Iraqi Journal of Statistical Sciences, Vol. 19, No. 2, 2022, pp(42 -55)

المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

استخدام نماذج ARIMA والغابة العشوائية للتنبؤ ببيانات الانواء الجوية

عدي زكي جرجيس الجبوري و أسامة بشير الحنون

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

معلومات النشر

تاريخ المقالة:
تم استلامه في 25 ايار 2022
تم القبول في 23 تموز 2022
متاح على الإنترنت في 1 كانون الاول 2022

الكلمات الدالة:
نموذج الانحدار الهرمي بواسون مع اعتراض عشوائي ، وطريقة الاحتمال الأقصى الكامل ، ومعامل الارتباط داخل الفصل ، والتأثيرات الثابتة والعشوائية

المراسلة:
عدي زكي جرجيس الجبوري
alastadhdyeljbwry@gmail.com

الخلاصة

ان التغيرات المناخ دوراً مؤثراً قد يؤدي الى مشكلات كثيرة على صحة الإنسان وبقية الكائنات الحية لذا فانه من الضروري دراستها والتنبؤ بها للحد او للتقليل من اضرارها من خلال التخطيط لها والسيطرة عليها. ان المشكلة الرئيسية تكمن في عدم خطية هذا النوع من البيانات و فوضويتها. ومن اشهر اساليب السلاسل الزمنية استخداما هي نماذج الانحدار الذاتي والمتوسطات المتحركة المندمجة Integrated Autoregressive and Moving Average model (ARIMA) كنماذج سلاسل زمنية تقليدية احادية المتغير. ان مثل هذه النماذج لا يمكنها التعامل بصورة سليمة مع البيانات غير الخطية فتُظهر نتائج تنبؤ قليلة الدقة. في هذه البحث تم استخدام بيانات الانواء الجوية متمثلة بدرجات الحرارة الصغرى وكميات التبخر لاحد محطات الانواء الجوية الزراعية في محافظة نينوى. تهدف هذه البحث الى تحقيق التجانس في البيانات خلال المواسم المختلفة وايجاد أنموذج يتعامل مع البيانات غير الخطية ويعطي اقل خطأ للتنبؤ مقارنة بالنموذج التقليدي ARIMA. لذلك فقد تم استخدام أنموذج اكثر تلاؤماً مع بيانات الانواء الجوية ليعطي تنبؤات غاية في الدقة يدعى نموذج الغابة العشوائية Random Forest (RF). ان من اهم اسباب تحسين نتائج التنبؤ هو اعتماد نموذج RF في اتخاذ القرار على العديد من أشجار الانحدار غير المترابطة والتي يؤدي كل منها الى قرار مستقل وأن القرار النهائي سيكون بالغالبية المطلقة لمجاميع أشجار الانحدار. تم الحصول على نتائج تنبؤ اكثر دقة باستخدام نموذج RF مقارنة بنتائج تنبؤات ARIMA في مرحلتي التدريب والاختبار. من ذلك فانه تم استنتاج افضلية مطلقة لنموذج RF اذا ما قورن مع نموذج ARIMA التقليدي عند التنبؤ بالبيانات المناخية.

DOI: https://doi.org/10.33899/iqjoss.2022.176203 ,

1. المقدمة Introduction

في هذه البحث تم التطرق الى دراسة التنبؤ ببعض متغيرات الانواء الجوية اذ تكمن أهمية هكذا تنبؤات من خلال معرفة مدى تأثيرها على الانسان والحيوان والنبات وسائر الكائنات الحية والتخطيط لمستقبل خالي من مشاكل التأثيرات السلبية لمتغيرات الانواء الجوية المختلفة وغني بتأثيراتها الإيجابية.

تم استخدام أنموذج ARIMA كأسلوب تقليدي شائع الاستخدام وبعد عدة محاولات تم الحصول على افضل نموذج ARIMA يلائم بيانات الدراسة. في هذه البحث تم استخدام بيانات الانواء الجوية متمثلة بدرجات الحرارة الصغرى وكميات التبخر لاحد محطات الانواء الجوية الزراعية في محافظة نينوى للفترة من (2018/5/15) ولغاية (2020/7/19). ان العديد من الباحثين في دراسات سابقة استخدموا بيانات الانواء الجوية على اختلافها للتنبؤ واستنتجوا عدم خطية بيانات الانواء الجوية على الاطلاق ولذلك قد يكون أنموذج ARIMA غير دقيق في نتائج التنبؤ لوجود تلك المشكلة في البيانات ولتلك الأسباب يقترح غالبا استخدام أساليب أخرى غير خطية تتعامل مع هكذا نوع من البيانات بشكل أفضل وبالتالي تعطي نتائج أفضل في التنبؤ مقارنة بنماذج ARIMA.

تعتبر نماذج الغابة العشوائية Random Forest طريقة دقيقة وقوية للغاية في التنبؤ بسبب اعتمادها في اتخاذ القرار على العديد من أشجار القرار حيث تكون أشجار القرار هذه غير مترابطة وكل منها تؤدي الى قرار مستقل وفي نهاية الامر فأن القرار النهائي لأسلوب الغابة العشوائية RF سيكون بالغالبية المطلقة لقرارات أشجار الانحدار التي تتكون منها الغابة العشوائية مما يجعل من اسلوب الغابة العشوائية اسلوباً حصينا ضد عدم خطية البيانات وكذلك عدم تجانسها.

ان بيانات الانواء الجوية تعد بشكل عام أحد أنواع السلاسل الزمنية التي تحتوي على العديد من المتغيرات الموسمية وكذلك الدورية التي قد تؤثر سلبا في جعل هذا النوع من البيانات غير متجانسة وكذلك تؤثر في نتائج التنبؤ ودقتها. لذلك ولتحقيق التجانس الى حد كبير في بيانات الدراسة المتمثلة بدرجات الحرارة الصغرى وكذلك كميات التبخر فقد قسمت البيانات الى قسمين وفقا لطبيعة الأجواء في محافظة نينوى. القسم الأول من البيانات يضم الأشهر الحارة ومشاهداتها في حين يضم القسم الثاني الأشهر الباردة. تضم الأشهر الحارة بيانات الأشهر (أيار، حزيران، تموز، آب، أيلول) فيما تضم الأشهر الباردة (تشرين الثاني، كانون الأول، كانون الثاني، شباط، اذار).

قام (Shukur and Lee, 2015) باستخدام أنموذج ARIMA للتنبؤ ببيانات السلسلة الزمنية الخاصة بسرعة الرياح وكذلك استخدم أنموذج ARIMA مع أساليب أخرى ذكائية ضمن نماذج هجينة للتنبؤ وكذلك لتقدير القيم المفقودة في السلاسل الزمنية وقد حصل الباحث على نتائج جيدة عند استخدامه أنموذج ARIMA. واقترح (Chen et. at, 2012) أنموذج للتنبؤ والذي اعتمد على طريقة الغابة العشوائية للتنبؤ ببيانات السلسلة الزمنية لمؤشر هطول الامطار في حوض نهر هايخة– الصين حيث أظهرت النتائج ان التنبؤ بأنموذج الغابة العشوائية RF يعطي قدرات تنبؤية افضل من أنموذج ARIMA. كما قدم (Kane et. at, 2015) مقترح بتطبيق أنموذج ARIMA وأنموذج الغابة العشوائية RF للتنبؤ ببيانات السلسلة الزمنية الخاصة بعرض انفلونزا الطيور (1N5H) في مصر حيث أظهرت نتائج الدراسة ان أنموذج RF تفوق في الأداء على أنموذج ARIMA .

تناولت هذه البحث التنبؤ لبيانات درجات الحرارة الصغرى وكميات التبخر للموسمين الحار والبارد ولفترتي التدريب والاختبار. تنوعت الأساليب المستخدمة في هذه البحث بهدف حل المشاكل. ان بيانات الانواء الجوية تعد وكما ذكرت دراسات سابقة من البيانات غير الخطية مما يتطلب الامر اقتراح أساليب أكثر تلائماً مع بيانات الدراسة ذلك ان استخدام الأساليب الشائعة مثل أنموذج ARIMA قد يؤدي غالباً الى نتائج غير دقيقة. كذلك فان عدم التجانس في بيانات الدراسة نتيجة لاحتوائها على العديد من الأنماط الموسمية والدورية قد يؤدي كذلك الى الحصول على نتائج غير سليمة.

يهدف هذا البحث على نحو رئيسي الى استخدام أساليب تؤدي للوصول الى تنبؤات أفضل دقة لمتغيرات الدراسة وتتلخص اهم الاهداف في استخدام أسلوب تقسيم البيانات الى قسمين أصغر لضمان تجانس البيانات وإعطاء نتائج ادق ويسمى هذا الأسلوب غالباً أسلوب التراصف الزمني Time stratified. كذلك يعد استخدام أنموذج الغابة العشوائية كطريقة تضمن تحسين دقة نتائج التنبؤ وذلك لاعتمادها في اتخاذ القرار النهائي على غالبية القرارات الفرعية للعديد من أشجار القرار المستقلة عن بعضها أي ان أنموذج الغابة العشوائية يعد اسلوباً محصناً في التعامل مع البيانات غير الخطية وقليلة التجانس مثل بيانات هذه الدراسة.

