المجلة العراقية للعلوم الاحصائية (18) 2010
ص ص [195-208]

التحليل المميز واستعمالاته طريقة للتصنيف

خولة مصطفى صادق* مي مصطفى صادق**

الملخص

قمنا ولفترة ثلاث سنوات بدراسة حجم معين من الضباب والغازات المختلفة التي تم رصدها في وادي وانة في السليمانية .

كذلك حاولت الدراسة تحديد العوامل المؤثرة في عملية تصنيف البيانات والعوامل التي تجعل الحادثة تنمو نموا اكثر أي تزداد عددا واختبار هذا التأثير باستعمال طرائق احصائية متعددة .تم استخدام التحليل المميز غير المعلمي كطريقة تحليلية متضمنا استخدام الطريقة اللبية في طريقة الكثافة المقدرة غير المعلمية . ولوحظ ان احسن نتيجة للتصنيف كانت بدمج عامل الرطوبة النسبية وعامل نسبة غاز الاوزون مع عامل الاشعاعات المنبثقة من التربة .

كما ان عامل الرطوبة النسبية RH ظهر ليكون مؤثراً معيقاً للجسيم المتكون اذ ان تاثيرات الاوزون والاشعاعات مجتمعة كانت موصلة اكثر .

وقد تبين لنا بان تركيز SO_2 ، NO_2 يعطينا مؤثراً معنوياً لظهور نواة الحدث ولكن بسبب كثرة الملاحظات الضائعة سوف نحاول اشتقاقها.

Discriminate analysis and it's application as a classification method

Abstract

More than three years of measurements of aerosol size-distribution and different gas and meteorological parameters made in a specific region were analyzed for this study to examine which of the meteorological and trace gas variables effect on the emergence of nucleation events. As an analysis method, we used discriminate analysis with non-parametric density estimation method. The best classification results in our data was reached with the combination of relative humidity ozone concentration and a third degree polynomial of radiation. RH appeared to have a preventing effect on the new particle formation whereas the effects of O_3 and radiation were more conductive. The concentration of SO_2 and NO_2 also appeared to have significant effect on the emergence of nucleation events but because of the

* كلية التربية / قسم الرياضيات /جامعة الموصل
** مدرس/ هيئة التعليم التقني – بغداد
تاريخ التسلم :2009/ 10/18 ---------------- تاريخ القبول : 2010/4/21

great amount of missing observation, we had to exclude them from the final analysis

المقدمة :

تم التنسيق مع مديرية الارصاد الجوية على اخذ البيانات التي سيرد ذكرها لاحقا من وادي وانه في محافظة السليمانية وبصورة يومية علـى امتـداد ثـلاث سنوات وذلك اعتبارا من 2000/2/24 ولغاية 2003/2/24 اذ ان عملنا كان على مدار (814) يوماً خلال هذه الفترة الزمنية والتي تضمنت (293) من ايام احـداث و(270) من ايام ليس بها احداث يضاف الـى ذلـك ان (251) يومـاً لا يمكـن تصنيفها .

يعد اليوم حدثا اذا دل على تكوينات جديدة أي هناك تغيرات جويـة تـشير الى نمو او خفوت نسبة الاوزون او الرطوبة النسبية او الاشعاع أي ايام يحدث فيها تغير.

اما اذا لم يلاحظ أي تغير فسوف نطلق على هذه الايام انها ليس بها احداث ومن الملاحظ وجود عدد كبير من الايام ليس بالامكان تميزها هل انها بها احـداث ام غير ذلك للتشابه الكبير فيما بينها . وقد شـملت دراسـتنا تـاثير الاشـعاعات والاوزون O_3 والرطوبة النسبية واخيرا NO_2 . فبعد اختبار التـأثيرات المعنويـة تبين ان تاثير الرطوبة النسبية اولا ثم الاشعاعات واخيـرا نـسبة الاوزون تـاتي بالتناوب ، ولاحظنا ان اكبر تاثير على الجسيم المتكون الجديد (الضباب) كان مـن الرطوبة النسبية ومن ثم تاثير الاوزون O_3 والاشعاعات التي لم تكن بهذا التـاثير المعيق أي انها موصلة اكثر من الرطوبة النسبية ، كذلك ظهر لدينا ان تركيز SO_2 و NO_2 في الجو له تاثير معنوي كبير متعلق بانبثاق او نشوء أساسـي للـضباب وبمرافقة أي تغير ملحوظ ومن المؤسف ان غالبية او العدد الاكبر من البيانات التي تم تسجيلها كانت غير مستوفاة للشروط أي خالية من بعض الملاحظات الضرورية

