ESAIM: PS
September 2005, Vol. 9, p. 230–240
DOI: 10.1051/ps:2005013

ESAIM: Probability and Statistics

DETECTING ATYPICAL DATA IN AIR POLLUTION STUDIES BY USING SHORTH INTERVALS FOR REGRESSION

Cécile Durot[1] and Karelle Thiébot[1, 2]

Abstract. To validate pollution data, subject-matter experts in Airpl (an organization that maintains a network of air pollution monitoring stations in western France) daily perform visual examinations of the data and check their consistency. In this paper, we describe these visual examinations and propose a formalization for this problem. The examinations consist in comparisons of so-called shorth intervals so we build a statistical test that compares such intervals in a nonparametric regression model. This allows to detect atypical data. A practical application of the test is given.

Mathematics Subject Classification. 62G08, 62G09, 62G10, 62P12.

Received December 15, 2003. Revised April 8, 2005.

1. Introduction

This paper and its companion (Durot and Thiébot [2]) are motivated by a problem raised by "Air Pays de la Loire" (Airpl in the sequel), an organization that maintains a network of air pollution monitoring stations in western France.

It is now well known that there exist relationships between air pollution, human health and environmental matters so study of pollution is of a great importance. That is the reason why studies of air pollution are daily carried out by Airpl. The main task of Airpl is to detect when air pollution achieves a critical level in the Loire valley and if necessary, to inform people and industries. If air pollution exceeds a critical level then for instance, procedures for limiting polluting waste have to be set up in polluting industries, and it is recommended to avoid the use of personal cars.

Pollution studies performed by Airpl are based on measures of the concentration in the ambient air of several pollutants, and the first problem with these pollution data concerns their validity. In some cases indeed, collected data are invalid. This may happen for instance following the failure of monitoring device. Beyond failure problems, it may also happen that pollution data are well measured by monitoring device but are not representative of the area where they are collected: this happens for instance if a car which motor is in operation is parked for a long time just near a pollution sensor in a urban site. Validity of pollution data thus have to be checked before the data are used to perform air pollution studies. Until now, validation of data was daily performed by subject-matter experts, mainly through visual examination of the data. Because this job took a long time every day, Airpl wished to implement statistical tests that would aid experts to validate pollution

Keywords and phrases. Air pollution, validation, regression, bootstrap, shorth.

[1] Université Paris Sud, Bâtiment 425, 91405 Orsay Cedex, France; cecile.durot@math.u-psud.fr

[2] *Supported by Air Pays De La Loire*; Air Pays de la Loire, 2 rue A. Kastler, BP 30723, 44307 Nantes Cedex 3, France.

data. It is the goal of this paper and its companion to build such statistical tools. To be more specific, a nonparametric statistical model is investigated in the companion paper and is applied in the present paper to the problem of validating pollution data.

The paper is organized as follows. In Section 2, the problem of validation of pollution data is more precisely described and our statistical goal is defined. Statistical model and testing procedure are given in Sections 3 and 4 respectively. Some simulations are reported in Section 5 and a practical application of the method is given in Section 6. We conclude the paper with remarks on possible generalizations of the method in Section 7.

2. Validation of pollution data

To understand the problem of validation, a description of the network maintained by Airpl is informative. This is given in the following paragraph. We then describe experts validation and we define our statistical goal.

Airpl network

Airpl network is composed of eight subnetworks spread out over the Loire valley, each subnetwork being in turn composed of several monitoring sites. At each site, monitors continuously measure the level (in micrograms per cubic metre) of a given pollutant in ambient air. Pollutant which level is measured could be for instance ozone, nitrogen monoxide or carbon dioxide (twelve pollutants are studied by Airpl in total). Between four to twenty measures are collected by a given monitor per minute, depending on the considered pollutant. Each fifteen minutes, the monitor computes the mean of the measures collected during the preceding fifteen minutes and transmits this mean to a database. Only these means are saved so, in the sequel, we refer to them as concentration measures. At the end of a day d, the database thus contains sets of observations, each of them consisting of the ninety-six concentration measures computed that day for a given pollutant p at a given site s. Such a set of concentration measures is called a profile and is denoted by (p, d, s) in the sequel. In most cases, sites are not equipped to measure all of the twelve pollutants and are only equipped to measure some of them. However, numerous profiles (about 125) are collected per day by Airpl, see Figures 3 and 4 for examples of such profiles. These data are subjected to validation as described below and studied in order for instance to detect oversteppings of an alarm threshold. Moreover, when validated, concentration measures are averaged in order to obtain hourly average concentrations that are spread to people, see `http://www.airpl.org`.

Experts validation

Validation of pollution data is an intricate task. On the one hand, monitoring devices are often checked so that possible failure is detected. On the other hand, examination of the data is carried out by subject-matter experts in order to check consistency of the data. Concentration measures are first considered individually. At this stage, a concentration measure could be invalidated for instance if it exceeds a given threshold or if it is significantly greater than the other concentration measures in the profile. Profiles are then considered as a whole and visually compared with other profiles to check consistency of the measures. Other examinations that take into account medium and long term trends are carried out. Many kinds of examinations are daily performed by experts of Airpl, and it is not the aim of this paper to explain all of them: we refer to Thiébot [4] for more details. In this paper, we are only interested in one kind of examination that we explain now.

