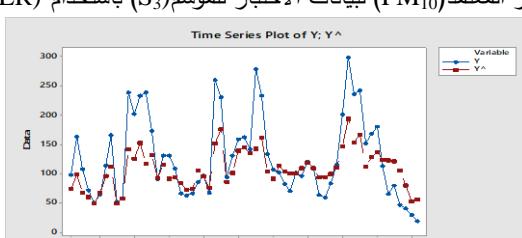
الشكل (12): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات التدريب للموسم (3) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (3) باستخدام (MLR)



الشكل (13): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات الاختبار للموسم (3) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (3) باستخدام (MLR)



الشكل (14): التطابق بين البيانات الأصلية مع البيانات التقديرية للمتغير المعتمد (PM_{10}) لبيانات التدريب للموسم (4) باستخدام (MLR) للمتغير المعتمد (MLR) لبيانات الاختبار للموسم (4) باستخدام (MLR)



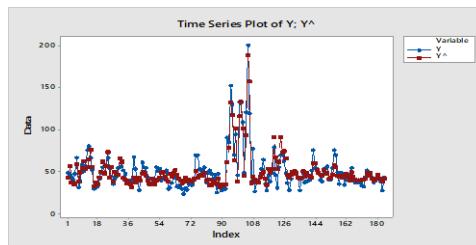
الشكل (15): التطابق بين البيانات الأصلية مع البيانات التقديرية (سلسلة التباين) للمتغير المعتمد (PM_{10}) لبيانات الاختبار للموسم (4) في (MLR)

ومن الاشكال (8-15) اعلاه فمن الممكن استنتاج انه باستخدام عملية التراصيف الزمني وتقسيم البيانات الى مواسم فأن الفرق بين الاحطاء الحقيقة والاحطاء التقديرية تكون اقل مما كانت عليه في حالة البيانات الكاملة باستخدام (MLR). بعد ان تم تطبيق (MLR) على البيانات الموسمية سيتم تهجين للشبكات العصبية المعاودة (RNN) باستخدام نموذج (MLR) وتنبئ هذه الطريقة (MLR-RNN) كل موسم من هذه المواسم على حدة والجدول (5) يوضح قيم مقياس الخطأ (MAPE) لبيانات التراصيف الزمني للتدريب والاختبار باستخدام طريقة (MLR-RNN) الهجينة .

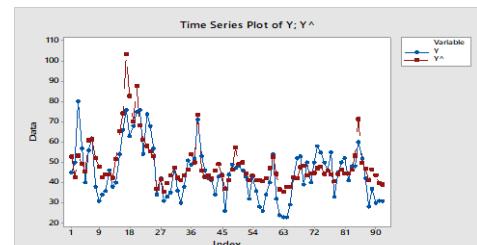
الجدول(5): قيم (MAPE) في حالة استخدام طريقة MLR-RNN الهجينة .

| | بيانات التدريب | بيانات الاختبار |
|-------|----------------|-----------------|
| S_1 | 21.486 | 20.641 |
| S_2 | 15.942 | 18.203 |
| S_3 | 16.433 | 20.239 |
| S_4 | 23.950 | 37.750 |

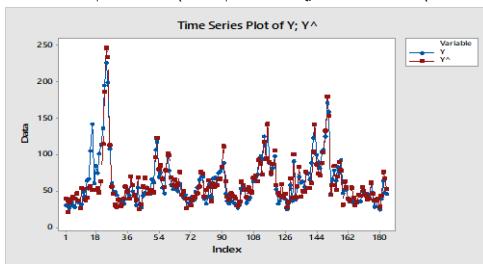
و عند مقارنة نتائج الخطأ (MAPE) للطريقة الهجينة (MLR-RNN) مع نتائج (MAPE) لنموذج الانحدار (MLR) في الجدولين (4 و 5) يتبيّن ان معيّن نتائج الطريقة الهجينة خصوصاً لبيانات التدريب تفوقت على نتائج نموذج MLR و اعطت نتائج اكثـر دقة .
 الأشكال (16-23) ادناه توضّح مدى التطابق والانسجام بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) للمتغير المعتمد (PM_{10}) لبيانات التدريب والاختبار لجميع البيانات المتراصفة زمنياً للموسم الاربعة باستخدام الطريقة الهجينة (MLR-RNN) .