واضطررنا الى اعتبارها ملاحظات ضائعة ومن ثم تم حذف اعداد كبيرة جدا منها من التحاليل النهائية .

التحليل المميز : Discriminate analysis

عند دراسة تصنيف الاعداد نلاحظ وجود عدة طرائق تعتبر مقبولة تصنف وتشرح لنا المتغيرات العددية ، وفي هذا البحث تم استخدام التحليل المميـز (DA) الخالي من المعالم والذي يطلق عليه التحليل اللبي [5] لايجاد العوامل التي تـصنف الايام كنواة للاحداث المصاحبة او كأيام لا توجد فيها أية حوادث او أية متغيرات . وتم تفسير الحالة الحسابية والرياضية لهذه الايام اعتماداً على برنـامج الحاسـوب الجاهز (SAS) وتم استخدام طريقتين مختلفتين لاختبار كفـاءة الجـودة للنمـاذج المستخدمة [3] :

1. اعادة التعويض : حيث ان النموذج المحسوب له نفـس المجموعـة مـن المعلومات سواء تم تقديره من أي مصدر.
2. تحليل العبور Cross – Validation : حيـث ان اي نمـوذج سـيطابق مجاميع بيانات مختلفة افضل من مجموعة واحدة مستخدمة في التقدير .

ان التحليل المميز هو طريقة احصائية لمتعدد المتغيرات والتـي تـستعمل لبناء نموذج وصفي او مجاميع مميزة تعتمد علـى مجموعـة متغيـرات تنبؤيـة ولوصف أية ملاحظات في المجاميع اذا كان التوزيع داخل كل مجموعة هو توزيع طبيعي متعدد ، يمكن في هذه الحالة استخدام طريقة معلمية (خطيـة او تكعيبيـة) لانشاء دالة مميزة .

تستخدم طرائق التحليل المميز غير المعلمي عند عدم وجـود فرضـية ان المتغيرات الداخلة في التحليل المميز تتبع التوزيع الطبيعي وهي فرضـية رئيـسة للتحليل المميز.

ان الطريقة غير المعلمية تعتمد على مجموعات وصفية ذوات دالة احتمالية معينة والتي تستخدم لانتاج مقاييس معيارية تعتمد على هذه الاحتمـالات ، وفيمـا يتعلق بهذا البحث فان التحليل مكون من بيانات مقاسـة وان الفرضـية الطبيعيـة (التوزيع الطبيعي) غير موجودة لذلك سيكون ثمة حاجة الى استخدام طريقة اللبـي غير المعلمية "Non-parametric Kernel method" اذ انها ستعتبر قوية فـي الدوال او البيانات متعددة الخطية والتي غالبا ما توجد في هكذا نوع من التحاليل .

التقدير اللبي : Kernel estimation [3][6]

ان الفرض الاساسي في التقدير اللبي هـو تقـدير دالـة الكثافـة لجميـع الملاحظات التي تم الحصول عليها دون الحاجة الى فرضية وجود أي نـوع مـن التوزيع. حيث ان تقارب المشاهدات من حيث القيمة او الكمية يستدعي التعامل مع التقدير اللبي والذي ينص على ان مربع المسافة بين أيـة ملاحظتـين متمـاثلتين بالمتجه x والمتجه y في المجموعة t سيعطى بـالعلاقة الاتية :

$$d_t^2(x,y)=(x-y)'V_t^{-1}(x-y)$$

V_t ستكون مصفوفة التغاير التي تحتوي على حوادث مصنفة في المجموعة t .