Profiles of some pollutants (such as nitrogen or carbon monoxides) are characterized by a peak that corresponds to the period of the day when pollution is maximal. We are concerned here with comparisons of such profiles. More precisely, we are interested in a criterion for validation that is based on the fact that the peak of a given profile (p, d, s) should occur at the same period as peaks of other profiles do. For instance, peak of (p, d, s) should occur at the same period as that of (p, d, s'), where s' denotes a neighbouring site to s (spatial consistency). It should also occur at the same period as that of (p, d', s), where d' denotes a well chosen date (temporal consistency). Likewise, if s is a urban site then peaks of (p, d, s) and (p', d, s) should occur at the same period if p and p' denote for instance nitrogen monoxide and nitrogen dioxide respectively (chemical consistency). It is worth noticing that several comparisons are performed to check spatial, temporal and chemical

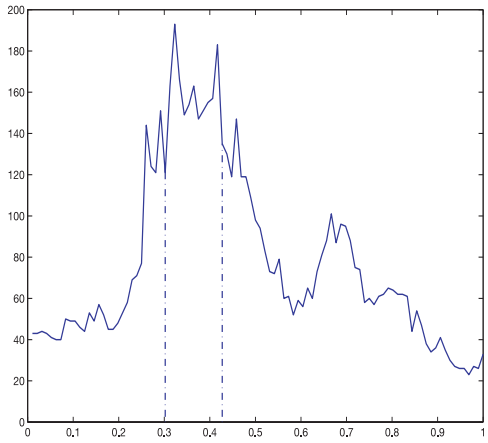

FIGURE 1. Profile of nitrogen dioxide (in micrograms per cubic metre) on December 24th, 1998, in the urban site called Théâtre Graslin (Nantes). An interval $[\mu - r, \mu + r]$ that satisfies (1) with $\eta = 0.25$ is also drawn.

consistency, and that a profile is not rejected as soon as one of the consistency criterion is in fault. For instance, a profile may be temporally inconsistent and chemically consistent if it was collected on a day when the weather was atypical. In that case, the profile is called temporally atypical but is not invalidated by the expert. Beyond comparisons of profiles, experts thus have to decide whether lack of consistency from one or several criterion should invalidate the profile in question.

Let us describe more precisely the way Airpl experts daily carry out comparisons of peaks. Thanks to a software package implemented by Airpl, the profile in question is drawn on a computer screen together with several other profiles to be chosen by the expert. Since observation times are equally spaced, each profile is represented by the points $(t_i, h(t_i))$ and linear interpolation, where for every $i \in \{1, \ldots, 96\}$, $t_i = i/96$ and $h(t_i)$ is the ith concentration measure in the profile, see Figure 1 for an example. Thanks to the software package, the expert plots an interval $[\mu - r, \mu + r]$ such that the surface of the area delimited by the restriction of h to $[\mu - r, \mu + r]$ is a proportion η of the surface of the area delimited by h, that is

$$\int_{\mu-r}^{\mu+r} h(t)\mathrm{d}t = \eta \int h. \tag{1}$$

The most common value considered for η is 0.25 and the interval has to be chosen as short as possible so that it localizes the peak of the profile (an example of such an interval is drawn in Fig. 1). The profile in question is consistent with another profile if the two associated intervals are (approximately) equal.

Statistical goal

Our goal in this paper is to formalize the visual comparisons that are performed by subject-matter experts, that is to build a statistical test for comparing the peaks of two given profiles. This allows to detect profiles that are atypical from a given criterion. Recall however that this does not suffice to validate data: validation of pollution data requires a further study that is out of the scope of this paper, see comment 4 in Section 3.

3. Statistical model and null hypothesis

Consider a profile $\mathcal{P} = (p, d, s)$ consisting of $n = 96$ concentration measures and fix $\eta \in (0, 1)$ (the most common value considered for η is 0.25). Abstractly, we may view concentration measures as random variables, so we denote by $y_i^{\mathcal{P}}$ a random variable that represents the ith concentration measure in profile $\mathcal{P}$. Observation

times are fixed and equally spaced so the ith observation time is (possibly changing scale and origin) $t_i = i/n$. Each concentration measure is obtained as the mean of at least sixty measures on a given period, see Section 2, so we assume $y_1^{\mathcal{P}}, \ldots, y_n^{\mathcal{P}}$ to be independent Gaussian variables. We assume furthermore that they possess the same standard deviation that we denote by $\sigma^{\mathcal{P}}$. The statistical model we consider is then a semiparametric regression model with fixed design:

$$y_i^{\mathcal{P}} = f^{\mathcal{P}}(t_i) + \varepsilon_i^{\mathcal{P}},\ i = 1, \ldots, n,$$

where $t_i = i/n$, $f^{\mathcal{P}}$ is an unknown regression function defined on $[0,1]$ and the $\varepsilon_i^{\mathcal{P}}$'s are independent and identically distributed Gaussian variables with mean zero and (unknown) standard-deviation $\sigma^{\mathcal{P}} > 0$. The model is semiparametric since we assume the $\varepsilon_i^{\mathcal{P}}$'s common distribution to be known up to one parameter but we do not assume that $f^{\mathcal{P}}$ belongs to a given parametric set of functions. We only assume here that $f^{\mathcal{P}}$ is smooth, positive (recall $f^{\mathcal{P}}(t)$ is the expectation of a random variable that represents a concentration measure) and possesses a peak. More specifically, we assume that