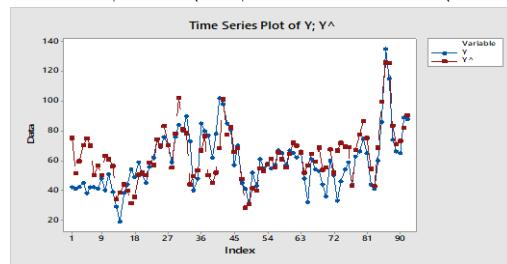
الشكل(18): التطابق بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) (PM_{10}) لبيانات التدريب للموسم (S_1) باستخدام (MLR-RNN).



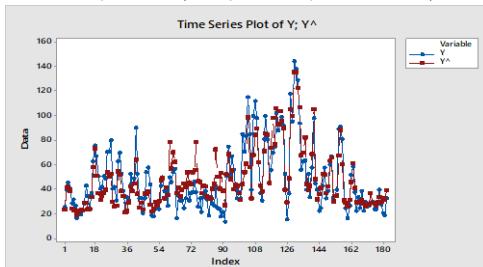
الشكل(19): التطابق بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) (PM_{10}) لبيانات الاختبار للموسم (S_2) باستخدام (MLR-RNN).



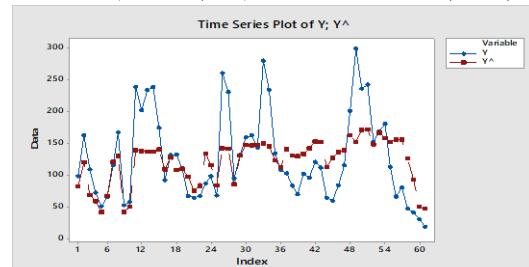
الشكل(20): التطابق بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) (PM_{10}) لبيانات التدريب للموسم (S_3) باستخدام (MLR-RNN).



الشكل(21): التطابق بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) (PM_{10}) لبيانات الاختبار للموسم (S_3) باستخدام (MLR-RNN).



الشكل(22): التطابق بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) (PM_{10}) لبيانات التدريب للموسم (S_4) باستخدام (MLR-RNN).



الشكل(23): التطابق بين البيانات الأصلية مع البيانات التقدّيرية (سلسلة التنبؤ) (PM_{10}) لبيانات الاختبار للموسم (S_4) باستخدام (MLR-RNN).

ومن الاشكال (16-23) اعلاه وكذلك من خلال المقارنة بين الطريقتين المقترجتين فأن التطابق بين السلسليتين باستخدام MLR-RNN افضل من النموذج MLR للبيانات المتراصفة زمنياً وكذلك افضلية نتائج التنبؤ للبيانات المتراصفة زمنياً مع حالة البيانات الكلية حيث تم بناء افضل النماذج من خلال استخدام بيانات التدريب ثم التنبؤ بها اما بيانات الاختبار فتم التنبؤ بها باستخدام نفس النموذج الذي تم انشاؤه من بيانا التدريب، ولهذا السبب فأن دقة التنبؤ لبيانات التدريب سيكون ادق من التنبؤ بالمشاهدات المستقبلية للسلسلة في حين لا يمكن مقارنة النتائج للتنبؤ لبيانات التدريب والاختبار لاختلاف حجم العينة بينهما، ولنفس

السبب فإن أفضلية نتائج التنبؤ لبيانات التدريب باستخدام MLR-RNN مقارنة مع MLR وهي التي يمكن الحكم بها على الطريق الأفضل وليس من خلال نتائج بيانات الاختبار.

6. الاستنتاجات

من خلال ما تم عرضه من نتائج ومناقشات للطرق المقترنة وتطبيقاتها على بيانات الدراسة فمن الممكن استنتاج أفضلية لاستخدام منهجية الطريقة المهجنة المقترنة MLR-RNN لتحسين نتائج التنبؤ لتلوث الهواء وخصوصا في حالة استخدام اسلوب الترافق الزمني لبيانات الدراسة والذي يحسن نتائج التنبؤ لتلوث الهواء.