2. نموذج ARIMA ونموذج الغابة العشوائية

1.2. نموذج ARIMA(p,d q)

سيتم التطرق هنا الى التنبؤ باستخدام أنموذج (ARIMA) وأنموذج الغابة العشوائية. يعد أسلوب بوكس جنكيز (Box-Jenkins) اساسا في تحليل السلاسل الزمنية والتعرف على أنموذج (ARIMA) للتنبؤ في بيانات السلسلة الزمنية. ومن ثم التطرق الى مفهوم الغابة العشوائية والطرق المستخدمة للتنبؤ مع كيفية استخدام بعض المقاييس لحساب دقة التنبؤات (Box ,et. at, 2015).

تعرف السلسلة الزمنية بأنها مجموعة من المشاهدات يتم جمعها من ظاهرة معينة في فترات زمنية معينة وغالباً ما تكون هذه الفترات متساوية كأن تكون (يوم، أسبوع شهر، سنة، ، ... الخ) وتكون من متغيرين احدهما (مستقل) وهو متغير الزمن والأخر تابع (معتمد) حسب الظاهرة المدروسة، حيث ان الهدف من تحليل السلاسل الزمنية هو تكوين أنموذج لتفسير سلوك السلسلة الزمنية واستحصال النتائج وذلك بالتنبؤ بسلوك السلسة المستقبلي وبالاعتماد على البيانات الماضية.

يعد أنموذج ARIMA (p,d q) من ابرز واشهر السلاسل الزمنية الغير المستقرة (wei, 2006) حيث ان (p) يشير الى رتبة أنموذج الانحدار الذاتي و(d) يمثل الفروق اللازمة لتحقيق الاستقرارية و (q) يمثل الى رتبة المتوسطات المتحركة والصيغة العامة له:

$$\emptyset_p(B)(1-B)^d Z_t = \theta_q(b)a_t \quad (1)$$

$$\emptyset(B)W_t = \theta(B)a_t \quad (2)$$

اذ ان

$$W_t = (1-B)^d Z_t \quad (3)$$

حيث ان $\emptyset_p$هي معلمة أنموذج الانحدار الذاتي (MA) وان B هو عامل الازاحة الخلفي وان a_{t-k}تمثل الأخطاء او التغيرات العشوائية اعتماداً على a_tبفرض ان التغيرات العشوائية هي عمليات تشويش ابيض بوسط حسابي صفر وتباين ثابت ويمكن كتابته:

$$(1-\emptyset_1 B-\emptyset_2 B^2-\cdots-\emptyset_P B^p)W_t=\left(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q\right)a_t \qquad (4)$$

Or

$$W_t=\emptyset_1 W_{t-1}+\emptyset_2 W_{t-2}+\cdots+\emptyset_p W_{t-p}-\theta_1 a_{t-1}-\theta_2 a_{t-2}-\cdots-\theta_q a_{t-q}+a_t \qquad (5)$$

حيث تعتبر النماذج AR وMA وARMA حالات خاصة من نماذج (ARIMA) على فرض السلسلة الزمنية مستقرة بثبات التباين وخلوها من الاتجاه العام.

ويمكننا ان نعبر عن أنموذج AR(p) بأنه أنموذج ARIMA (p, 0, 0) وعن أنموذج MA(q) بأنه أنموذج ARIMA(0, 0, q) وبعد اخذ الفروق الملائمة وبرتبة ملائمة لها يتم اللجوء الى استخدام الأساليب نفسها لنماذج السلاسل الزمنية المستقرة (Wei,1990) .(Liu,2006)

2.2. نموذج $\mathbf{ARIMA}(P,D,Q)_S$ الموسمي

هو أحد نماذج ARIMA الذي يستخدم عندما تكون البيانات غير مستقرة وذلك باحتوائها على تأثيرات موسمية وتتم ازالتهما الفروق الموسمية حيث ان:

(P) يشير الى عدد معلمات الانحدار الذاتي الموسمي

(D) يمثل عدد الفروق الموسمية

(Q) يشير الى عدد معلمات المتوسطات المتحركة الموسمية، وان:

(S) تمثل الفترة الدورية الموسمية التي تعيد السلسلة فيها نفس الدورة الموسمية والصيغة العامة لها:

$$\Phi_P(B^s)(1-B^s)^D Z_t=\Theta_Q(B^s)a_t \qquad (6)$$

Or

$$\Phi_P(B^s)W_t=\Theta_Q(B^s)a_t \qquad (7)$$

حيث ان:

$$W_t=(1-B^s)^D Z_t \qquad (8)$$

حيث ان $\emptyset_p$هي معلمة أنموذج الانحدار الذاتي الموسمي وان (Θ_Q)هي معلمة أنموذج المتوسطات المتحركة الموسمية أي ان:

$$\Phi(B^s)=\left(1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_p B^{PS}\right)$$
$$\Theta(B^s)=\left(1-\Theta_1 B^s-\Theta_2 B^{2s}-\cdots-\Theta_p {B_Q}^s\right)$$

وبفرض ان التغيرات العشوائية هي عمليات تشويش ابيض بوسط حسابي صفر وتباين ثابت $a_t\sim i.i.d.N(0,{\sigma_a}^2)$ ويستخدم أنموذج ARIMA الموسمي مع التغيرات الموسمية والتي تتغير بتكرار بانتظام خلال فترة زمنية لا تتعدى السنة اما تكون يومية او أسبوعية او شهرية او فصلية (ربع سنوية) ويرجع ظهور هذه التغيرات الى الظروف الطبيعية على مدار السنة ويرمز لها بالرمز (S)

3.2. نموذج ARIMA (p, d, q) $(P,D,Q)_S$ المضاعف

هو احد نماذج (ARIMA) الأكثر تعميما وشمولا حيث يضم فيه المعلمات الموسمية وغير الموسمية والفروقات الخاصة بالنمطين. ويمكن كتابته بشكل عام وكالاتي: −

$$\emptyset(B)\ \Phi(B)((1-B)^d z_t=\theta(B)\ \Theta(B)a_t \qquad (9)$$
$$\emptyset(B)\ \Phi(B)W_t=\theta(B)\ \Theta(B)a_t \qquad (10)$$

حيث ان:

$$W_t=(1-B)^D(1-B)^d Z_t$$
$$\emptyset(B)=(1-\emptyset_1 B-\emptyset_2 B^2-\cdots-\emptyset_p B^p)$$
$$\Phi(B^S)=(1-\Phi_1 B^S-\Phi_2 B^{2S}-\cdots-\Phi_P B^{PS})$$
$$\theta(B)=(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_P B^q)$$
$$\Theta(B^S)=(1-\Theta_1 B^S-\Theta_2 B^{2S}-\cdots-\Theta_q B^{QS})$$

4.2. منهجية بوكس جنكنز في تحليل السلاسل الزمنية

(Barker, 1998), (Box and Jenkins, 1976)

تسمى هذه المنهجية بأسلوب (بوكس−جنكنز) المتكرر في نمذجة السلاسل الزمنية

وقدم كل من بوكس وجنكز عام (1976) أربع خطوات منهجية مميزة ومتسلسلة تباعا هي:

الخطوة الأولى: −التعرف على أنموذج افتراضي تجريبي للسلسلة الزمنية، فأن التعرف يشمل تحقيق شروط الاستقرارية الضعيفة للسلسلة تحت البحث، تم تحديد رتب متعددات الحدود لنماذج السلاسل الزمنية ARIMA(p,d,q)(P,D,Q) المحولة.

الخطوة الثانية: –تقدير معلمات الأنموذج التجريبي الذي تم تحديده والتعرف عليه في الخطوة الأولى.

الخطوة الثالثة: –اجراء فحوص تشخيصية عديدة على الأنموذج لاختبار مدى ملائمته التي اجتازها فهو الأنموذج المطلوب وان كان غير ذلك وظهر نقص في تطابقه فتعاد دورة تكرارية أخرى.

(تعرف←تقدير←فحوص تشخيصية)

الخطوة الرابعة: –تطبيق أنموذج السلسلة الملائم بعد اجتيازه الخطوات الثلاثة السابقة جميعها والتنبؤ لبيانات السلاسل الزمنية.

اولا: التعرف: Identification: (Liu, 2006) (فاندل، 1992) (pankratz, 1983)

للتعرف على الأنموذج الأفضل من نماذج ARIMA وتحديده فلا بد ان يضم اقل عدد ممكن من المعلمات ويمكن تلخيص الخطوات الأنموذجية للتعرف على أي أنموذج على النحو التالي: –

التوقيع البياني للسلسلة الزمنية: عند تحليل السلسلة الزمنية فمن الضروري رسم السلسلة الزمنية بيانياً وذلك للتعرف على العديد من ملامحها ولاسيما تحديد فيما إذا كانت السلسلة الزمنية مستقرة او غير مستقرة إضافة الى الملامح الأخرى. ويعد الارتباط الذاتي والارتباط الذاتي الجزئي من الأدوات المفيدة لبيان مدى استقرارية السلسلة الزمنية.