ان التصانيف [9] التي تعود الى متجه الملاحظات المتعددة x تعتمد على كثافة المجموعة المقدرة التي تم الحصول عليها من البيانات . ومن هذه الكثافـات المقدرة سيكون الاحتمال الاولى من عضوية المجموعة في x التي قد تم اشتقاقها .

نشاهد ان قيمة x تم تصنيفها في المجموعة u اذا كانت $t=u$ والتي تنتج اعلى قيمة من الدالة الاحتمالية الشرطية $P(t/x)$.

ان طريقة النواة [1][2] تستعمل اشعاعات ثابتة r وقيمة محصلة kt لكي يتم تقدير الكثافة t في كل ملاحظة من المتجه x . والان لندع z يكون P من المتجهات "P-dimensional vector" اذن سيكون مقدار الـ (P-dimensional) وحدة والتي تكون ضمن القيد $z'z=1$ هي [10] :

$$V_0 = \frac{\pi^{\frac{P}{2}}}{\Gamma\left(\frac{P}{2}+1\right)}$$

Γ : تمثل دالة كاما

وكذلك في المجموعة t قيمة القيد الناقص او المقطوع للـ P-dimensional لـ

$$\left\{Z/Z' \quad V_t^{-1} \quad Z = r^2\right\}$$

متمثلة بما ياتي [5] :

$$V_r(t) = r^P |Vt|^{\frac{1}{2}} V_0$$

وتم تعريف عدة دوال مقبولة خاصة بالكثافة الاحتمالية لـ Kernel فضلاً عن ذلك تم اختيار كل واحدة لمعرفة الاختلافات العددية الاحصائية لكل منها والتي هي كالاتي [8]:

$$k_t(z) = \begin{cases} c_1(t)\left(1 - \frac{1}{r^2} z'V_t^{-1}z\right) if \ z'V_t^{-1}z \le r^2 \\ 0 \qquad\qquad\qquad\qquad else \ where \end{cases}$$

حيث

$$c_1(t)=\frac{1}{V_r(t)}\left(1+\frac{P}{2}\right)$$

المجموعة t التي لها كثافة احتمالية في x تقدر بـ

$$f_t(x)=\frac{1}{\eta_t}\sum_y k_t(x-y)$$

حيث ان η_t هي عدد الملاحظات في المجموعة t .

اما اشارة الجمع فهي تخضع لها كل ملاحظات y في المجموعة t , k_t هي دالة Kernel التي تم الحصول عليها . الاحتمال اللاحق لجمع البيانات في المجموعة t سيعطى بـ [6][7]:

$$P(t/x)=\frac{q_t f_t(x)}{f(x)}$$

حيث ان

$$f(x)=\sum_u q_u f_u(x)$$

هو الكثافة الاحتمالية غير الشرطية المقدرة وان q_t هو الاحتمال السابق للمجموعة t .

الجانب العملي :

هناك عدة عوامل تؤثر في كمية او كثافة الضباب الذي تمت دراسته في منطقة (وانه) في السليمانية حسب البيانات التي تم الحصول عليها من الارصاد الجوية وهي :

1. تركيز الاوزون O_3 .
2. الاشعاعات وتشمل أية اشعاعات سواء في الشمس او بتأثير الآلات من المعامل القريبة .

3. الرطوبة النسبية .

4. تركيز SO_3 , NO_2 في الجو .

وواجهت الباحثتان صعوبة بالغة في الحفاظ على جميع هذه الارقام في المشاهدة الواحدة ولسلسلة طويلة مؤلفة من ثلاث سنوات مما حدا بهما الى حذف كثير من المشاهدات غير المكتملة المعلومات او (غير الحادثة) لجميع المتغيرات وتم استخدام طريقة اعادة التعويض لمعرفة التأثير المعنوي للمشاهدات الثلاثة فقد تم في المرحلة الاولى ترك الايام غير المصنفة خارج التحليل .

وحاولنا التحليل باستخدام الايام الحادثة وغير الحادثة.