7. Reference

1. Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R., & Chehaibi, S. (2019). Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). *Soil and Tillage Research*, 190, 202-208.
2. Ahmad, M., Alam, K., Tariq, S., Anwar, S., Nasir, J., & Mansha, M. (2019). Estimating fine particulate concentration using a combined approach of linear regression and artificial neural network. *Atmospheric Environment*. 117050 :219 .
3. Dawson, C., & Wilby, R. (2001). Hydrological modelling using artificial neural networks. *Progress in physical Geography*, 25(1), 80-108 .
4. Honarasa, F., Yousefinejad, S., Nasr, S., & Nekoeinia, M. (2015). Structure–electrochemistry relationship in non-aqueous solutions: predicting the reduction potential of anthraquinones derivatives in some organic solvents. *Journal of Molecular Liquids*, 212, 52-57 .
5. Jahandideh, S., Jahandideh, S., Asadabadi, E. B., Askarian, M., Movahedi ,M. M., Hosseini, S., & Jahandideh, M. (2009). The use of artificial neural networks and multiple linear regression to predict rate of medical waste generation. *Waste management*, 29(11), 2874-2879 .
6. Janssen, N. A., Hoek, G., Simic-Lawson, M., Fischer, P., Van Bree, L., Ten Brink, H., . . . Brunekreef, B. (2011). Black carbon as an additional indicator of the adverse health effects of airborne particles compared with PM10 and PM2. 5. *Environmental health perspectives*, 119(12), 1691-1699 .
7. Lin ,L., Dekkers, I. A., Tao, Q., & Lamb, H. J. (2020). Novel artificial neural network and linear regression based equation for estimating visceral adipose tissue volume. *Clinical Nutrition* .
8. Malig, B. J., Pearson, D. L., Chang, Y. B., Broadwin, R., Basu, R., Green, R. S., & Ostro, B. (2015). A time-stratified case-crossover study of ambient ozone exposure and emergency department visits for specific respiratory diagnoses in California (2005–2008). *Environmental health perspectives*, 124(6), 745-753 .
9. Palit, A. K., & Popovic, D. (2006). *Computational intelligence in time series forecasting: theory and engineering applications*: Springer Science & Business Media.
10. Sheela, K. G., & Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*, 2013 .
11. Shrestha, R. R., Theobald, S., & Nestmann, F. (2005). Simulation of flood flow in a river system using artificial neural networks. *Hydrology and Earth System Sciences Discussions*, 9(4), 313-321 .
12. Tobias, A., Armstrong, B., & Gasparini, A. (2014). *Analysis of time-stratified case-crossover studies in environmental epidemiology using Stata*. Paper presented at the United Kingdom Stata Users' Group Meetings 2014.
13. Torkashvand, A. M., Ahmadi, A., & Nikravesh, N. L. (2017). Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). *Journal of integrative agriculture*, 16(7), 1634-1 .644
14. Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., & Kukkonen, J. (2011). Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. *Science of the total environment*, 409(8), 1559-1571 .
15. Yonaba, H., Anctil, F., & Fortin, V. (2010). Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. *Journal of Hydrologic Engineering*, 15(4), 275-283 .
16. Zhou, F., Liu, B., & Duan, K. (2020). Coupling wavelet transform and artificial neural network for forecasting estuarine salinity. *Journal of Hydrology*, 125127 .

Air Pollution Forecasting using Hybrid MLR-RNN Method with Time-Stratified Method

Khitam Walid Qader

Osama Bashir Shukr

College of Computer Science and Mathematics, University of Mosul

Abstract: Studying and forecasting Particular matter (PM_{10}) is necessary to control and reduce the damage of environment and human health. There are many pollutants as sources of air pollution may effect on PM_{10} variable. This type of dataset can be classified as an non-linear. Studied datasets have been taken from climate station in Malaysia. Multiple Linear Regression (MLR) is used as a linear statistical method for PM_{10} forecasting through its influencing by corresponding climate variables, therefore it may reflect inaccurate results when used with nonlinear datasets. Time stratified (TS) method in different styles is implemental for satisfying more homogeneity of datasets. It includes ordering similar seasons in different years together to formulate anew variable smoother than their original. To improve the results of forecasting, Recurrent Neural Network (RNN) has been suggested to be used after combining with MLR in hybrid MLR-RNN method in this study. In general, the results of forecasting were the best with using time stratified approach. In addition, the results of hybrid method were outperformed comparing to MLR model. As conclusion in this study, RNN and TS can be used as active approaches to obtain better forecasting results with nonlinear datasets in which PM_{10} is to dependent variable.

Keywords: Multiple linear regression (MLR), Time Stratified (TS), Particular Matter (PM_{10}), Forecasting, Air Pollution, hybrid MLR-RNN.