(a) $f^{\mathcal{P}}$ is twice differentiable on $[0,1]$ with bounded second derivative,
(b) $\inf_{t\in[0,1]} f^{\mathcal{P}}(t) > 0$,
(c) there exists a unique shortest interval $[\mu^{\mathcal{P}} - r^{\mathcal{P}}, \mu^{\mathcal{P}} + r^{\mathcal{P}}] \subset (0,1)$ that satisfies

$$\int_{\mu^{\mathcal{P}}-r^{\mathcal{P}}}^{\mu^{\mathcal{P}}+r^{\mathcal{P}}} f^{\mathcal{P}}(s)\mathrm{d}s \geqslant \eta \int_0^1 f^{\mathcal{P}}(s)\mathrm{d}s,$$

(d) $f'^{\mathcal{P}}(\mu^{\mathcal{P}} - r^{\mathcal{P}}) > f'^{\mathcal{P}}(\mu^{\mathcal{P}} + r^{\mathcal{P}})$, where $f'^{\mathcal{P}}$ denotes the first derivative of $f^{\mathcal{P}}$.

The so-called η-shorth interval $[\mu^{\mathcal{P}} - r^{\mathcal{P}}, \mu^{\mathcal{P}} + r^{\mathcal{P}}]$ then localizes the peak of the profile $\mathcal{P}$ and our objective is to build statistical tests that compare η-shorth intervals associated to two different profiles $\mathcal{P}_0$ and $\mathcal{P}_1$. We thus wish to build a statistical test for the null hypothesis

$$\mathcal{H}_0 :\ \text{“}\mu^{\mathcal{P}_0} - r^{\mathcal{P}_0} = \mu^{\mathcal{P}_1} - r^{\mathcal{P}_1} \text{ and } \mu^{\mathcal{P}_0} + r^{\mathcal{P}_0} = \mu^{\mathcal{P}_1} + r^{\mathcal{P}_1}\text{”}, \tag{2}$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are two profiles that satisfy the above conditions. The profiles $\mathcal{P}_0$ and $\mathcal{P}_1$ are furthermore assumed independent from each other.

Comments

1. In applications, the η-shorth interval localizes a peak so $f^{\mathcal{P}\prime}(\mu^{\mathcal{P}} - r^{\mathcal{P}}) > 0 > f^{\mathcal{P}\prime}(\mu^{\mathcal{P}} + r^{\mathcal{P}})$ and assumption (d) holds.

2. The η-shorth interval does not depend on the vertical scale choice. This scale invariance is a minimum requirement since in applications, we compare profiles that are not necessarily defined at the same scale. Consider for instance two neighbouring urban sites s and s', where s is located in the centre of the town and s' is not. Then for given p and d, peaks of (p,d,s) and (p,d,s') should occur at the same period, but concentration measures are certainly greater in s than in s'.

3. Regression models are frequently used for pollution data, see *e.g.* Bell *et al.* [1]. In many cases, one wishes to predict pollution or to explain pollution in terms of given factors such as maximum temperature on a given day or force of the wind. In those cases, one has to consider a regression model with covariables. The pollution measure is then modeled by a random error plus an unknown function of these covariables. The main difficulty with this kind of models lies in the choice of the covariables and the choice of a model for the regression mean (one may assume for instance that the regression mean belongs to a given parametric set of functions). On the contrary, our aim in this paper is to describe profiles, which allows to consider more flexible models. We thus chose a regression model that does not involve covariables, and where the regression mean is only assumed to be smooth.

4. In order to validate a given profile using our method, one has to perform several tests: at present, a subject-matter expert of Airpl performs about twelve comparison tests to check spatial, temporal and chemical consistency of a given profile. Moreover it can be seen on the first example in Section 6 that a profile could be

valid even if it is atypical from a given criterion, that is, even if some of the comparison tests reject the null hypothesis. Thus in order to invalidate profiles using our method, one has to work out a strategy for combining the results of the tests and getting a decision. In particular, one has to calibrate the level of the tests. This problem heavily relies on the network of air pollution monitoring stations and is out of the scope of this paper.

4. Testing procedure

Fix $\eta \in (0,1)$ and assume we are given two independent profiles $\mathcal{P}_0$ and $\mathcal{P}_1$ that satisfy conditions of the preceding section. We wish to test the null hypothesis $\mathcal{H}_0$, see (2). We need some more notations. For every real-valued function H defined on $[0,1]$ let r_H and μ_H be defined by

$$r_H = \inf\left\{ r \geqslant 0 : \sup_{\mu}\{H(\mu+r) - H(\mu-r)\} \geqslant \eta H(1) \right\}$$

and

$$\mu_H = \operatorname*{argmax}_{\mu}\{H(\mu+r_H) - H(\mu-r_H)\}.$$

For $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_1\}$, let $F^{\mathcal{P}}$ be defined by $F^{\mathcal{P}}(t) = \int_0^t f^{\mathcal{P}}(s)ds$ for every $t \in [0,1]$ and let define $F_n^{\mathcal{P}}$ by

$$F_n^{\mathcal{P}}(t) = \frac{1}{n}\sum_{i \leqslant nt} y_i^{\mathcal{P}},\ t \in [0,1].$$