لقد وضع العالم Hyvonenet al. (2005) نموذجين متغيرين للبيانات التي تم اخذها من العالمين Hyytiala and Finland بنموذجهما الخاص بالتكرار النسبي اذ شرح كثافة الانحدار 88% من الايام الحادثة بخطأ معياري كلي مقداره 12% ولكن بالنسبة لبياناتنا هذا النموذج كان بسيطا جدا اذ انه لا يزال يمثل 88% من الحوادث في اعادة التعويض ولكنه يعطينا عددا كبيرا من الحوادث الكاذبة (false events) (أي تتنبأ بيوم غير حادث ليكون يوما حادثا) والذي سيزيد الخطأ التصنيفي الكلي الى 22% . كما في الجدول رقم (1) .

كذلك نشاهد ان ذاك كان احسن موديلين متغيرين تم ايجادهما . اذ لابد لنا ان نزيد عدد التنبؤات لكي نحصل على تصنيف مقبول . وبالنسبة الى موديلنا نستطيع ان نشرح تقريبا 99% من الايام الحادثة في اعادة التعويض بالمعلومات المستنقاة من ثلاثة متجهات مختلفة النسبية اذ كانت الاشعاعات وتركيز الاوزون افضل متغير معنوي.

ان تأثير الاشعاعات تم تقديره بثلاث درجات معيارية ، تركيز NO_2 , SO_2 كان ملائما ايضا للتنبؤ للنموذج ولكن بسبب العدد الكبير من الملاحظات الضائعة كان الخطأ التصنيفي الكلي للنموذج 3.36% واضحاً.

ان عملية اعادة التعويض هي طريقة جيدة لايجاد احسن نمـوذج للبيانـات الجارية امامنا ولكن اذا كان أي واحد يحتاج الى معرفة كيف ان النمـوذج يتكـون بمجموعتين مختلفتين من البيانات هناك طريقة يتم استعمالها وهي ما تعـرف بــ (Cross - validation) . وفي هذه الطريقـة سـوف يـتم انـشاء 1000 مـن المجموعات التي تم التدريب عليها أي المجموعات الكاملـة التـي تـم اسـتعمال التحويلات الرياضية عليها و 1000 مجموعة من البيانات الاصـلية بمـا يعـرف بطريقة Bootstrap أي طريقة اعادة المعاينة [Efron and Tibshirani, 1993]

نحن نحسب كلا النموذجين لكل المجاميع التي نطلق عليهـا المجموعـات المتدربة ونحاول التنبؤ بتوزيع الاحداث لمجاميع الاختبارات مع النتائج . نحـسب المتوسط ، الانحراف المعياري و95% من حدود الثقة من الخطأ المصنف النهائي ونسب عدم التصنيف من الـ 1000 تقدير المكررة . النموذجان المتغيران ينعـدم التصنيف بينهما بمعدل مقداره 13.7% في الحدث اليومي ولكن ايضا سـوف يـتم التنبؤ باعداد كبيرة من الحوادث الكاذبة والتي سوف تزيد الخطأ التصنيفي النهـائي الى 23.2% (الجدول رقم 2) .

لو اخذنا النموذج من ناحية الاشعاعات RH والاوزون O_3 متضمنا ضياع عدد من الحوادث التي ينتجها هذان المتغيران سنصل الى عدد قليل من الحـوادث الكاذبة . ان متوسط الخطأ التصنيفي النهائي لـ 1000 مكرر هو 17.6% والـذي هو بصوره واضحة اعلى من نموذج المتغيرين وكذلك نستطيع القـول ان نمـوذج المتغيرات الثلاث التنبؤي يتطابق اكثر مـع بيانـات مـا تعـرف بــ (Cross-validation) .

