Listy Biometryczne - Biometrical Letters
Vol. 35(1998), No. 1, 37-66

On forecasting ozone episodes in the Paris area*

Liliane Bel[1], Lise Bellanger[1,2], Michel Bobbia[1], Gabriela Ciuperca[1,2], Didier Dacunha-Castelle[1], Elisabeth Gilibert[4], Patrick Jakubowicz[1], Georges Oppenheim[1] and Richard Tomassone[2,3]

[1]Modélisation Stochastique et Statistique dOrsay, Bât.425, Université Paris XI, 91405 Orsay Cedex

[2]Mathématique et Informatique, Institut National Agronomique, 16 rue Claude Bernard, 75231 Paris Cedex 05

[3]CNRS-UMR 5558, Université Claude Bernard Lyon I, 43 Bd du 11 novembre 1918, 69622 Villeurbanne Cedex

[4]AIRPARIF, Surveillance de la Qualité de lAir en Ile-de-France, 10 rue Crillon, 75004 Paris

SUMMARY

This paper presents two different statistical strategies to elucidate the dependence of ozone on primary pollutants (nitrogen oxides) and on meteorology. The aim is to forecast, at 8am of a current day, the maximum ozone value occurring in the afternoon, using 6 years (1990-95) of pollutant and meteorological data for Paris. The first method is based on classical methods using simultaneously cluster analysis, analysis of variance, discriminant analysis and stepwise regression. We identify three distinct and homogeneous groups in Paris area. Within these groups, daily curves of ozone pollution form clusters of decreasing levels; these clusters are well discriminated by previous ozone, primary pollutant and meteorological data. The second method is based on nonparametric methods using a kernel estimator of an autoregressive function with exogenous variables. It works by analogy on climatic and pollution conditions. The forecast is a weighted sum of maxima observed in the past. We compare the two methods on 1996 data, and propose some improvements to avoid forecast errors in particular cases.

KEY WORDS: air pollution, ozone concentration, prediction, linear model, kernel nonparametric forecasting.

*The paper was submitted on the occasion of 70-th birthday of Professor Tadeusz Caliński.

1. Introduction

As it occurs in all great cities, Paris has a serious photochemical ozone (O_3) air pollution problem. When the urban emission pattern of O_3-forming pollutants is fairly uniform, it seems that the variations of O_3 concentrations are controlled by meteorological factors and by a series of atmospheric reactions between primary pollutants such as nonmethane hydrocarbons (HC) in the presence of nitric oxide (NO) and nitrogen dioxide (NO_2).

Accurate O_3 forecasts can be used to protect sensitive individuals from excessive concentrations of O_3, when sufficient leading time is given to the public. They can also be used as a guidance in air pollution advisory committees, giving to the administration some tools to help to decide whether short-term control strategies need to be considered during air pollution episodes. Some similar studies were undertaken by Clark (1982), Eder et al. (1994), Rhodes and Miller-Gonzalez (1994) and Ryan (1995).

Our modelling approach consists in determining a statistical relationship between a predictand (O_3) and variables (henceforth called predictors) from either nitrogen oxides (NO and NO_2) or meteorological measurements (temperatures and wind velocity at different heights). The following data are available:

– one-hour pollution measurements from 8 air quality stations measuring O_3, NO and NO_2 (measurements initiated at different dates from 1990 to 1995, see Table 1).

– one-hour meteorological measurements from Saclay site (in the suburbs of Paris); these data are supposed to be representative of climatic conditions all over the studied Ile-de-France area.

We must specify here that the period under study is a summer one (from May 1st to September 15th), as air pollution occurs during it.

The methods we used were either "classical" or "more advanced". By classical, we mean methods already used in similar situations (clustering, regression, discrimination) adapted to our case study; by more advanced, we mean nonparametric ones based on kernel estimator of an autoregressive function with exogenous variables. All variables used start from midday on the day before prediction to 8am on the day of prediction.

First, we shall present available data (section 2); in a second step the different methods (section 3) and, at last, results with possible improvements (section 4).