تحقيق الاستقرارية للسلسلة الزمنية: –تكون السلسلة الزمنية مستقرة إذا امتلكت وسطاً حسابياً وتبايناً ثابتاً في كثير من الحالات تكون السلاسل الزمنية غير مستقرة ويعود السبب في ذلك اما في سبب تغيير في الوسط الحسابي عبر الزمن أي تمتلك اتجاهاً عاماً او بسبب تغيير في تباين السلسلة عبر الزمن. فاذا كانت السلسلة الزمنية غير مستقرة يمكننا تحقيق الاستقرارية الضعيفة فيها او في بعض الأحيان نسميها الاستقرارية من الدرجة الثانية (Chan, 2004, kitagawa; 2010; palma; 2007)

ويمكن تلخيص شروط الاستقرارية الضعيفة بالنقاط التالية:

أ– استقرارية الوسط الحسابي واستقلاليته عن الزمن:

$$E(Z_t) = \mu_t = \mu_{t+k}\, \mu = \frac{1}{n}\sum_{t=1}^{n} Z_t \qquad (11)$$

حيث ان μ هو الوسط الحسابي و $\mu = \mu_{t-k}$هو الوسط الحسابي لكل من Z, Z_{t+k}

ب– استقرار التباين واستقلاله عن الزمن

$$E(Z_t - \mu)^2 = Var\,(Z_t) = {\sigma_{Z_t}}^2 = {\sigma_{Z_{t+k}}}^2 = \sigma_Z{}^2 = \frac{1}{n}\sum_{t=1}^{n}(Z_t - \mu)^2 \qquad (12)$$

حيث ان ${\sigma_z}^2$هو التباين المحدد وان${\sigma_{Z_{t+k}}}^2, {\sigma_{Z_t}}^2$يمثلان التباين للمتغيرين z_{t-k}, Z_t

ت– استقرار دالة التغاير الذاتي بحسب الزمن واعتمادها فقط على الفجوة الزمنية بين المشاهدات.

$$E[(Z_t - \mu)(Z_{t+k} - \mu)] = Cov(Z_t, Z_{t-k}) = \gamma_k \qquad (13)$$

$$\gamma_{(k)} = Cov(Z_t, Z_{t-k}) = \gamma_{(k)} = \gamma \qquad (14)$$

حيث ان γ هي التباين المشترك وان $\gamma_{(k)}$هي التباين المشترك بين المتغيرين Z_t, Z_{t-k}

3 –تحديد رتب متعدد الحدود (q,p): بعد تحقيق استقرارية السلسلة يتم البدء بالتعرف على ملامح السلسلة وتحديد رتب متعددات الحدود في أنموذج (ARIMA) وعدد المعلمات (p, q, P, Q) والجدول ادناه يوضح منهجية مبسطة لتحديد رتب متعددات الحدود في (ARIMA) وعدد المعلمات في الأنموذج من خلال دالتي (PACF) (ACF)

جدول (1) دالتي الارتباط الذاتي والارتباط الذاتي الجزئي لأنواع نماذج (ARMA)

الأنموذج	(ACF)	(PACF)
AR (P)	تقترب من الصفر تدريجياً وتساوي الصفر بعد الارتباط الذاتي (q)	تساوي الصفر فجأة بعد الفجوة الزمنية p
MA(q)	تساوي الصفر فجأة بعد الفجوة الزمنية q	تقترب من الصفر تدريجياً
ARMA(p, q)	تقترب من الصفر تدريجياً وتنقطع الفجوة الزمنية q بعد اخذ d من الفروق	تقترب من الصفر تدريجياً وتنقطع الفجوة الزمنية p بعد اخذ d من الفروق

جدول (2) دالتي الارتباط الذاتي والارتباط الذاتي الجزئي لأنواع نماذج (ARMA)الموسمي

الأنموذج	(ACF)	(PACF)
AR (P)	تقترب من الصفر تدريجياً	تنقطع الى الصفر فجأة بعد الارتباطات الذاتية الجزئية (PS)

MA(Q)	تنقطع الى الصفر فجأة بعد الارتباطات الذاتية (QS)	تقترب من الصفر تدريجياً
SARMA(P,Q)	تنقطع الى الصفر فجأة بعد الارتباطات الذاتية (QS)	تنقطع الى الصفر فجأة بعد الارتباطات الذاتية الجزئية (PS)

قدم (Box and Jenkins 1976) الارتباط الذاتي الجزئي كأداة ضرورية لتحديد افضل رتب لنماذج (ARIMA) ينطوي بمفهوم الارتباط الذاتي الجزئي على الارتباط الشرطي بين Z_t, Z_{t-k} فقط بوجود وبثبوت بقية المتغيرات أي من دون تأثيرات ويرمز له بالرمز ϕ_{kk}.

الخطوة الثانية: تقديرات معلمات الأنموذج: Estimating the parameters of the model

بعد قيامنا بالمرحلة او الخطوة الأولى وهي التعرف على أنموذج (ARIMA) الافتراضي بطريقة بوكس جنكيز بعد ذلك سوف نقوم بالخطوة الثانية الا وهي تقدير معالم الأنموذج وذلك بتنظيم دالة الإمكان (Likelihood Function) للأنموذج، يشار الى مثل هذه التقديرات بتقديرات الإمكان الأعظم (Maximum Likelihood Estimates) حيث يمكن كتابة أنموذج (ARIMA) بالصيغة العامة له:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} + a_t \qquad (15)$$

حيث ان $W_t = (1-B)^d Z_t$ للسلسلة غير المستقرة Z_t ويستخدم المتجه (W) لـ (n) من المشاهدات حيث ان (n) تمثل عدد المشاهدات بعد تحقيق استقرارية السلسلة. والأنموذج السابق يمكن كتابته بالصيغة التالية:

$$a_t = W_t - \phi_1 W_{t-1} - \phi_2 W_{t-2} - \cdots - \phi_P W_{t-P} + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} \qquad (16)$$

حيث ان (a_t) يمثل التشويش الأبيض او الخطأ العشوائي

عندما يكون التباين ثابتاً والوسط الحسابي صفرا حيث ان دالة الكثافة الاحتمالية للأخطاء هي:

$$P(a|\phi,\theta,\sigma_a{}^2) = (2\pi\sigma_a{}^2)^{\frac{-n}{2}} \exp[-\frac{1}{2\sigma_a{}^2}\sum_{t=1}^{n} a_t{}^2] \qquad (17)$$

حيث ان

$$\phi = \phi_1, \phi_2, \ldots, \phi_P,$$
$$\theta = \theta_1, \theta_2, \ldots, \theta_q$$
$$a = a_1, a_2, \ldots, a_n,$$

وان دالة الكثافة الاحتمالية لـ W يمكن كتابتها بالصيغة التالية:

$$P(W|\phi,\theta,\sigma_a{}^2) = (2\pi\sigma_a{}^2)^{\frac{-n}{2}} |\Sigma|^{\frac{-1}{2}} \exp[-\frac{1}{2\sigma_a{}^2} w' \Sigma^{-1} w] \qquad (18)$$

حيث ان (Σ) هي دالة θ و Φ وان $(w^{-1}\Sigma^{-1}w)$ هي مجموعة المربعات للدالة التي تحوي ϕ و θ ويرمز لها بالرمز $S(\emptyset,\theta)$ وان $\Sigma = E(ww')$ $\sigma_a{}^2$ هي مصفوفة التباين والتباين المشترك للمتجه (w) حيث يتم الحصول على مقدرات الإمكان الأعظم (MLE) بتعظيم دالة الإمكان او نحصل عليها بعد اخذ اللوغاريتم الطبيعي لدالة الإمكان:

$$\ln L(\phi,\theta,\sigma_a{}^2|W) = \frac{-n}{2}\ln(2\pi\sigma_a{}^2) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2\sigma_a{}^2}S(\phi,\theta) \qquad (19)$$

(Cryer and Chan,2008)

الخطوة الثالثة: الفحص التشخيصي: Diagnostic checking

بعد ان تعرفنا على الأنموذج وتم تقدير معلماته ففي هذه الخطوة سيتم التأكد من دقة الأنموذج وملائمته ومعرفة فيما إذا كانت المعلمات الأنموذجية معنوية حيث ان هنالك العديد من الأدوات للفحص التشخيصي منها:

1- معنوية المعلمات المقدرة من الجانب الاحصائي يشترط معنوية مقدرات معلمات الأنموذج جميعها حيث ان المعلمات غير المعنوية تعتبر من الأسباب المخلة بدقة الأنموذج حيث سيتم اختبار فرضية العدم والتي تنص على ان مقدرات المعلمات لا تختلف معنوياً عن الصفر أي تساوي الصفر اذ ان القيمة الحرجة لاختبار (t) هي القيمة الجدولية مضروبة بالخطأ المعياري المقدر للمعلمة، وان القيم الجدولية تختلف باختلاف مستوى المعنوية والذي يختلف باختلاف حجم السلسلة الزمنية، وغالباً ما تستخدم (∝= 0.05) والقيمة الجدولية لها هو (1.96) في الاختبارات والتي تناسب البيانات الكبيرة جداً فاذا كانت القيمة المطلقة للقيمة المحسوبة لاختبار (t) لكل مقدر تساوي على الأقل القيمة الحرجة فعند ذلك سوف نرفض فرضية العدم أي ان (المقدر المعنوي)، اما اذا كانت القيمة المحسوبة لاختبار (t) اكبر من القيمة الحرجة فعند ذلك سوف ترفض فرضية العدم وتقبل الفرضية البديلة. وبالتالي يعتبر هذا مؤشر على إمكانية تبسيط الأنموذج وذلك بتخفيض عدد معلماته من خلال حذف المعلمات غير المعنونة من الأنموذج (المقدر غير المعنوي) وهو المقدر ذو الرتبة الأعلى في الأنموذج فيتم تبسيط الأنموذج وذلك بحذف هذا المقدر.