Then, $r^{\mathcal{P}} = r_{F^{\mathcal{P}}}$, $\mu^{\mathcal{P}} = \mu_{F^{\mathcal{P}}}$ and for notational convenience, we denote $r_{F_n^{\mathcal{P}}}$ and $\mu_{F_n^{\mathcal{P}}}$ by $r_n^{\mathcal{P}}$ and $\mu_n^{\mathcal{P}}$ respectively. Let K be the quartic function:

$$K(x) = \frac{15}{16}(1-x^2)^2 \mathbb{1}_{[-1,1]}(x).$$

Finally, let $g_n^{\mathcal{P}}$ be the first derivative of $G_n^{\mathcal{P}}$, the smoothed version of $F_n^{\mathcal{P}}$ given by

$$G_n^{\mathcal{P}}(t) = \int_{-1}^{1} F_n^{\mathcal{P}}(t - xh_n^{\mathcal{P}})K(x)\ \mathrm{d}x,\ t \in [0,1], \tag{3}$$

where $h_n^{\mathcal{P}}$ is a smoothing parameter to be fixed (see Sect. 5 for calibration) and where we set $F_n^{\mathcal{P}}(t) = 0$ for every $t \leqslant 0$ and $F_n^{\mathcal{P}}(t) = F_n^{\mathcal{P}}(1)$ for every $t \geqslant 1$. The effective construction of a statistical test for $\mathcal{H}_0$ is as follows. Fix $\alpha \in (0,1)$, the asymptotic level of the test. For $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_1\}$, compute $r_n^{\mathcal{P}}$, $\mu_n^{\mathcal{P}}$, $\mu_{G_n^{\mathcal{P}}}$, $r_{G_n^{\mathcal{P}}}$ and a consistent estimator $\widehat{\sigma_n^{\mathcal{P}}}^2$ of $(\sigma^{\mathcal{P}})^2$ (see *e.g.* Hall *et al.* [3] for examples of such estimators). Condition on $(y_1^{\mathcal{P}}, \ldots, y_n^{\mathcal{P}})$, generate bootstrap residuals $(\varepsilon_1^{\mathcal{P}*}, \ldots, \varepsilon_n^{\mathcal{P}*})$ as i.i.d. Gaussian variables with mean zero and standard deviation $\widehat{\sigma_n^{\mathcal{P}}}$ and compute $\mu_{G_n^{\mathcal{P}*}}$ and $r_{G_n^{\mathcal{P}*}}$, where

$$G_n^{\mathcal{P}*}(t) = \frac{1}{n}\sum_{i \leqslant nt} g_n^{\mathcal{P}}(t_i) + \frac{1}{n}\sum_{i \leqslant nt} \varepsilon_i^{\mathcal{P}*},\ t \in [0,1].$$

Repeat the last step S times (where S is some positive integer to be fixed, see Sect. 5 for calibration) in order to get, conditionally on $(y_1^{\mathcal{P}}, \ldots, y_n^{\mathcal{P}})$, S independent copies of the bootstrap estimators $\mu_{G_n^{\mathcal{P}*}}$ and $r_{G_n^{\mathcal{P}*}}$. For $\beta \in \{\alpha/4, 1-\alpha/4\}$, compute the empirical β-quantile $q^*_{\beta,n,S}$ obtained from the S copies of

$$\left(\mu_{G_n^{\mathcal{P}_0*}} + r_{G_n^{\mathcal{P}_0*}} - \mu_{G_n^{\mathcal{P}_0}} - r_{G_n^{\mathcal{P}_0}}\right) - \left(\mu_{G_n^{\mathcal{P}_1*}} + r_{G_n^{\mathcal{P}_1*}} - \mu_{G_n^{\mathcal{P}_1}} - r_{G_n^{\mathcal{P}_1}}\right)$$

and the empirical β-quantile $s^*_{\beta,n,S}$ obtained from the S copies of

$$\left(\mu_{G_n^{\mathcal{P}_0*}} - r_{G_n^{\mathcal{P}_0*}} - \mu_{G_n^{\mathcal{P}_0}} + r_{G_n^{\mathcal{P}_0}}\right) - \left(\mu_{G_n^{\mathcal{P}_1*}} - r_{G_n^{\mathcal{P}_1*}} - \mu_{G_n^{\mathcal{P}_1}} + r_{G_n^{\mathcal{P}_1}}\right).$$

Finally, reject H_0 if either

$$(\mu_n^{\mathcal{P}_0} + r_n^{\mathcal{P}_0}) - (\mu_n^{\mathcal{P}_1} + r_n^{\mathcal{P}_1}) \notin [q^*_{\alpha/4,n,S}, q^*_{1-\alpha/4,n,S}]$$

or

$$(\mu_n^{\mathcal{P}_0} - r_n^{\mathcal{P}_0}) - (\mu_n^{\mathcal{P}_1} - r_n^{\mathcal{P}_1}) \notin [s^*_{\alpha/4,n,S}, s^*_{1-\alpha/4,n,S}].$$

It follows from Theorems 2.1 and 2.2 of the companion paper Durot and Thiébot [2] that this test has asymptotic level α as S and n go to infinity.

5. Simulations

The testing procedure described in Section 4 involves two parameters to be chosen in a somewhat arbitrary way: the bandwidth $h_n^{\mathcal{P}}$ involved in the bootstrap step and the number of bootstrap replications S. The aim of this section is to provide, *via* simulations, values for S and $h_n^{\mathcal{P}}$ that can be used when applying the method to profiles of pollutants. To run the testing procedure, we need a consistent estimator for $(\sigma^{\mathcal{P}})^2$. In this section and the following one, we consider a difference-based estimator (see Hall *et al.* [3]) that takes the form

$$\widehat{\sigma_n^{\mathcal{P}}}^2 = \frac{1}{n-m}\sum_{k=1}^{n-m}\left(\sum_{j=0}^{m} d_j y^{\mathcal{P}}_{j+k}\right)^2.$$

We fixed $m = 5$ and we chose $(d_0, \ldots, d_5)$ to be the optimal difference sequence (see Tab. 1 of Hall *et al.* [3]), that is

$$(d_0, \ldots, d_5) = (0.9064, -0.2600, -0.2167, -0.1774, -0.1420, -0.1103).$$

Number of bootstrap replications

Concerning the number of bootstrap replications S, it is known that it should be greater than the number of observations n. On the other hand, S should not be too large since computation cost increases with S. We thus carried out simulations as described below with various values for S. We observed that results obtained with $S > 250$ are identical to that obtained with $S = 250$ (which, roughly speaking, means that asymptotic is achieved when $S = 250$) and are in some cases different from that obtained with $S \leqslant 200$. Number of replications $S = 250$ thus seems relevant here.