The study was administered through a contract with the Orsay University Statistical Laboratory (Paris XI University) and sponsored by AIRPARIF (Surveillance de la Qualité de lAir en Ile-de-France).

2. Data

2.1. The 8 stations for pollutants

The one-hour pollution measurements were made on 8 stations, as indicated in Table 1 and Figure 1. One of the major, and primary, problems is existence of missing data, due to failures of sensors; these missing data represent 10% of the data base. When we have few such data in a sequence (one or two hours) we reconstructed them by spline interpolations; when the gap is greater we treated them as missing data.

All three pollutants (O_3 as the predictand, NO and NO_2 as predictors) were measured simultaneously. The following notations are used with the common measurement unit ($\mu g/m^3$): P (O_3 concentration), N (NO concentration) and M (NO_2 concentration). We shall also use global nitrogen oxides MN = N/30 + M/46 (30 and 46 corresponding, respectively, to molecular weights of the two components).

Table 1. Data for the 8 stations with starting year of measurements

Station	Code	Starting year
Neuilly	1F92	1990
Paris, 13th district	13F75	1991
Tour Eiffel, Paris 7th district	7F75	1992
Aubervilliers	1F93	1990
Créteil	1F94	1990
Montgeron	4F91	1994
Champ-sur-Marne	1F77	1990
Tour Saint Jacques, Paris 4th district	4F75	1991

2.2. Saclay site for meteorological variables

Temperatures were measured in °C at three heights: 0m, 40m and 100m. We used:

- T : soil-temperature (0m), and two *gradients*
- G : (Temperature at 40m - T)/0.4,
- H : (Temperature at 100m - T),
- V and W : wind velocity (unit: m/s) measured at 110m (V) and at 58m (W).

In the following observations of each variable will be indexed by i denoting time (hour) so we shall use the symbols P_i, N_i, M_i, T_i, G_i, H_i, V_i and W_i. Except if differently indicated in text, for pollutant variables, $i = 0$ corresponds to midnight and $i = 7$ to 7am the current day; for meteorological variables $i = 1$ corresponds to midday the day before prediction, and $i = 21$ to 8am the current day.

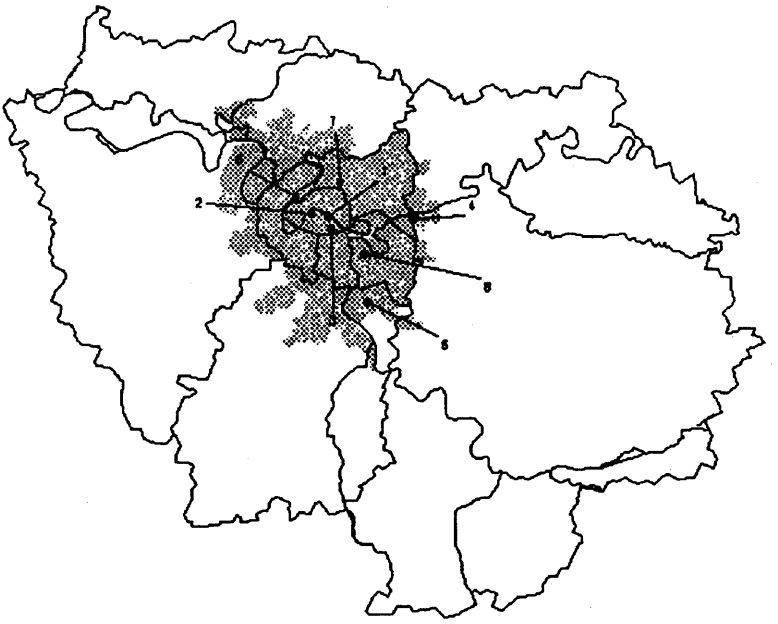

Figure 1. Network AIRPARIF. Positions of the 8 stations in Ile-de-France.

2.3. AIRPARIF alarm rules

AIRPARIF had already defined some alarm rules concerning O_3 (Table 2); in fact level 3 was never attained, and is out of our topic.