2- حالة الارتباط الذاتي لسلسلة البواقي: SACF of Residuals Series

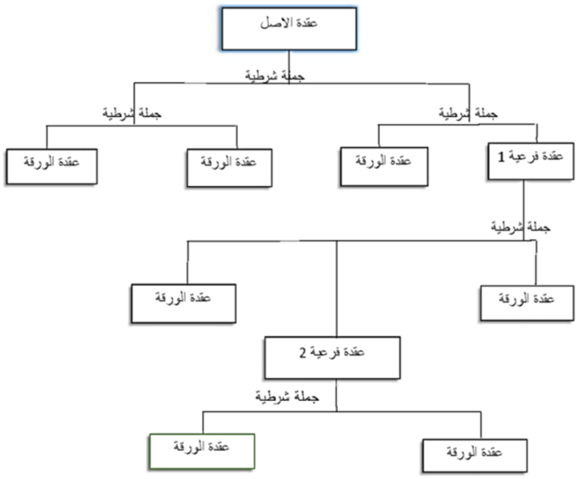
من الممكن ان نستخدم (ACF) للبواقي وذلك لاختبار فيما إذا كانت سلسلة البواقي مطابقة وموافقة لعملية التشويش الأبيض $[a_t \sim \mathrm{i.i.}d\ N(0,\sigma_a^{\,2})]$ فاذا كانت سلسلة البواقي ذات تشويش ابيض فيتوجب ذلك بان دالة الارتباط الذاتي للبواقي ان لا تحتوي على معاملات ارتباط معنوية كذلك لاختبار معنوية معاملات الارتباط الذاتي يجب تقدير الانحراف المعياري لمعاملات الارتباط الذاتي للبواقي ثم بعد ذلك ضربها بالقيمة الجدولية (1.96) لتحديد مدى المعنوية عند ثقة (0.95) (Shukur,2015).

الخطوة الرابعة: التنبؤ Forecasting

في هذه المرحلة سوف نقوم بالتنبؤ بالمشاهدات المستقبلية للسلسلة الزمنية بعد عبور او اجتياز مرحلة الفحص التشخيصي بنجاح. على فرض ان (n) تشير او تمثل الفترة الزمنية الحالية لذا يجب ان يكون التنبؤ لمشاهدة تحدث بعد (1) فترة زمنية الى الامام وان هذه المشاهدة سوف يرمز لها بالرمز (Z_{n+1})والتي لم تحدث بعد، علماً ان التنبؤ لقيمة منفردة لكل فترة زمنية يسمى بـ (التنبؤ بنقطة Point Forecasting) كما يمكن تنبؤ بحدود ثقة حول كل تنبؤ نقطي والذي يدعى (التنبؤ بفترة Interval Forecasting) وسوف نختار طريقة تنبؤات اقل متوسط مربعات خطأ (Minimum Mean Squares Error) (MMSE) والتي تستخدم أنموذج ARIMA (p, d, q) العام.

5.2. الغابة العشوائية (RF) Random Forest

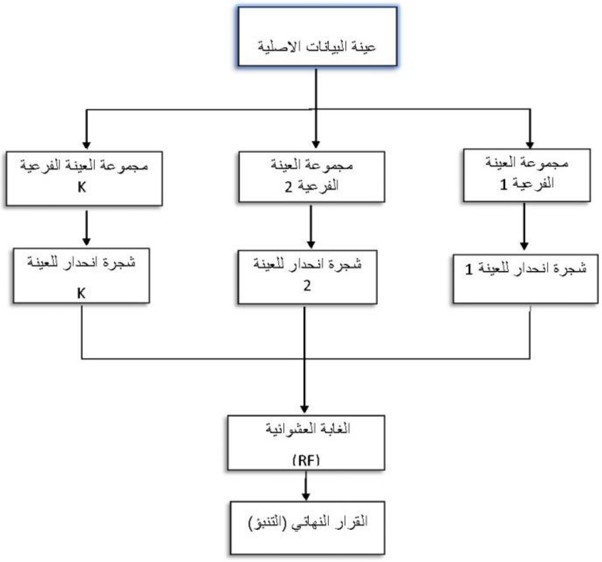
الغابة العشوائية هي احدى خوارزميات التعلم الخاضعة للأشراف Supervised أي ان مخرجات الغابة العشوائية يجب ان تتطابق مع متغيرات الهدف وبمقارنتها تنتج أخطاء التنبؤ وتعتمد على مبدأ تقنيات أشجار التصنيف والانحدار ومن مميزاتها انها دقيقة حسابياً وتعمل بسرعة وذلك عبر بيانات كبيرة نسبياً وهي من التقنيات الحديثة حيث يتم استخدامها في العديد من التطبيقات في مجالات متنوعة لاعتمادها على مبدأ التصنيف والانحدار فهي عبارة عن مخطط لمجموعة أشجار تستخدم لبناء أنموذج يعطي تنبؤات من خلال اوراقها الناتجة عن مساحات وتفرعات مختارة عشوائياً من البيانات بمبدأ مشابه لبديهيات أشجار الانحدار (Shumway et. At. 2010)، الشكل(1) يوضح هيكلية الغابة العشوائية كأحد أنواع أشجار الانحدار .



شكل(1) :هيكلية الغابة العشوائية كأحد أنواع أشجار الانحدار

كل تفرع في الشجرة في الشكل (1) يمثل نقطة قرار تم اتخاذها على أساس جملة شرطية وهكذا تستمر التفرعات لحين الوصول الى القرارات النهائية المتمثلة بعقد الأوراق حيث ان كل ورقة تعتبر كعقدة منفصلة من قرار منفصل عن باقي الأوراق وان هذه الأشجار تعطي تطابق امثل بين المخرجات المتمثلة بالتنبؤات بالمقارنة مع المتغير الأصلي الذي تم اعتباره كمتغير هدف، أي سيتم تطوير أسلوب التنبؤ والحصول على تنبؤات مثلى بأقل أخطاء للتنبؤ عند استخدام أسلوب (RF) كأحد تقنيات أشجار التصنيف والانحدار مقارنة بالأساليب التقليدية للتنبؤ. توفر نمذجة السلاسل الزمنية باستخدام الغابات العشوائية قدرة تنبؤية معززة وأكثر دقة مقارنة بنماذج السلاسل الزمنية التقليدية للتنبؤ خصوصاً ببيانات الأرصاد الجوية وبيانات أخرى كثيرة على العموم. ان الشكل(1) السابق يوضح مبدأ عمل خوارزمية الغابة العشوائية كاحد أشجار التصنيف والانحدار الذي يستخدم في التنبؤ اما الإطار الاشمل الذي يتميز به أسلوب الغابة العشوائية فهو أكثر تعميماً عن أشجار الانحدار والتصنيف وذلك لاعتماده على مبدأ تقسيم عينة بيانات الدراسة الى عدة عينات فرعية (Bootstrap Sample Sets) وذلك لإخذ جميع الأنماط السلوكية لعينة الدراسة في جميع الفترات المختلفة والحصول على شجرة انحدار لكل عينة فرعية ومن ثم فأن مجاميع هذه الأشجار سوية سوف تمثل ما يسمى بالغابة العشوائية (RF) وان القرار النهائي

يكون مستنبطاً من خلال غالبية عقد الأوراق لجميع أشجار الانحدار، الشكل (2) يوضح الإطار العام لخوارزمية عمل الغابة العشوائية . (RF)



الشكل (2): الإطار العام لخوارزمية عمل الغابة العشوائية (RF) .