Bandwidth

Our objective now is to provide values for $h_n^{\mathcal{P}}$ such that the distributions of variables

$$A_n^{\mathcal{P}} = n^{1/3}(\mu_n^{\mathcal{P}} + r_n^{\mathcal{P}} - \mu^{\mathcal{P}} - r^{\mathcal{P}}) \text{ and } B_n^{\mathcal{P}} = n^{1/3}(\mu_n^{\mathcal{P}} - r_n^{\mathcal{P}} - \mu^{\mathcal{P}} + r^{\mathcal{P}}) \tag{4}$$

are close to the distributions of their bootstrap versions

$$A_n^{\mathcal{P}*} = n^{1/3}(\mu_{G_n^{\mathcal{P}*}} + r_{G_n^{\mathcal{P}*}} - \mu_{G_n^{\mathcal{P}}} - r_{G_n^{\mathcal{P}}}) \text{ and } B_n^{\mathcal{P}*} = n^{1/3}(\mu_{G_n^{\mathcal{P}*}} - r_{G_n^{\mathcal{P}*}} - \mu_{G_n^{\mathcal{P}}} + r_{G_n^{\mathcal{P}}}) \tag{5}$$

for a given profile $\mathcal{P}$. Indeed, the testing procedure described in Section 4 is based on the fact that $A_n^{\mathcal{P}}$ and $A_n^{\mathcal{P}*}$ (resp. $B_n^{\mathcal{P}}$ and $B_n^{\mathcal{P}*}$) possess the same asymptotic distribution: it is proved in Durot and Thiébot [2] that these random variables converge in distribution to $\mathcal{C}^{\mathcal{P}}\tau$, where τ is the location of the maximum of a standard two-sided Brownian motion with parabolic drift and $\mathcal{C}^{\mathcal{P}}$ is some positive constant that only depends on the regression mean and on the error variance. The testing procedure thus performs well if the distribution of $A_n^{\mathcal{P}}$

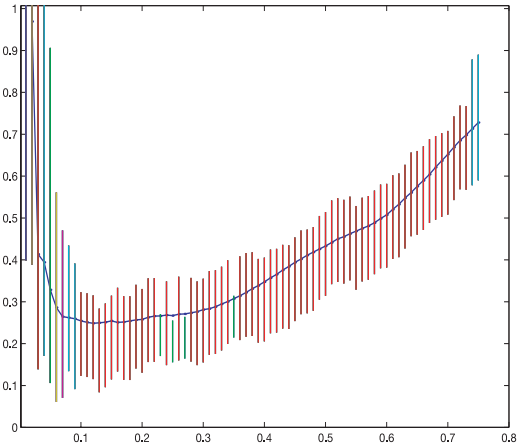

FIGURE 2. Simulated distances. The solid curve is $\bar{d}$ and for each $h_n \in \mathcal{S}_{h_n}$, the line segment through the points $(h_n, m(h_n))$ and $(h_n, M(h_n))$ is drawn.

is close to that of $A_n^{\mathcal{P}*}$ and the distribution of $B_n^{\mathcal{P}}$ is close to that of $B_n^{\mathcal{P}*}$ for each $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_1\}$. For notational convenience and because we consider only one profile at once, we omit superscript $\mathcal{P}$ in the sequel. Moreover, for the sake of simplicity, we only describe in the sequel simulations related to the distributions of A_n and A_n^* (results obtained about distributions of B_n and B_n^* are similar).

In order to calibrate h_n, we carried out simulations as follows. We fixed $\eta = 0.25$, $n = 100$ and $S = 250$ (recall that, when studying profiles of pollutants, experts consider profiles consisting of 96 concentration measures and the most common value they consider for η is 0.25). We fixed f, σ and we set $\mathcal{S}_{h_n} = \{0.01, 0.02, \ldots, 0.75\}$ and $L = 130$. For every $l \in \{1, \ldots, L\}$, we performed Step l in the following way:

- Step l − Draw a n-sample s_l from the Gaussian distribution with mean zero and variance σ^2. Compute the y_i's, A_n and $\widehat{\sigma}_n^2$. Draw independent n-samples $s_{1,l}^*, \ldots, s_{S,l}^*$ from the Gaussian distribution with mean zero and variance $\widehat{\sigma}_n^2$. For every fixed $h_n \in \mathcal{S}_{h_n}$, compute A_n^* in each sample $s_{1,l}^*, \ldots, s_{S,l}^*$ and compute the empirical distribution function $F_{A_n^*}^{(h_n,l)}$ from the S independent copies of A_n^* thus obtained.