Table 2. AIRPARIF alarm rules

Alarm level	Rule
0	7 stations with O_3 concentrations less than 130 g/m^3
1	at least 2 stations with O_3 concentrations exceeding 130 g/m^3
2	at least 2 stations with O_3 concentrations exceeding 180 g/m^3
3	at least 2 stations with O_3 concentrations exceeding 360 g/m^3

One of the constraints of our study was to apply the same rules after having made O_3 predictions for the eight stations. So, we are going to mimic the same strategy, even if we may ulteriorly propose an other one for later studies on alarm probability level.

3. Methodology

As indicated above, we shall propose two different methodologies to forecast O_3, linear and non-parametric methods.

3.1. Linear methods

For O_3 the total number of available observations (day×station) is 828 for the period from 1990 to 1995 during the pollution period (May-September). Clearly, the data base being not sufficient for every station, it was difficult to develop a statistical model for each of them; so we decided to cluster similar stations. This was done in two major steps:

– a *reduction step* in which we summarize variables and groups, in order to have easily interpretable variables and homogeneous subgroups,

– a *modelling step* in which we shall furnish models for O_3 forecast in subgroups.

A description of all methods used may be found in books on the linear model and multivariate analysis, see for example Tomassone et al. (1993); all computations were made using SAS (1985).

3.1.1. Reduction step

3.1.1.1. Existence of a "virtual Paris"

The first step consists in an ANOVA (Analysis of Variance) for every P_i. The statistical model is the following:

$$P_{ijhk} = \mu_i + sta_{ih} + an_{ik} + jour_{ijk} + e_{ijhk} \tag{1}$$

where:

– i corresponds to the hour ($i = 1, ..., 24$),

– j corresponds to the day ($j = 1, ..., 828$, the total number of days in the data base),

– h corresponds to the station ($h = 1, ..., 8$),

– k corresponds to the year ($k = 1$ for 1990, 6 for 1995),

and

– P_{ijhk} is the observed value of O_3,

– μ_i is the mean effect for hour i (if the "design" were balanced it would be the mean value for O_3 at this hour),

– sta_{ih} is the station effect (a qualitative one),

– an_{ik} is the year effect,

– $jour_{ijk}$ is the day effect (a qualitative one),

– e_{ijhk} is a random effect with classical assumptions of independence and Normal distribution with zero mean and constant variance σ_i^2.

Model (1) is not a complete one. Due to the lack of data we are not able to insert in it a day×station interaction. By estimating station effect, it is possible to analyze similarity of the stations and to cluster them.

3.1.1.2. Clusters of stations

Each station effect is associated with a Student statistic t_{ih} which permits us to analyze a matrix (24 rows×8 columns) by a Principal Component Analysis (PCA). A graphical analysis of PCA results leads us to cluster the 8 stations in 3 subgroups (Table 3). With only 2 subgroups, the heterogeneity is too large within one of them; with 4 subgroups the only difference is due to the splitting of subgroup 2 (1F93, 4F91 on one part and 1F94 on the other).

Table 3. Stations clustered in three subgroups

Subgroup	Stations
PM1	13F75, 1F92, 7F75
PM2	1F93, 1F94, 4F91
PM3	1F77, 4F75

Having this result, we do once again an ANOVA similar to (1) for each subgroup g (g = PM1, PM2, PM3). Eliminating station and year effects, we are able to estimate $P_{ij++(g)}$ by $\hat{P}_{ij++(g)}$ for each hour (i) and for each day (j). These values give a curve (with hour in abscissa) characteristic of each day under analysis (Figure 2).

For other pollutants (NO and NO_2) it is also possible to obtain similar curves using ANOVA modelling by computing estimated values $\hat{N}_{ij++(g)}$ and $\hat{M}_{ij++(g)}$. We are now in situation to analyze O_3 evolution by modelling $\hat{P}_{j++(g)}$ (the predictand) as a statistical function of $\hat{N}_{ij++(g)}$, $\hat{M}_{ij++(g)}$, T_{ij}, G_{ij}, H_{ij} and V_{ij} (the predictors).