هنالك احتمال ان تكون الأشجار في الغابة العشوائية مترابطة فيما بينها بحسب الشكل (2) فأنها عائدة الى نفس نوع البيانات وكذلك تم اعتماد مبدأ التعبئة (bagging principle) الذي اساسه هو عملية المعاينة التمهيدية (Bootstrap Sampling) اذ تعمل طريقة التعبئة على تحسين أداء أشجار التصنيف والانحدار وتجعل (RF) أكثر حصانة عند تجميعها مع بعضها. يتم معالجة ذلك بجعل الاشجار في الغابة العشوائية غير مترابطة مع بعضها (مختلفة)لذلك فقد قدم (Breiman, 2001) مقترحاً لان تنمو كل شجرة بشكل منفصل وكذلك بشكل عشوائي وباجتماع هذين المبدأين ستحدد ملامح وعدد مجموعات العينات الفرعية المشار اليها في الشكل (2) .بعد تحويل الأشجار في الغابة العشوائية من مترابطة الى غير مترابطة (مختلفة) مما سيضمن زيادة ملحوظة في دقة تنبؤ الغابة العشوائية.

يتم بناء خوارزمية الغابة العشوائية باستخدام الخطوات الثلاثة ادناه:

1- من بيانات التدريب يتم استخراج B من العينات التمهيدية والتي هي في الأصل مترابطة فيما بينها اذ ان B تمثل حجم الغابة او عدد الأشجار المتعددة المشار اليها في الشكل (2)

2- لكل مجموعة من مجموعات البيانات B فأن نمو الشجرة T_bسيتم باتباع خطوات متسلسلة في كل عقدة من عقد الشجرة لحين الوصول الى n_{min} والتي تمثل الحد الأدنى من أوراق الأشجار او عدد العقد وكما يلي:

أ- اختيار m والتي تمثل العدد المختار عشوائياً من التنبؤات في كل قسم من العدد الكلي للمتغيرات p.

ب- اختيار أفضل التنبؤات من التنبؤات المختارة في (أ) وقد تم الإشارة اليها بالرمز m مع اختيار القسم العائدة اليه بهدف تقليل قيمة Mse للتنبؤات المختارة في (أ).

ج- فصل العقدة الى عقدتين فرعيتين تبعاً للمعيار المستخدم او القيم التنبؤية الأفضل التي تم اختيارها في (ب).

3- استخلاص المخرجات من جميع الأشجار من خلال إيجاد المجموعة $\{T_b\}_1^B$وأخيرا فأنه عند نقطة معينة X فأن التنبؤ ممكن ان حسب المعادلة التالية: (Noureen,et,at.2019)

$$f_{RF} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \ldots \qquad (20)$$

2.8 مقاييس خطأ التنبؤ Foracasting Error Measurements

للمقارنة بين الطرائق المقترحة سيتم استخدام العديد من مقاييس الخطأ وفي اغلب الدراسات يتم استخدام مقياس للخطأ RMSE الجذر التربيعي لمتوسط المربعات الخطأ وMAEمتوسط الخطأ المطلق النسبي.

وهذه المقاييس يمكن ان تقسم الى مقاييس تصف تشتت البيانات وأخرى تصف الدقة والنسبة المئوية للخطأ. RMSE يقيس عادة التشتت و MAPE يمثل عادة النسبة المئوية لخطأ التكهن ودقته.

بحسب مقياس mean absolute percentage error (MAPE) متوسط القيمة المطلقة للنسبة المئوية للخطأ على النحو التالي:

$$MAPE = \frac{1}{n}\sum_{i+1}^{a}\left|\frac{e_i}{z_i}\right| \times 100 \tag{21}$$

اذ ان r تمثل خطأ التكهن، n عدد المشاهدات و z_i هو السلسلة الحقيقة او الاصلية المستعملة كهدف. اما مقياسي mean absolute error (MAE) متوسط القيمة المطلقة للخطأ و root mean squares error (RMSE) الجذر التربيعي لمتوسط مربعات الخطأ فيمكن كتابة الصيغة الرياضية لهما كما يلي: (Hyndman & Koehler, 2006)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i)^2} \tag{22}$$

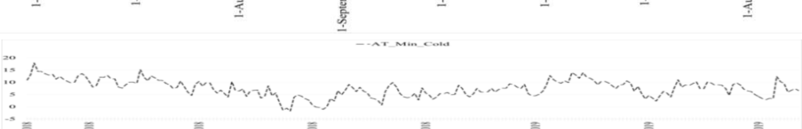
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|e_i| \tag{23}$$

عندما N:عدد مشاهدات العينة و e_i مقدار الخطأ والذي يمثل الفرق بين متغير القيم الحقيقية ومتغير القيم التنبؤية.

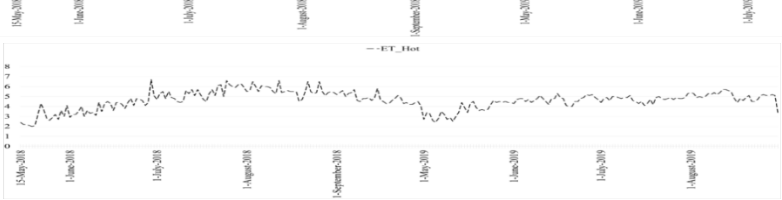
3. النتائج والمناقشة

تم تناول نوعين من البيانات تضمنت المجموعة الأولى درجات الحرارة الصغرى لمدينة الموصل والتي نم اخذها من مركز الأرصاد الجوية الزراعية/ محافظة نينوى/ محطة الموصل التابعة لوزارة الزراعة في الموقع المحدد بخط الطول E : 43.16وخط العرضN : 36.33. وتضمنت المجموعة الثانية كمية التبخر (mm) مأخوذة من نفس المحطة المشار اليها سابقاً تضمنت مجموعتي البيانات (675) مشاهدة للفترة من (15/5/2018) ولغاية (19/7/2020) ولوحظ احتوائها على بيانات يمكن وصفها بانها غير متجانسة وذلك للتنوع الذ تحتويه البيانات يمكن وصفها بانها غير متجانسة وذلك للتنوع الذي تحتويه البيانات من خلال مرورها بالفصول الرسمية الأربعة وتقلباتها من حيث البرودة والحرارة وغيرها من التقلبات الجوية كما ان ذلك واضح بعد رسم الاتجاه العام. ولتحقيق انسجام أكبر للبيانات فقد تم تقسيمها الى مجموعتين الأولى للموسم البارد ويضم الأشهر (تشرين الثاني–كانون الأول–كانون الثاني–شباط–اذار) والمجموعة الثانية خاصة بالموسم الحار والذي يضم الأشهر (أيار–حزيران–تموز–اب–أيلول). تم تقسيم البيانات في كل مجموعة الى مجموعتين جزئيتين هما التدريب والاختبار وذلك للتحقق من صحة ثبوتها من خلال اختبار الأنموذج الذي يتم بناءه ببيانات بالتدريب وذلك باستخدام بيانات مجموعة الاختبار للتحقق من صحة أداء الأنموذج عادة ما يتم افتراض النسبتين 70% و 30% لمجموعتي بيانات التدريب والاختبار على التوالي من العدد الكلي لمشاهدات السلسلة الزمنية. لذلك تم تقسيم بيانات الموسم البارد الذي يضم (303) مشاهدة الى (212) مشاهدة لمجموعات التدريب و (91) مشاهدة لمجموعات الاختبار. اما فيما يخص الموسم الحار فقد تم تقسيم بياناته التي تضم (372) مشاهدة الى (262) مشاهدة لمجموعة التدريب و (110) مشاهدة لمجموعة الاختبار. أن استقرارية الأنموذج يتم التحقق منها من خلال رسم السلسلة الزمنية وكل من دالتي الارتباط الذاتي والارتباط الذاتي الجزئي. اما رسم السلسلة وتوقيعها بيانياً يجب ان تظهر من خلالها السلسلة الزمنية منسجمة ومتناسقة وخالية من القيم الشاذة والمتطرفة ويكون فيها المتوسط والتباين مستقرين اما دالتي الارتباط الذاتي والارتباط الذاتي الجزئي فتستخدمان لتأكيد التحقق من الاستقرارية من خلال نوع الاضمحلال فعندما يكون الاضمحلال عند نحو عدم المعنوية بطيئاً أي بعد أكثر من (6) ارتباطات فعندئذ نتأكد ان السلسلة غير مستقرة اما الاضمحلال السريع فغالبا ما يدل على استقرارية السلسلة. التوقيع البياني للسلاسل الزمنية لفترات التدريب لدرجة الحرارة الصغرى في الموسمين الحار والبارد وكمية التبخر للموسمين الحار والبارد على التوالي مدرجة كما في الشكل(3) ادناه.