These L steps being performed, we got L independent copies of A_n from which we computed the empirical distribution function F_{A_n}. We then computed for every $h_n \in \mathcal{S}_{h_n}$ and $l \in \{1, \ldots, L\}$ the L_2-distance between F_{A_n} and $F_{A_n^*}^{(h_n,l)}$, which is denoted in the sequel by $d(h_n, l)$. Finally, we computed

$$m(h_n) = \min_{1 \leqslant l \leqslant L} d(h_n, l), \; M(h_n) = \max_{1 \leqslant l \leqslant L} d(h_n, l) \text{ and } \bar{d}(h_n) = \frac{1}{L} \sum_{l=1}^{L} d(h_n, l).$$

Our aim is to provide values for h_n where $m(h_n)$, $M(h_n)$ and $\bar{d}(h_n)$ are minimum.

Let us describe more precisely the results we obtained for a given regression function f and a given standard deviation σ. Fix $\sigma = 0.1$ and let f be defined by

$$f(t) = \frac{4}{\sqrt{2\pi}} \exp\left(-8(t - 0.5)^2\right), \; t \in [0, 1]. \tag{6}$$

Performing simulations just described, we obtained results summarized in Figure 2. It is seen in Figure 2 that the most relevant values for h_n belong to $[0.06, 0.3]$ when $\sigma = 0.1$ and f is given by (6). This means that the procedure is not very sensitive to the choice of the bandwidth h_n: one can choose h_n in a quite large interval so that the distance between the distributions of A_n and A_n^* is (approximately) minimal.

TABLE 1. Belonging to bootstrap confidence intervals. We reported the proportion of steps when $\mu+r$ belongs to (7) on line $\in$, the proportion of steps when $\mu+r < \mu_n+r_n-n^{-1/3}q^*_{1-\alpha/2}$ on line $<$ and proportion of steps when $\mu+r > \mu_n+r_n-n^{-1/3}q^*_{\alpha/2}$ on line $>$.

$1-\alpha$	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91	0.90	0.85	0.80
$<$	0.012	0.019	0.022	0.026	0.030	0.036	0.040	0.046	0.054	0.084	0.110
$\in$	0.981	0.970	0.962	0.953	0.946	0.938	0.928	0.917	0.906	0.858	0.813
$>$	0.007	0.011	0.016	0.021	0.024	0.026	0.032	0.037	0.040	0.058	0.077

We repeated simulations just described with various f and σ that are close to regression means and error variances we observe on typical profiles. It emerges from these simulations that the optimal choice for h_n depends on the ratio f/σ. We drawn from these simulations a criterion for choosing h_n: choose $h_n = 0.22\sqrt{\widehat{\sigma}_n/F_n(1)}$ if $\widehat{\sigma}_n/F_n(1) \leqslant 0.15$ and $h_n = 0.32\sqrt{\widehat{\sigma}_n/F_n(1)}$ otherwise.

To illustrate this choice for h_n, we reported here another simulation study. We considered the regression function (6) and we fixed $\eta = 1/4$, $\sigma = 0.1$, $n = 100$, $S = 500$. Since $\sigma/F(1) \leqslant 0.15$, we chose h_n once and for all to be $0.22\sqrt{\sigma/F(1)}$, that is $h_n = 0.07$. We repeated 1000 times the following step.

- Step $-$ Draw a n-sample from the Gaussian distribution with mean zero and variance σ^2. Compute the y_i's and $\widehat{\sigma}_n^2$. Draw independent n-samples $s_1^*, \ldots, s_S^*$ from the Gaussian distribution with mean zero and variance $\widehat{\sigma}_n^2$. Compute the empirical $(1-\alpha/2)$ quantile $q^*_{1-\alpha/2}$ and the empirical $\alpha/2$ quantile $q^*_{\alpha/2}$ of the S copies of A_n^* obtained from these S samples.

We reported on Table 1, line "$\in$", the proportion of steps when $\mu + r$ belongs to

$$\left[\mu_n + r_n - n^{-1/3}q^*_{1-\alpha/2}, \mu_n + r_n - n^{-1/3}q^*_{\alpha/2}\right]. \tag{7}$$

Interval (7) is an asymptotic $(1-\alpha)$-confidence interval for $\mu + r$ as S and n go to infinity, so it is expected that the proportion of steps when $\mu + r$ lies in (7) is approximately $1-\alpha$. It is seen on Table 1 that it is indeed the case.

6. SOME APPLICATIONS

We illustrate the proposed testing procedure through examples of checks on spatial consistency. We consider here three urban sites in the town of Nantes. These sites are called Théâtre Graslin, Sainte Luce and Bellevue and we refer to them as Tgra, Stel and Bell in the sequel. The site Tgra is located in the centre of the town and is more submitted to urban pollution than the two other sites. Sites Stel and Bell are located respectively at the north-eastern and the western peripheries of Nantes. The three sites are subjected to industrial pollution generated by sites located to the west and to the north-west of Nantes. Data that were available to us are profiles from September 1993 until February 1999, so we consider here profiles that were collected on December 1998.