This analysis was performed also with other available statistical tools. By using influence diagnostics, we were faced with outliers. We decided to delete all observations in 1991 for 1F92 station, because of errors in sensors.

3.1.1.3. The intermediary predictors

At this step we deal with the problem of estimating the maximum O_3 value for a day (always encountered between 2pm and 8pm) as a function of:

– pollutants values during the night preceding the day for which we want to forecast maximum O_3 value: P_0 to P_7 (autoregression), N_0 to N_7 and M_0 to M_7; the values before seem to be not significant, except a remanence of atmospheric conditions captured by $P_{\max 0}$, the maximum observed value of the previous day (0 indicates observed value at midnight and 7 the observed value at 7am).

– temperature and gradient values T_1 to T_{21}, G_1 to G_{21}, H_1 to H_{21} (1 indicates observed value at midday the previous day and 21 at 8am).

– wind velocity values V_4 to V_8 (4 indicates observed value at 4am and 8 at 8am).

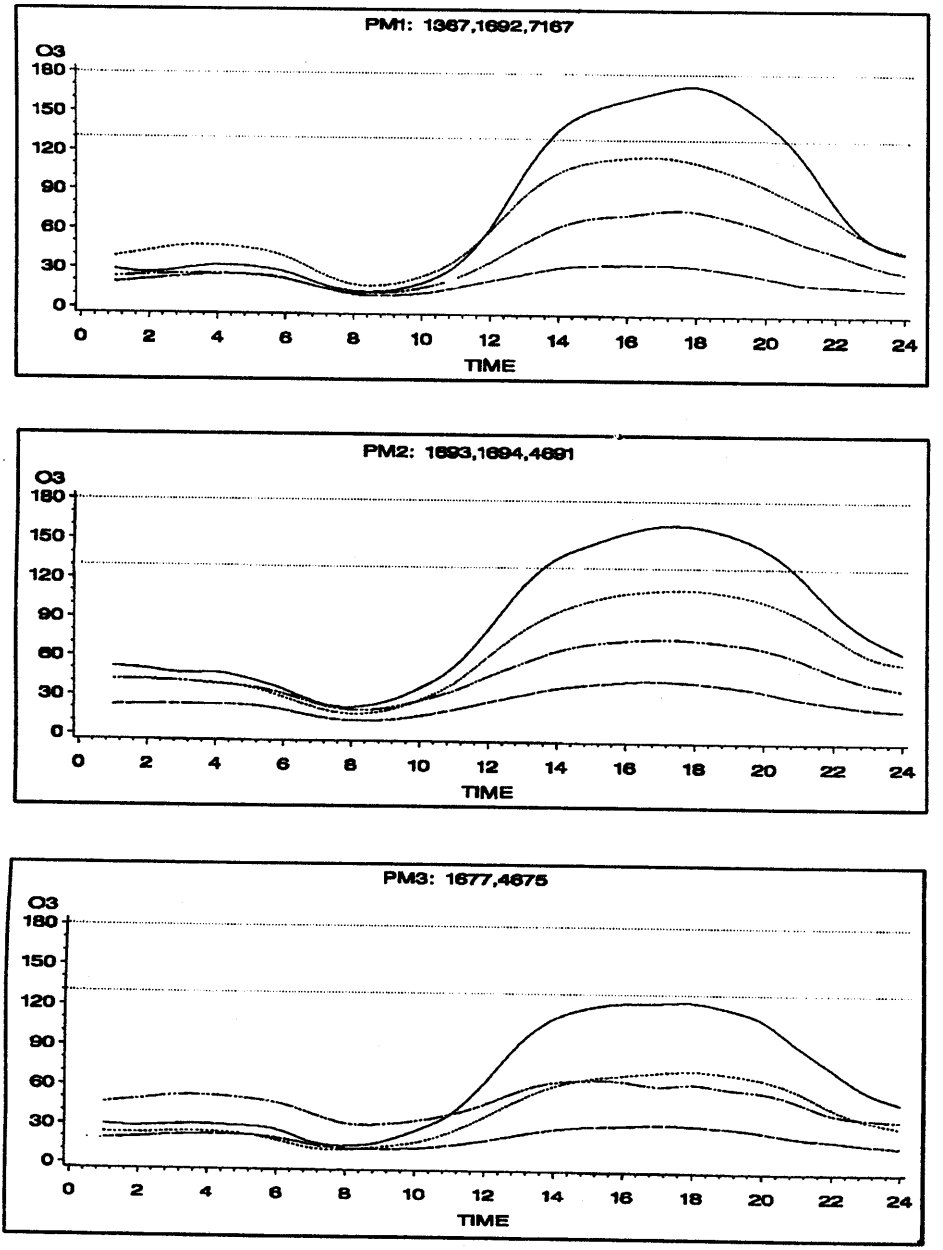

Figure 2. Graphs characteristic of Ozone evolution from 1 am to 12 pm

The total number of predictors is 91. It was difficult to use directly these variables because their number is too large and all of them are too much correlated. This multicollinearity induces very instable results. It was decided to transform them by using linear combinations; the transformations are based either on physical knowledge, as the molar sum for nitrogen oxides

$$MN_i = M_i/46 + N_i/30, \tag{2}$$

or on linear transformations given by PCA. But, when we did so for each variable, we saw that the successive principal components (accounting for more than 95% of total variability) had very simple interpretations: the first was practically the mean, the second a linear tendency, the third a quadratic component and the fourth, if