ان السلبية الرئيسة في هذا النموذج هي انه لم نأخذ في الحسبان الايام التي لم يتم تصنيفها في البيانات . اذ ان القيود التي من الواجب اتباعها سوف تقودنا الى ان نخسر غالبا واحد من ثلاثة من البيانات التي تم جمعهـا. اذن سـوف نخطـط لاستعمال معلومات التصنيف في تحليلات اضافية وكذلك نحتاج الى الاخـذ بنظـر

الاعتبار الايام غير المصنفة . وهذا يتم باستعمال التحليل المميز لثلاثة – تصنيف الحدث – المتغير ، حيث ان التصانيف هي حوادث ، حوادث كاذبة وعدم التصنيف، بدلا من متغيرين مصنفين ، حوادث وايام ذوات حوادث كاذبة . اذ ان أي صنف من العناصر غير مصنف سابقا ليس بالضبط مستقلا عن اصناف الحوادث وانعدامها والتحليل المميز سيكون ذا خاصية ضعيفة وغير مناسب والتصنيف هو ليس جيد كما في البيانات المقيدة الخطأ التصنيفي النهائي لـ [التنبؤات الثلاثة وثلاثة نماذج المصنفة] هو 13.5% ونسبة عدم التصنيف للايام ذوات الحوادث كانت 1.9% [جدول رقم 3] .

ان للنموذج بعض الصعوبات في نشر الايام غير الحادثة من الايام غير المصنفة وهذا يحدث غالبا بسبب انه اعداد كبيرة من الايام غير المصنفة هي غالبا تشابه الايام غير الحادثة.

ان انشاء النموذج بالمتغيرين وبثلاثة تصانيف هو ايضا أسوأ من الذي بتصنيفين . اذ ان الخطأ التصنيفي النهائي هو 43.2% حتى لو يتم المتنبأ بصورة صحيحة بـ 83.3% من الايام الحادثة .

ان النموذجين لا يستطيعان تمييز الايام غير المصنفة من الايام الحادثة والايام غير الحادثة في بياناتنا وكما معلوم سيكون نحو نصف من الايام غير المصنفة في الايام الحادثة. وهي كذلك تصنف 10.8% من الايام الحادثة في الايام غير الحادثة 31.2% من الايام غير الحادثة في الايام الحادثة .

Cross-validation لبيانات التصانيف الثلاثة في الجدول (4) ترينا اضافة الايام غير المصنفة للتحليل والتي تجعل التحليل غير مناسب ويؤدي الى زيادة الخطأ التصنيفي . ان نموذج التنبؤات الثلاثة قد فقد مرة اخرى كثيرا من الحوادث ولكن غالبيتها هي حوادث ضائعة صنفت في المجموعة غير المصنفة . بينما النموذج التنبؤي للمتغيرين جعل كثيرا من الحوادث الضائعة في الحوادث غير

الضائعة فضلا عن ذلك فان النموذجين المتغيرين ينتجان عددا كبيرا من الحـوادث الكاذبة .

كما انه من الممكن لدعم او تقوية النموذج وذلك ليقوم باعادة تصنيف كـل الايام غير المصنفة في احدى المجموعتين من الايام اما حادثة او غير حادثة ، كما ان بيانات الايام غير المصنفة هي فقط نتيجة لطريقة التصنيف الكفوءة او المثالية ، وكنتيجة لهذه الفرضية نستطيع ان نعمل اعادة تعويض حيث ان الايام الحادثة وغير الحادثة تصنف كما تعمل فيما لو كان هناك تصنيفان في التحليل وان 27.86% من الايام غير المصنفة تم تصنيفها كأيام حادثة كما في الجدول (5) .

ونلاحظ ان النموذج قد فشل في عملية اعادة التعويض لـ 23NC مـن الايام وهذا يعزى الى تأثير لمجموعة الاحتمالية التصنيفية الكبيرة . كما ان فـشل اعادة التعويض ممكن تجنبه باستبدال الـ width لـ kernel ، ولكـن للمقارنـة سوف نحتاج الى استخدام نفس الـ kernel في كل نموذج كما رأيناها في الجدول (3) حيث ان نموذج المتغيرين فرض اغلبية الايام غير المصنفة (54.29%) هـي ايام حادثة .

ان اعادة التعويض تعطينا تقديرا لعدد الايام غير المصنفة والتي من الممكن تصنيفها كأيام حادثة وغير حادثة ، ومن الشائع دائما ان اغلبية الايام غير المصنفة هي ايام غيرحادثة ، كما اقترح النموذج ذو التنبؤات الثلاثة . وقد قادنـا النمـوذج ذات المتغيرين الى تغطية جميع التقدير لعدد الايام الحادثة وتم تصنيف اكثـر مـن نصف الايام غير المصنفة الى ايام حادثة .