Consider first the profiles of nitrogen dioxide collected on December 24th 1998, in the sites Tgra, Stel and Bell. These profiles are drawn in Figure 3 together with the associated estimated η-shorth intervals (we fixed here $\eta = 0.25$, as experts do). When comparing the η-shorth intervals of Stel and Tgra, we obtained a p-value equal to 0.98, so the null hypothesis that the two η-shorth intervals are equal is validated and the two profiles are consistent with each other. On the other hand, we obtained a p-value equal to 0.01 (respectively 0.01) when comparing η-shorth interval associated to Bell to the one associated to Tgra (respectively Stel). Therefore,

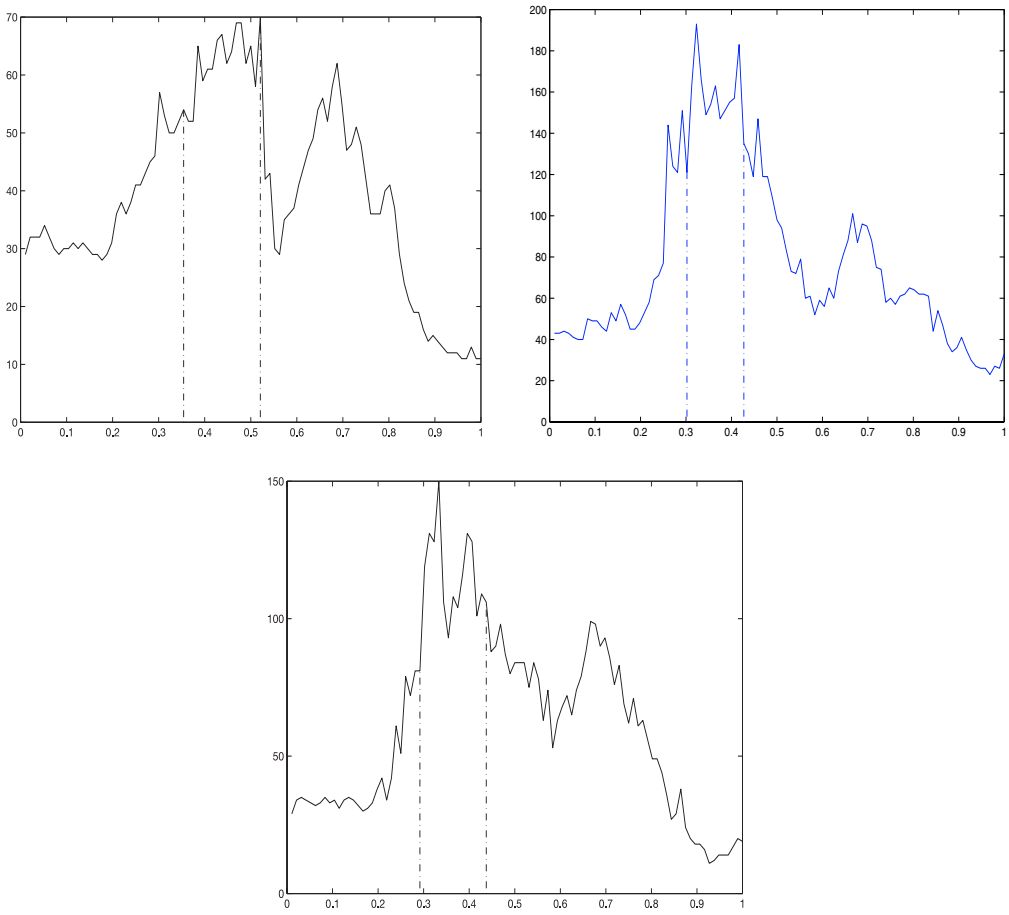

FIGURE 3. Profile of nitrogen dioxide on December 24th, 1998, in Bell (up left), Tgra (up right) and Stel (down).

profile in Bell is atypical and has to be submitted to experts. Note that the profile in Bell was not invalidated by experts: the difference between this site and the others was explained by a sudden change in the wind direction (it changed from north-west to west/south-west) which led to a delay of the peak in Bell.

Consider now the profiles of nitrogen monoxide collected on December 4th 1998, in the sites Tgra and Bell. These profiles are drawn in Figure 4 together with the associated estimated η-shorth intervals (we fixed again $\eta = 0.25$). When comparing the η-shorth intervals we obtained a p-value equal to 0.00, so profiles are not consistent with each other. Performing other comparison tests, it appears that the profile in Tgra is atypical and have to be submitted to experts. The profile in Tgra was indeed invalidated by experts: data collected between times 0.2 and 0.3 were not valid following a momentary failure of monitoring device.

7. Concluding remarks

In this section, we discuss possible improvements and generalizations of the method.

Null hypothesis

The more relevant null hypothesis to be considered for comparing two profiles is the one given in (2). However, our method can easily be adapted to tests for other null hypotheses using the fact that for a given profile $\mathcal{P}$,

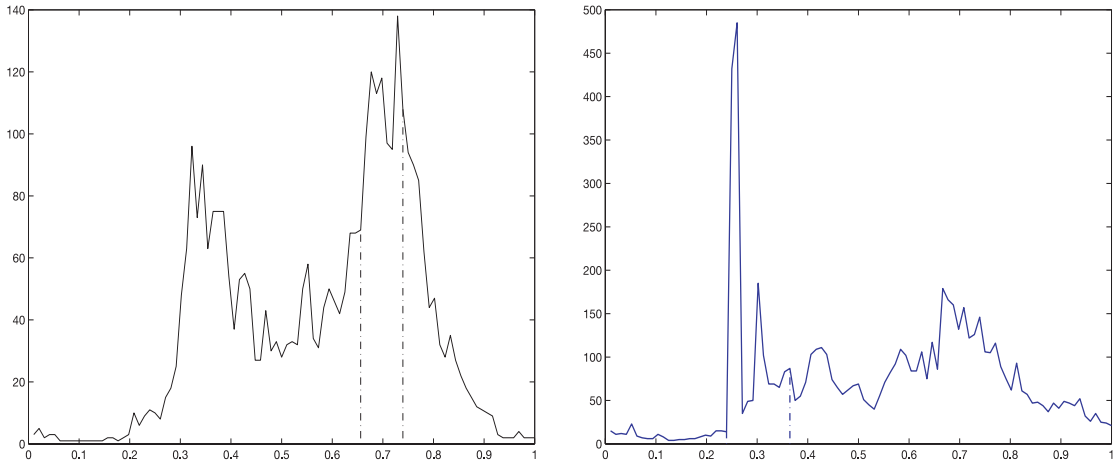

FIGURE 4. Profile of nitrogen monoxide on December 4th, 1998, in Bell (left) and Tgra (right).