any, a cubic (Table 4). So, having these so simple interpretations in mind, we thought it was better to introduce these natural transformations, classical in ANOVA for studying components effects. One of the primary reason was justified by the ability to estimate them even with partially missing data; in fact this possibility was not used until now. When the percentages of variability accounted for by these linear transformations were not sufficient, we added to them some other simple linear transformations such as mean on an interval between two different hours.

The total number of variables is now 48. We may guarantee that they contain practically the same information as the basic ones, and have the advantage of being well understood from a physical point of view. We may say that they represent, for each set (P, N, M, T, G, H, V and W), a good summary of the 8 curves. Now we are able to focus our interest on our principal aim: how these predictors, which contain practically all our available information in a more stable form than the basic data, may help us in forecasting maximum pollution, knowing a short history of pollution and meteorological evolution?

But these predictors are still highly correlated: until now transformations concern only components within the eight sets of data (P, M, N, T, G, H, V and W). So, multicollinearity still exists between the sets, which may cause difficulties in obtaining stable results. Therefore we decided to introduce new predictors to protect against this fact.

3.1.1.4. The final predictors

The idea is simple: we do some regressions and use the residuals which have the advantage of being uncorrelated with variables used as regressors. More formally, this means that having a variable we compute first its regression equatio.. on a specific set of other variables; as an example MP0203 may be regressed on MP0607 and MP0405, giving:

$$E(MP0203) = b_0 + b_1 MP0607 + b_2 MP0405.$$

Table 4. First transformations from basic variables

New variables	Linear combinations
Ozone pollutant (9 autopredictors)	
LP2407	$(-7*P_0-5*P_1-3*P_2-1*P_3+1*P_4+3*P_5+5*P_6+7*P_7)/8$
QP2407	$(+7*P_0+1*P_1-3*P_2-5*P_3-5*P_4-3*P_5+1*P_6+7*P_7)/8$
CP2407	$(-7*P_0+5*P_1+7*P_2+3*P_3-3*P_4-7*P_5-5*P_6+7*P_7)/8$
MP2407	$(P_0+P_1+P_2+P_3+P_4+P_5+P_6+P_7)/8$
MP2401	$(P_0+P_1)/2$
MP0203	$(P_2+P_3)/2$
MP0405	$(P_4+P_5)/2$
MP0607	$(P_6+P_7)/2$
PMAX0	Maximum Ozone for the previous day
Nitrogen pollutant (9 predictors)	
MM0607	$(M_6+M_7)/2$
MN0	$N_0/30+M_0/46$
MN1	$N_1/30+M_1/46$
MN2	$N_2/30+M_2/46$
MN3	$N_3/30+M_3/46$
MN4	$N_4/30+M_4/46$
MN5	$N_5/30+M_5/46$
MN6	$N_6/30+M_6/46$
MN7	$N_7/30+M_7/46$
Temperature (7 predictors)	
MT1518	$(T_4+T_5+T_6+T_7)/4$
MT1921	$(T_8+T_9+T_{10})/3$
MT2207	$(T_{11}+T_{12}+T_{13}+T_{14}+T_{15}+T_{16}+T_{17}+T_{18}+T_{19}+T_{20})/10$
QT1421	$(7*T_3+1*T_4-3*T_5-5*T_6-5*T_7-3*T_8+1*T_9+7*T_{10})/8$
DT1907	$(T_8-T_{20})/12$
T20	
T21	
Gradient 40m (9 predictors)	
DG2319	$(G_{12}-G_8)/4$
LG2407	$(-7*G_{13}-5*G_{14}-3*G_{15}-1*G_{16}+1*G_{17}+3*G_{18}+5*G_{19}+7*G_{20})/8$
QG2407	$(+7*G_{13}+1*G_{14}-3*G_{15}-5*G_{16}-5*G_{17}-3*G_{18}+1*G_{19}+7*G_{20})/8$
MG2407	$(G_{13}+G_{14}+G_{15}+G_{16}+G_{17}+G_{18}+G_{19}+G_{20})/8$
DG0723	$(G_{20}-G_{12})/8$
MG2402	$(G_{13}+G_{14}+G_{15})/3$
MG1416	$(G_3+G_4+G_5)/3$
G20	
G21	
Gradient 100m (9 predictors)	
DH2319	$(H_{12}-H_8)/4$
LH2407	$(-7*H_{13}-5*H_{14}-3*H_{15}-1*H_{16}+1*H_{17}+3*H_{18}+5*H_{19}+7*H_{20})/8$
QH2407	$(+7*H_{13}+1*H_{14}-3*H_{15}-5*H_{16}-5*H_{17}-3*H_{18}+1*H_{19}+7*H_{20})/8$
MH2407	$(H_{13}+H_{14}+H_{15}+H_{16}+H_{17}+H_{18}+H_{19}+H_{20})/8$
DH0723	$(H_{20}-H_{12})/8$
MH0307	$(H_{16}+H_{17}+H_{18}+H_{19}+H_{20})/5$
MH1416	$(H_3+H_4+H_5)/3$
H20	
H21	
Wind velocity (5 predictors)	
MV0407	$(V_4+V_5+V_6+V_7)/4$
LV0407	$(-3*V_4-1*V_5+1*V_6+3*V_7)/4$
QV0407	$(V_4-V_5-V_6-V_7)/4$
V7	
V8	