الاستنتاجات :

قمنا تحليل البيانات الى حصلنا عليها من وادي يقع بالقرب من وانه فـي محافظة السليمانية للفترة من 24 شباط 2000 لغاية 2003/3/14 وقد كانت النتائج

التي حصلنا عليها ترينا ان مناطق التلوث مثل وادي وانه يمتاز بعدة عمليات معقدة تسيطر على نشوء او انبثاق الحدث اكثر بكثير من المناطق النظيفة . حيث وجـدنا ان الرطوبة النسبية الكبيرة تعطينا تأثيرا يحدث من معالم جديدة ، حيث ان الاشعاع الكبير والتركيز الكبير للاوزون له تأثير ايجابي في احتمالية الحدث الذري وتساعد الوسائل المؤثرة لكي تنمو الى احجام معروفة. وان طريقة التحليل المميز مع دالـة كيرنل متضمنة طريقة تقدير الكثافة غير المعلمية . ولاختبـار أي مـن متغيـري المؤثرات الجوية وآثار الغازات لها تأثير في نشؤ الحدث الذري (nucleation) . ان احسن نتيجة تصنيف في بياناتنا تم التوصل اليها بضم RH , O_3 وثلاث درجات من الاشعاع .

ان تركيز SO_2 , NO_2 قد ظهرت ايضا كتأثيرات معنوية في ظهور نـواة الحدث ولكن بسبب الكمية الكبيرة من الملاحظات الضائعة ، سوف يتم اشتقاقها من التحليل . وفي بعض الاحيان يكون كـلا مـن الاشـعاعات والاوزون متفـاعلان وينتميان الى مجموعة ذات ثلاث متغيرات والتي تعطينا احسن شرح احصائي لاي نواة حدث واقعة ، اذ ان الاوزون يتولد كتفاعل كيمـائي كـذلك تركيـز الاوزون الواضح يعتمد على بقية العوامل اكثر من الاشعاع وعند ذلـك تعطينـا معلومـات اضافية للنموذج الاحصائي والذي هو مشابه بالعائدية لتفاعل الاكسدة العضوية الذي يؤول الى تصنيفات على نمو البيانات .

الجدول (1)

إعادة التعويض لنموذجين مختلفين

الحوادث الفاشلة	الحوادث الضائعة	الخطأ التصنيفي	المتنبأ به
اعادة التعويض			
5.67%	1.27%	3.36%	RH, O_3, Radiation
34.7%	12.1%	22.82%	RH, log(cs)

الجدول (2)

الخطأ التصنيفي الكلي في Cross-Validation ونسبة عدم التصنيف للنماذج

التنبؤات	Cross-Validation 1000 Simulations				
		Mean	**Std Dev**	**Lower 95% Cl for mean**	**Upper 95% Cl for mean**
	Total error	17.63%	0.0277	17.46	17.80
RH, O_3, Radiation	missed events	15.24%	0.0436	14.97	15.52
	false events	20.29%	0.0531	19.96	20.62
	Total error	23.21%	0.0249	23.05	23.36
	missed events	13.67%	0.0395	13.42	13.91
RH, log(cs)	false events	33.83%	0.0495	33.52	34.14

الجدول (3)

اعادة التعويض للنماذج لبيانات ذوات ثلاثة تصانيف

Resubstitution table for the models for three-class data

المتنبأ به	الخطأ التصنيفي	اعادة التعويض لثلاثة تصانيف في البيانات			
		الحوادث الضائعة	غير حادثة الى حادثة	**NC to event**	**NC to nonevent**
RH, O_3, Radiation	13.47%	1.91%	4.96%	7.14%	2.86%

RH, log(cs)	43.15%	16.56%	31.21%	64.43%	29.29%

الجدول (4)

Total classification errors in cross-validation and misclassification rates for the models for the three-class data