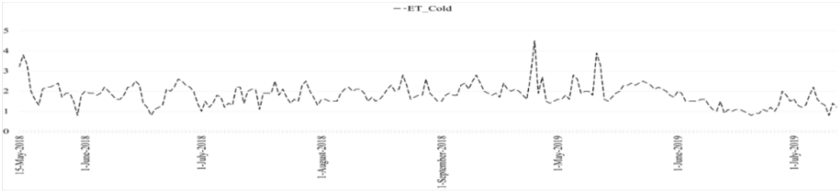
a



b



c



d

الشكل (3) التوقيع البياني لفترات التدريب لدرجة الحرارة الصغرى في الموسمين الحار والبارد وكمية التبخر للموسمين الحار والبارد على التوالي

بعد اخذ العديد من الفروقات الاعتيادية والموسمية مع اختبار السلاسل بعد كل فرق فقد تم التوصل الى الفروقات التالية التي تكفي للوصول الى الاستقرارية للسلاسل الزمنية ولما يلي:

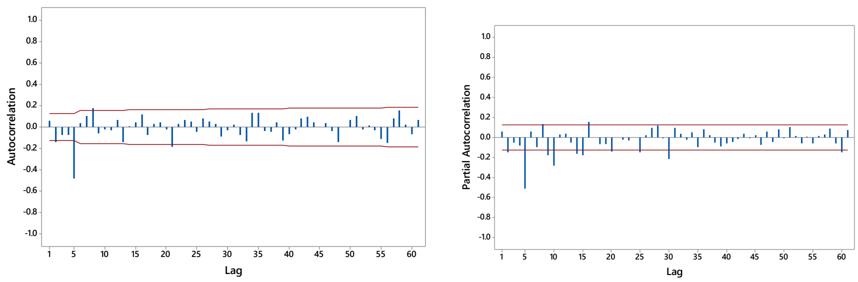
1- السلسة الزمنية لدرجات الحرارة الصغرى/ الموسم الحار: فرق اعتيادي اول d=1 بالإضافة الى فرق موسمي اول D=1 عند S=5

2- السلسلة الزمنية لدرجات الحرارة الصغرى/ الموسم البارد: فرق اعتيادي ثاني d=2 بالإضافة الى فرق موسمي ثاني D=2 عند S=5

3- السلسلة الزمنية لكميات التبخر/ الموسم الحار: فرق اعتيادي ثاني d=2 بالإضافة الى فرق موسمي ثاني D=2 عند S=5

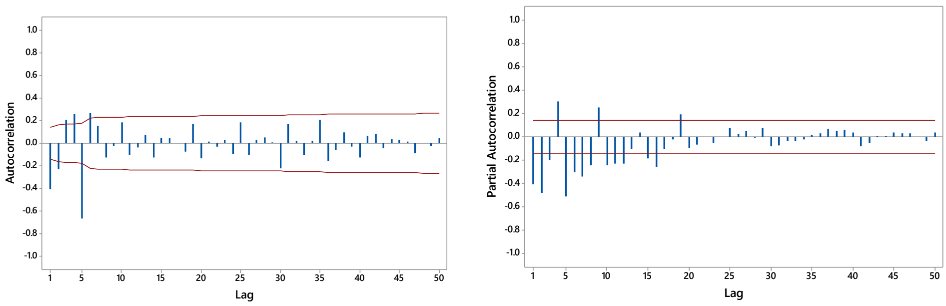
4- السلسلة الزمنية لكميات التبخر/ الموسم البارد: فرق اعتيادي ثاني d=2 بالإضافة الى فرق موسمي ثاني D=2 عند S=5

الاشكال (4) (5) (6) (7) تمثل دالتي الارتباط الذاتي والارتباط الذاتي الجزئي لدرجة الحرارة الصغرى في الموسمين الحار والبارد وكمية التبخر للموسمين الحار والبارد على التوالي للسلاسل الزمنية بعد اخذ الفروقات المشار اليها أعلاه أي بعد تحقيق الاستقرارية للسلاسل الزمنية.



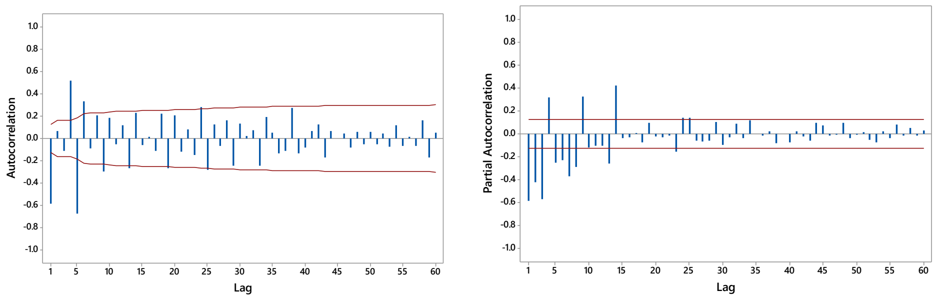
a b

الشكل (4): دالتي الارتباط الذاتي والارتباط الذاتي الجزئي للموسم الحار على التوالي لدرجة الحرارة الصغرى عندما S=5, D=1, d=1



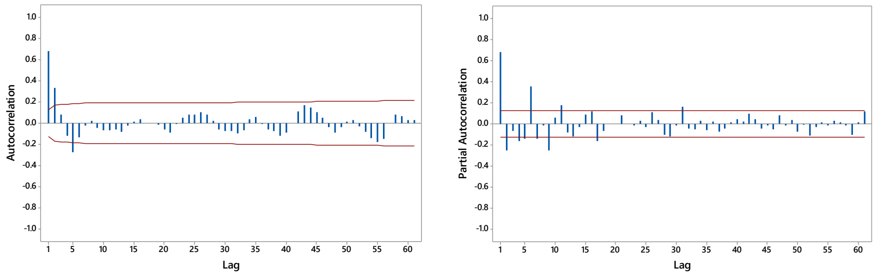
a b

الشكل (5): دالتي الارتباط الذاتي والارتباط الذاتي الجزئي للموسم البارد على التوالي لدرجة الحرارة الصغرى عندما S=5, D=2, d=2



a b

الشكل (6): دالتي الارتباط الذاتي والارتباط الذاتي الجزئي للموسم الحار على التوالي لكمية التبخر عندما S=5, D=2, d=2



a b

الشكل (7): دالتي الارتباط الذاتي والارتباط الذاتي الجزئي للموسم البارد على التوالي لكمية التبخر عندما S=5, D=2, d=2

من خلال الاشكال (4) الى (7) من الممكن استنتاج نماذج ARIMA المناسبة لكل مجموعة من البيانات وكما يلي:

الأنموذج الأول: –ان الأنموذج المناسب لدرجة الحرارة الصغرى للموسم الحار من الممكن استنتاجه من خلال الشكل(4) a و b حيث تشير دالة الارتباط الذاتي الى إمكانية معنوية معلمة واحدة للمتوسطات المتحركة الاعتيادية ومعلمة واحدة لمتوسطات المتحركة الموسمية. اما دالة الارتباط الذاتي الجزئي فتشير الى إمكانية وجود معلمة واحدة معنوية للانحدار الذاتي الموسمي عندما S=5 وبذلك فأن الأنموذج المناسب هو $ARIMA\ (0,1,1)(1,1,1)_5$ والذي يمكن تمثيله حسب الصيغة في المعادلة ادناه:

$$(1-\Phi_1B^5)(1-B)^d(1-B)^D Z_t = (1-\theta_1B)(1-\Theta_1B^S)a_t \qquad (24)$$

عندما $\Phi_1 = -0.0415\,, \theta_1 = 0.6610\,, \Theta_1 = 0.9574, d = 1,\ D = 1,\ S = 5$

حيث ظهر ان المعلمات $\Phi_1, \theta_1, \Theta_1$ معنوية.

الأنموذج الثاني: –ان الأنموذج المناسب لدرجة الحرارة الصغرى للموسم البارد من الممكن استنتاجه من خلال الشكل(5) a و b حيث تشير دالة الارتباط الذاتي الى إمكانية معنوية معلمتين للمتوسطات المتحركة الاعتيادية ومعلمة واحدة للمتوسطات المتحركة الموسمية. اما دالة الارتباط الذاتي الجزئي فتشير الى إمكانية وجود معلمتين معنويتين للانحدار الذاتي الاعتيادي ومعلمتين للانحدار الذاتي الموسمي عندما S=5 وبذلك فأن الأنموذج المناسب هو: $ARIMA\ (2,2,2),(2,2,1)_5$

والذي يمكن تمثيله حسب الصيغة في المعادلة ادناه:

$$(1-\phi_1B_1-\phi_2B_2)(1-\Phi_1B^2-\Phi_2B^{2S})(1-B)^d(1-B)^D Z_t = (1-\theta_1B-\theta_2B)(1-\Theta_1B^s)a_t \qquad (25)$$

عندما $\phi_1 = -0.5011, \phi_2 = -0.1416, \Phi_1 = -0.7405, \Phi_2 = -0.4169,\ \theta_1 = 0.5674,\ \theta_2 = 0.434, \Theta_1 = 0.9489$

حيث ظهر ان جميع المعلمات معنوية.