the variables $A_n^{\mathcal{P}}$ and $B_n^{\mathcal{P}}$ have approximately the same distributions as their bootstrap versions $A_n^{\mathcal{P}*}$ and $B_n^{\mathcal{P}*}$, see (4) and (5). Assume for instance we are interested in a single profile $\mathcal{P}$ and we whish to test the null hypothesis "$\mu^{\mathcal{P}} - r^{\mathcal{P}} \leqslant c$" for a given $c \in (0,1)$. Let $t_{1-\alpha}^*$ be the empirical $(1-\alpha)$-quantile obtained from S copies of $n^{-1/3}B_n^{\mathcal{P}*}$. Then the test that rejects the null if $\mu_n^{\mathcal{P}} - r_n^{\mathcal{P}} - c > t_{1-\alpha}^*$ has asymptotic level α.

Bootstrap

The bootstrap method involved in this paper rests on the assumption that the error terms $\varepsilon_1^{\mathcal{P}}, \ldots, \varepsilon_n^{\mathcal{P}}$ in each profile $\mathcal{P}$ are i.i.d. Gaussian so one may wonder whether the method is robust against departures from either normality or homoscedasticity. It can be shown that it is indeed the case, provided the bootstrap residuals $\varepsilon_1^{\mathcal{P}*}, \ldots, \varepsilon_n^{\mathcal{P}*}$ are generated with an adequate bootstrap method, see Durot and Thiébot [2]. To give examples, let us omit subscript $\mathcal{P}$ and set $\hat{\varepsilon}_i = y_i - \hat{f}_n(t_i)$, where $\hat{f}_n$ is a proper estimator of f (one can consider for instance the first derivative of G_n, see (3)). If the error terms are assumed i.i.d. but not necessarily Gaussian then one can generate the bootstrap residuals as a random sample of size n from the distribution that puts, conditionally on the y_i's, mass $1/n$ at each point $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ with $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \sum_j \hat{\varepsilon}_j/n$. If the error terms are not assumed homoscedastic then one can generate bootstrap residuals as $\varepsilon_i^* = \hat{\varepsilon}_i V_i$, where $V_1, \ldots, V_n$ are i.i.d. random variables with zero mean that are independent of $y_1, \ldots, y_n$. Both these bootstrap methods lead to tests with prescribed asymptotic level under appropriate assumptions.

Bandwidth

One may wonder whether the bandwidth h_n could be calibrated from real data rather than from simulations. We think this is possible if we have at hand a sufficiently large history of profiles together with a classification such that the estimators $\mu_n^{\mathcal{P}} - r_n^{\mathcal{P}}$ (resp. $\mu_n^{\mathcal{P}} + r_n^{\mathcal{P}}$) have approximately the same distribution for all profiles $\mathcal{P}$ of a given class. In such a case, we can perform simulations as in Section 5 with the following slight changes: $n = 96$, L is the number of profiles in a given class $\mathcal{C}$ and in Step l, y_i is the ith concentration measure in the l-th profile of $\mathcal{C}$. Then we can choose a value $h_n^{\mathcal{C}}$ for h_n where $\bar{d}(h_n)$, $m(h_n)$ and $M(h_n)$ are (almost) minimal. Now in order to study a new profile $\mathcal{P}$, one has to determine the class $\mathcal{C}$ that better represents $\mathcal{P}$ and then to apply the method

of Section 4 with the smoothing parameter $h_n^{\mathcal{C}}$. However, we do not known how to draw such a classification of profiles. According to Theorem 2.1 in Durot and Thiébot [2], a possible requirement is that all the profiles in a given class possess the same parameters $\mu^{\mathcal{P}}$, $r^{\mathcal{P}}$, $f^{\mathcal{P}}(\mu^{\mathcal{P}} + r^{\mathcal{P}})/\sigma^{\mathcal{P}}$ and $(f'^{\mathcal{P}}(\mu^{\mathcal{P}} - r^{\mathcal{P}}) - f'^{\mathcal{P}}(\mu^{\mathcal{P}} + r^{\mathcal{P}}))/\sigma^{\mathcal{P}}$.

References

[1] L. Bel, L. Bellanger, V. Bonneau, G. Ciuperca, D. Dacunha-Castelle, C. Deniau, B. Ghattas, Y. Misiti and G. Oppenheim, Éléments de comparaison de prévisions statistiques des pics d'ozone. *Rev. Statist. App.* **3** (1999) 7–25.

[2] C. Durot and K. Thiébot. *Bootstrapping the shorth for regression.* Submitted (2003).

[3] P. Hall, J.W. Kay and D.M. Titterington, Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** (1990) 521–529.

[4] K. Thiébot, Synthèse de l'enquête sur la procédure de validation de données dans les résaux de surveillance de pollution athmosphérique. *Technical report, Air Pays de la Loire* (1998).