التنبؤات	Cross-Validation 1000 Simulations, three-class data				
		Mean	**Std Dev**	**Lower 95% Cl for mean**	**Upper 95% Cl for mean**
RH, O_3, Radiation	Total error	31.78%	0.0276	31.61	31.95
	missed events	24.56%	0.0473	24.26	24.85
	event to nonevent	6.61%	0.0269	6.44	6.78
	nonevent to event	7.17%	0.0293	6.99	7.35
RH, log(cs)	Total error	45.08%	0.0242	44.93	45.23
	missed events	19.83%	0.0506	19.51	20.14
	event to nonevent	11.88%	0.0391	11.64	12.12
	nonevent to event	29.16%	0.0501	28.58	29.48

الجدول (5)

اعادة التعويض للنموذج (1) لثلاثة تصانيف من البيانات اذ ان كل الايام صنفت على انها أيام حادثة او غير حادثة

المتنبأ به	missed events	اعادة التعويض لثلاثة تصانيف في البيانات			
		nonevent to event	**NC to event**	**NC to nonevent**	**NC classification failed**

RH, O_3, Radiation	1.27%	5.67%	27.86%	55.71%	16.43%
RH, log(cs)	12.1%	33.33%	54.29%	44.29%	1.43%

References :

1. Birmili, W.; Berresheim, H., Plass-Dulmer, C., Elste, T., Gilge, S.; Wiedensohler, A., and Uhrner, U. The Hohenpeissenberg aerosol formation experiment (HAFEX): a long-term stubby study includeing size-resolved aerosol, H2So4, OH, and monoterpenes measurements, Atmos, Chem. Phys, 3, 361-376, 2003.
2. Boy, M, and Kulmala, M: Nucleation events in the continental boundary layer: influence of physical and meteprplogical parameters, Atmos, Chem. Phys. 2, 1-16, 2002.
3. Hyvonen, S, Juninen, H., Laakso, L., Dal Maso, M., Gronholm, T, Bonn, B., Keronen, P., Aalto, P., Hiltunen, V., pohja, T., Kauniainen, S., Han, P., Mannila, H., and Kulmala, M., A look at acrosol formation using data mining techniques, Atmos, Chem. Physa., 5, 3345-3356, 2005.
4. Etron, B and Tibshirani, R: An introducition of the Bootstrap, London Chapman and Hall, 1993.
5. Eqanechnikov, V. A.: Nonparametric estimation of a multimensional probability denaity, The-ory probab. Appl., 14, 153-158, 1969.
6. Kulmala, M., Kerminen, V. M., Anttila, T., Laaksonen, A., and ODowd, C.: Organic aerosol formation via sulphate cluster activation, J. Geophys. Res., 109, D04205, doi: 10, 1029/2003JD003961, 2004a.
7. Kulmala, M., Vehkamaki, H., petaja, T., Dal Maso, M., Lauri, A., Kerminen, V.M. Birmili, W., and McMurry, P.: formation and growth rates of ultrafine atmospheric particles: a review of observations, J. Aerosol Sci., 35, 143-176, doi: 10. 1016/j.jaerosci. 2003. 10. 003, 2004b.
8. Morrison, D. F: Multivariate Statistical Methods, New York, McGraw-Hill, 1976.
9. ODowd, C. D., Aalto, P., Hameri, K., Kulmala, M., and Hoffman, T.: Atmospheric particles from ofganic vapours, Nature, 416, 497-498, doi: 10. 1038/416497a, 2002.

10. Rodrinquez, S., van Dingenen R., putaud, J-P., Martins-Dos Santos, S., Roselli D.: Nucleation and growth of new particles in the rual atmosphere of Northern ltaly-relationship to air qual-lty monitoring, Atmos Environ., 39(36), 6734-6746, doi: 10. 1016/j. atmosenv. 2005. 07. 036, 2005.
11. SAS lnstitute inc.: SAS/STAT Users Guide, Version 8, SAS Publishing, 1999. Woo K. S., Chen D. R., Pui D. Y. H., and McMurry p, H.: Measurement of Atlanta aerosol size distnbutions: Observations of ultrafine particle events, Aerosol Sci, Technol., 34, 75-87, doi: 10. 1080/02786820120056, 2001.