الأنموذج الثالث: –ان الأنموذج المناسب لكمية التبخر للموسم الحار من الممكن استنتاجه من خلال الشكل(6) a و b حيث تشير دالة الارتباط الذاتي الى إمكانية معنوية معلمة واحدة للمتوسطات المتحركة الاعتيادية وثلاث معلمات للمتوسطات المتحركة الموسمية اما دالة الارتباط الذاتي الجزئي فتشير الى إمكانية وجود ثلاث معلمات معنوية للانحدار الذاتي الموسمي عند S=5 وبذلك فان الأنموذج هو $ARIMA\ (3,2,1)(3,2,3)_5$ و يمكن تمثيله حسب الصيغة في المعادلة ادناه:

$$(1-\phi_1B_1-\phi_2B_2-\phi_3B_3)(1-\Phi_1B^S-\Phi_2B^{2S}-\Phi_3B^{3S})(1-B)^d(1B)^D Z_t$$

$$= (1-\theta_1B)(1-\Theta_1B^S-\Theta_2B^{2S}-\Theta_3B^{3S})a_t \qquad (26)$$

حيث ان

$\phi_1 = -0.4294, \phi_2 = -0.3807, \phi_3 = -0.2638, \Phi_1 = -1.1653,$
$\Phi_2 = -0.4914, \Phi_3 = -0.1948, \theta_1 = 0.9796, \Theta_1 = 0.7858,$
$\Theta_2 = 0.6548, \Theta_3 = -0.4706, d = 2, D = 2, S = 5$

حيث ظهر ان جميع المعلمات معنوية.

الأنموذج الرابع: –ان الأنموذج المناسب لكمية التبخر للموسم البارد من الممكن استنتاجه من خلال الشكل(7) a و b حيث تشير دالة الارتباط الذاتي الى إمكانية معنوية معلمة واحدة للمتوسطات المتحركة الاعتيادية وثلاث معلمات للمتوسطات المتحركة الموسمية اما دالة الارتباط الذاتي الجزئي فتشير الى إمكانية وجود ثلاث معلمات معنوية للانحدار الذاتي الموسمي عندما S=5 وبذلك فأن الأنموذج المناسب هو: $ARIMA\ (3,2,1)(3,2,3)_5$ يمكن تمثيله كما في المعادلة ادناه

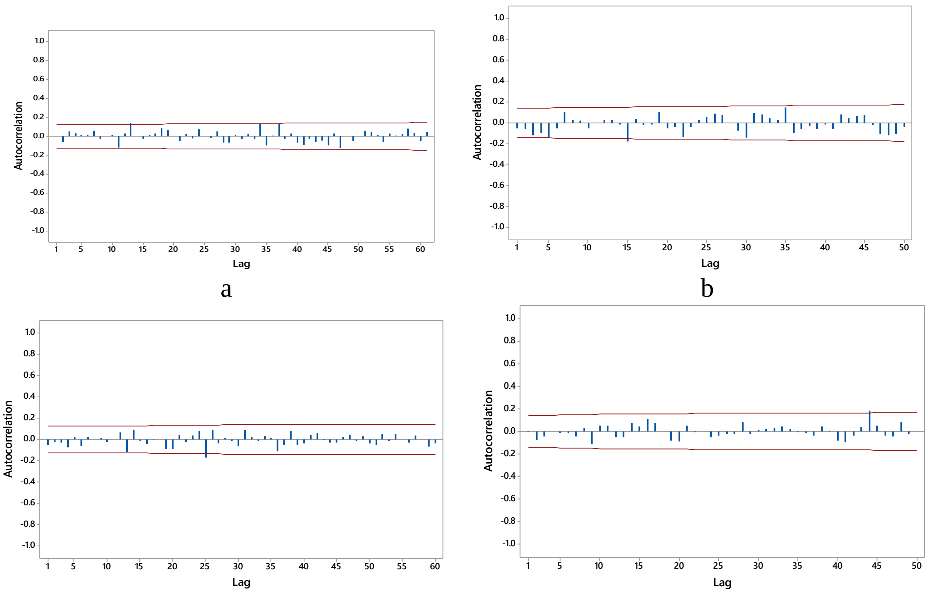
$$(1 - \phi_1 B_1 - \phi_2 B_2 - \phi_3 B_3)(1 - \Phi_1 B^S - \Phi_2 B^{2S} - \Phi_3 B^{3S})(1 - B)^d (1 - B)^D Z_t$$
$$= (1 - \theta_1 B)(1 - \Theta_1 B^S - \Theta_2 B^{2S} - \Theta_3 B^{3S}) a_t \qquad (27)$$

حيث ان:

$\phi_1 = -0.3188, \phi_2 = -0.2658, \phi_3 = -0.1790, \Phi_1 = -0.8660,$
$\Phi_2 = -0.0617, \Phi_3 = -0.0472, \theta_1 = 0.9285, \Theta_1 = 1.1038$
$\Theta_2 = 0.6046, \Theta_3 = -0.7220, d = 2, D = 2, S = 5$

حيث ظهر ان جميع المعلمات معنوية.

الشكل(8) يوضح الارتباطات غير المعنوية للبواقي لنماذج بيانات درجات الحرارة الصغرى في الموسمين الحار والبارد وكمية التبخر للموسمين الحار والبارد على التوالي مما يجعل من هذا الفحص التشخيصي دليل على سلامة النماذج الاربعة أعلاه.



الشكل (8) ACF للبواقي للنماذج الاربعة اعلاه على التوالي

تم احتساب قيم معيار متوسط القيمة المطلقة للنسبة المئوية للأخطاء Mean absolute percentage errors (MAPE) والجذر التربيعي لمتوسط مربعات الأخطاء Root of Mean squared errors (RMSE) ومتوسط القيمة المطلقة للأخطاء Mean absolute errors (MAE) الذي يقيس مدى دقة التنبؤات أي هو مؤشر لمدى أخطاء التنبؤ. والجدول (3) ادناه يوضح قيم معايير الاخطاء للتنبؤات لفترتي التدريب والاختبار باستخدام نماذج ARIMA الأربعة المشار اليها أعلاه.

الجدول (3) معايير(MAPE, RMSE, MAE) للتنبؤات لفترتي التدريب والاختبار للنماذج الاربعة

			MAPE	RMSE	MAE
AT Min	Hot	تدريب	6.9928	2.0711	1.6465
		اختبار	30.9986	6.6103	5.7293
	Cold	تدريب	79.4920	2.2434	1.6802
		اختبار	733.4600	66.5931	49.0294
ET	Hot	تدريب	7.8820	0.4866	0.3561

		اختبار	112.6959	10.8045	8.6510
	Cold	**تدريب**	19.2679	0.4664	0.3348
		اختبار	32.3209	1.1669	0.8229

سيتم الاعتماد على استخدام الايعاز(fitrensemple) في برنامج (MATLAB) لبناء أنموذج الانحدار التجميعي(Regression Ensemple Model) للغابة العشوائية RF باستخدام عدة متغيرات تفسيرية ومتغير واحد معتمد. ان بيانات هذا البحث تتضمن بيانات سلاسل زمنية احادية المتغير(درجات الحرارة الصغرى وكميات التبخر) وسيتم اعتماد مبدأ الارتباط الذاتي في السلاسل الزمنية لانشاء متغيرات تفسيرية من كل متغير من متغيرات الدراسة وذلك من خلال استخدام التخلفات الزمنية للمتغير الاصلي كمتغيرات تفسيرية حيث سيكون لكل متغير من متغيرات الدراسة ثلاث متغيرات تفسيرية (ثلاث تخلفات زمنية) فيما سيكون نفس المتغير الاصلي هو المعتمد. مبدأ عمل الايعاز (fitrensemple) في برنامج(MATLAB) هو كما تم ذكره انفا في بناء أنموذج انحدار تجميعي مع ملاحظة مايلي:

1- اعتماده على خوارزمية المربعات الصغرة التعزيزية Least-Squares Boosting والتي تتضمن ايجاد مجاميع(Ensemples) لأفضل معادلات تلائم بيانات الدراسة. وفي كل خطوة من هذه الخوارزمية(LS Boost) سيتم انجاز تعلم جديد وايجاد معادلة انحدار جديدة ثم ايجاد الفرق بين البيانات الحقيقية للمتغير المعتمد والتنبؤ التجميعي المتراكم من جميع خطوات التعلم السابقة. ان الفائدة المرجوة من هذه الخوارزمية هي تصغير مقياس(MSE) لأخطاء التنبؤ. ان اساس (LS Boost) يعتمد على مبدأ الخوارزمية التجميعية او التراكمية (Ensemple Algorithm) والتي تعرف بأنها احد تقنيات التعلم من خلال بناء نماذج عديدة وتوفيق تلك النماذج للحصول على نتائج افضل. عادة يؤدي استخدام النماذج المجمعة الى حلول ونتائج ادق مما لو استخدمت الاساليب التقليدية التي اساسها أنموذج واحد منفرد.

2- نظرا الى ان الطريقة تعتمد على مبدأ التجميع والتوفيق بين النماذج فان اشجار الغابة العشوائية باستخدام (10) تجزءات للبيانات كعدد افتراضي للإيعاز (fitrensemple) كحد اقصى والتي سيستفاد من توفيقها باستخلاص افضل النتائج. وسيتم استخدام (100) شجرة ثم توفيقها للحصول على افضل التنبؤات.

بعد الانتهاء من بناء أنموذج الغابة العشوائية باستخدام ايعاز(fitrensemple) فالخطوة التالية هي التنبؤ باستخدام هذا الأنموذج الذي يعد هو الأنموذج الامثل للبيانات الدراسة وذلك باستخدام يعاز (Bredict) والذي يتطلب الأنموذج الذي تم بنائه مع بيانات التدريب(المتغيرات التفسيرية فقط) للحصول على التنبؤات الداخلية المقابلة لفترة التدريب والتي تسمى تنبؤات التدريب(Training Forecast) وكذلك في خطوة تالية يتم ادخال نفس الأنموذج الذي تم بنائه مع بيانات الاختبار (المتغيرات التفسيرية فقط) للحصول على التنبؤات في فترة الاختبار والتي تسمى تنبؤات الاختبار (Testing Forecast) والجدول (4) يوضح قيم معايير اخطاء التنبؤ (MAPE) و(RMSE) و(MAE) لبيانات التدريب والاختبار.

الجدول(4) قيم معايير اخطاء التنبؤ (MAPE) و(RMSE) و(MAE) لبيانات التدريب والاختبار

البيانات	الموسم	الفترة	MAPE	RMSE	MAE
AT Min	**Hot**	تدريب	0.0912	0.0281	0.0215
		اختبار	13.0509	3.4350	2.6763
	Cold	تدريب	0.4544	0.0115	0.0085
		اختبار	75.1640	3.0296	2.4288
ET	**Hot**	تدريب	1.6726	0.1359	0.0739
		اختبار	20.7528	1.6414	1.4085
	Cold	تدريب	3.4651	0.1364	0.0635
		اختبار	29.9944	1.0435	0.7729

من خلال الجدولين (3) و(4) يتضح ان هنالك افضلية مطلقة لنتائج التنبؤ لبيانات درجة الحرارة الصغرى وكميات التبخر للموسمين الحار والبارد لفترتي التدريب والاختبار باستخدام أنموذج الغابة العشوائية مقارنة بنفس نتائج التنبؤ باستخدام الأنموذج التقليدي ARIMA اي ان أنموذج الغابة العشوائية اسهم كثيرا بتحسين نتائج التنبؤ وذلك لأنه يأخذ بنظر الاعتبار عدد اوسع من الاحتمالات باعتماده على اشجار تنبؤ عديدة ومن ثم اختيار افضل نتائج التنبؤ وبذلك يحقق تحسينا كبيرا في التنبؤ لبيانات الدراسة.

4. الاستنتاجات

على الرغم من ان أنموذج ARIMA يعد من النماذج شائعه الاستخدام في تطبيقات واسعه ومتنوعه للتنبؤ بالسلاسل الزمنية الا انه يفتقر الى التعامل مع البيانات غير الخطية وبالتالي سيؤدي استخدامه مع هكذا نوع من البيانات الى نتائج تنبؤ غير دقيقه خصوصا ان استخدامه مع بيانات الانواء الجوية مثل درجات الحرارة وكميات التبخر وغيرها التي تعد من البيانات غير الخطية كما اشار الى ذلك الكثير من الدراسات السابقة.

ان استخدام أنموذج الغابة العشوائية مع بيانات الانواء الجوية خصوصا بيانات درجات الحرارة الصغرى وكميات التبخر التي تعتبر من البيانات غير الخطية سيؤدي الى تحسينات ملحوظه في دقه نتائج التنبؤ والحصول على تنبؤات دقيقه جدا مقارنه بنتائج التنبؤ باستخدام الطرائق التقليدية مثل أنموذج ARIMA وذلك لان أنموذج الغابة العشوائية يعد من الاساليب غير الخطية بالإضافة الى اعتباره احد اساليب تعلم الاله الحديثة لذلك سيعطي دقه عالية في التنبؤ من خلال الاعتماد على اشجار انحدار عديده في وقت واحد وأنموذج واحد واختيار افضل القرارات التي تعطيها غابه الاشجار في أنموذج الغابة العشوائية.

المصادر References

1. AL-Badrani, Thafer & Slewa ,Rehad .(2014)." Evaluation of time series prediction of temperature rates using neural networks",Iraqi Journal of Statistical Science, Issue 14,N.26,PP.1-19.
2. Vandel ,Walter, (1992)." Applied time series and Box-Jenkins models",Rayed , Arabic Sudia King.
3. Barker, N. D. (1998). *Basic concepts of statistics* (Vol. 30). Oxford University Press, NY, USA.
4. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
5. Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control (rev. ed.) Holden-Day. San Francisco, 575.
6. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
7. Chan, N. H. (2004). Time series: applications to finance. John Wiley & Sons.
8. Chen, L., Omaye, S. T., Yang, W., Jennison, B. L., & Goodrich, A. (2001). A comparison of two statistical models for analyzing the association between PM10 and hospital admissions for chronic obstructive pulmonary disease. *Toxicology Methods*, *11*(4), 233-246.
9. Chen, J., Li, M., & Wang, W. (2012). Statistical uncertainty estimation using random forests and its application to drought forecast. Mathematical Problems in Engineering, 2012.
10. Cryer, J. D., & Chan, K. S. (2008). Time series analysis: with applications in R (Vol. 2). New York: Springer.
11. Díaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., & Moncada-Herrera, J. A. (2008). A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*, *42*(35), 8331-8340.
12. Fang, X., Liu, W., Ai, J., He, M., Wu, Y., Shi, Y., & Bao, C. (2020). Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. BMC infectious diseases, 20(1), 1-8.
13. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.
14. Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC bioinformatics, 15(1), 1-9.
15. Kitagawa, G. (2010). Introduction to time series modeling. Chapman and Hall/CRC.
16. Liu, L. M. (2006). *Time Series Analysis and Forecasting*. 2^{nd} ed. Scientific computing associates crop. Illinois, USA.
17. Noureen, S., Atique, S., Roy, V., & Bayne, S. (2019). A comparative forecasting analysis of ARIMA model vs random forest algorithm for a case study of small-scale industrial load. International Research Journal of Engineering and Technology, 6(09), 1812-1821.
18. Palma, W. (2007). Long-memory time series: theory and methods. John Wiley & Sons.

19. Pankratz, A. (1983). Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. John Wily & Sons. Inc. USA.
20. Petukhova, T., Ojkic, D., McEwen, B., Deardon, R., & Poljak, Z. (2018). Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. PloS one, 13(6), e0198313.
21. Shukur, O. B. (2015). Artifical Neural Network and Kalman Filter Approaches Based on ARIMA for Daily Wind Speed Forecasting (Doctoral dissertation, Universiti Teknologi Malaysia).
22. Shumway, R. H., and Stoffer, D. S. (2000). Time series analysis and its applications (Vol. 3). New York: springer.
23. Shukur, O. B., & Lee, M. H. (2015). Daily wind speed forecasting through hybrid KF-ANN model based on ARIMA. Renewable Energy, 76, 637-647.
24. Wei, W. W. S. (1990). Time series analysis: Univariate and multivariate methods. 478 pp. New York, Adisson–Wesley.
25. Wei, W. W. (2006). Time series analysis: univariate and multivariate. Methods. Boston, MA: Pearson Addison Wesley.
26. Zafra, C., Ángel, Y., & Torres, E. (2017). ARIMA analysis of the effect of land surface coverage on PM10 concentrations in a high-altitude megacity. *Atmospheric Pollution*

Using ARIMA and Random Forest Models for Climatic Datasets Forecasting

Oday Aljuborey; Osamah Basheer Shukur

Department of Informatics & Statistic, College of Computer & Mathematical Science, University of Mosul, Mosul, Iraq

Abstract

The damages through planning and controlling for these changes in the future. The main problem can be summarized in the nonlinearity of climatic dataset and its chaotic changes. The common approach is the integrated autoregressive and moving average model (ARIMA) as traditional univariate time series approach. Therefore, more appropriate model for studying the climatic data has been proposed for obtaining more accurate forecasting, it can be called random forest (RF) model. This model cannot deal with nonlinear data correctly and that may lead to inaccurate forecasting results . In this thesis, climatic datasets are studied represented by minimum air temperature and rational humidity for agricultural meteorological station in Nineveh. This thesis aims to satisfy data homogeneity through different seasons and find suitable model deal with nonlinear data correctly with minimal forecasting error comparing to ARIMA as traditional model. The research found the adequate of the model for this type of data, as it was found that there are some factors that contribute to the increase in the number of deaths in the epidemic, such as the advanced age of the patient, the length of stay in the hospital, the percentage of oxygen in the patient's blood, in addition to the incidence of some chronic diseases such as asthma. The study recommended a more in-depth study of other types of these models, and the use of other estimation methods, in addition to paying attention to the methods of data recording by the city health department.

Keywords: hierarchical Poisson regression model with random intercept, full maximum likelihood method, intraclass correlation coefficient, fixed and random effects