Table 5. The final predictors, residuals are coded with a leading R

Predictors	Regressors
Ozone pollutant (9 autopredictors)	
MP0607	$(P_6+P_7)/2$
PMAX0	Maximum Ozone for the previous day
MP0405	MP0607
MP0203	MP0607 MP0405
MP2401	MP0607 MP0405 MP0203
MP2407	MP0607
LP2407	MP0607 MP2407
QP2407	MP0607 MP2407 LP2407
CP2407	MP0607 MP2407 LP2407 QP2407
Nitrogen pollutant (9 predictors)	
MM0607	$(M_6+M_7)/2$
MN3	N3/30+M3/46
RMN2	MN3
RMN1	MN3 MN2
RMN0	MN3 MN2 MN1
RMN4	MN3
RMN5	MN3 MN4
RMN6	MN3 MN4 MN5
RMN7	MN3 MN4 MN5 MN6
Temperature (7 predictors)	
MT1518	$(T_4+T_5+T_6+T_7)/4$
QT1421	$(7*T_3+1*T_4-3*T_5-5*T_6-5*T_7-3*T_8+1*T_9+7*T_{10})/8$
DT1907	$(T_8-T_{20})/12$
T20	
T21	
MT2207	T20
RMT1921	T20 MT2207
Gradient 40m (9 predictors)	
MG2402	$(G_{13}+G_{14}+G_{15})/3$
G21	
MG2407	H20
LG2407	H20 MG2407
QG2407	H20 MG2407 LG2407
RG20	H20
DG2319	DH2319
DG0723	DH0723
MG1416	MH1416
Gradient 100m (9 predictors)	
DH2319	$(H_{12}-H_8)/4$
DH0723	$(H_{20}-H_{12})/8$
MH1416	$(H_3+H_4+H_5)/3$
H20	
H21	
MH2407	H20
LH2407	H20 MH2407
QH2407	H20 MH2407 LH2407
MH0307	H20
Wind velocity (5 predictors)	
V7	
V8	
MV0407	V7
LV0407	V7 MV0407
QV0407	V7 MV0407 LV0407

Then, we use as the final predictor not correlated with MP0607 and MP0405

$$RMP0203 = MP0203 - E(MP0203).$$

The final set of variables is given in Table 5.

3.1.2. Modelling step

Having the 3 subgroups (PM1, PM2 and PM3), we are going to analyze them separately in the following manner:

(1) use of the profile for O_3 curve to define classes of similar pollution,

(2) separation of classes through an DFA (Discriminant Factorial Analysis) using the best predictors within the 48 previous ones,

(3) building of a regression model within each class to forecast maximum O_3 pollution.

3.1.2.1. How to build O_3 pollution classes

Model (1) applied to each subgroup helps us to define a "mean-day", corrected for year and station. The data matrix (rows: mean-day; columns: hour) is clustered by rows using a method that Anderberg (1973) calls nearest centroid sorting. A set of points (mean-day) is selected (the cluster seeds) as a first guess of the means of the clusters. Each mean-day is assigned to the nearest seed to form temporary clusters. The seeds are then replaced by the means of the temporary clusters. The process is repeated until no changes occur in the clusters. The clustering is done on the basis of Euclidean distances computed on standardized variables (the 24 hours of a day). This process also permits to detect outliers: the corresponding mean-day often appears as a cluster with one member only (procedure FASTCLUST in SAS).

By trial and error, in every subgroup we obtain 4 classes C_t $(t = 1, 2, 3, 4)$ by decreasing order of maximum pollution from C_1 (most polluted) to C_4 (least polluted). The sizes of classes are given in Table 6, with years used in each station.

3.1.2.2. How to classify one day in a station in a pollution class

The following step consists in an DFA using the 48 predictors (autopredictors and predictors) as shown in Table 5. But, here again, we have tried to delete some predictors, having always in mind that it is better to use a minimal number of predictors to obtain stable results. The process was the following:

– *univariate* test on every predictor by an F-test in the 48 ANOVA, associated with *multivariate analysis* of the resulting discrimination (because a predictor may be of no interest alone, but extremely important if associated with one or more others),

– a *stepwise selection*, using STEPDISC procedure in SAS, to remove non-significant predictors,

– common sense rules to minimise misclassification, as we shall see further.

The selected predictors are furnished in Table 7.

Table 6. Classes size in each subgroup

Subgroup	Class C_t	Size (#mean-day)	O_3 maximum	Station	Year					
					90	91	92	93	94	95
PM1	1	23	171	1F92	*		*	*	*	*
	2	70	117	13F75			*	*	*	*
	3	157	76	7F75			*	*	*	*
	4	326	35							
Total		576								
PM2	1	19	162	1F93	*	*	*	*	*	*
	2	109	113	1F94	*	*	*	*	*	*
	3	172	76	1F91					*	*
	4	460	43							
Total		660								
PM3	1	39	125	1F77	*	*	*	*	*	*
	2	186	74	4F75		*	*	*	*	*
	3	44	66							
	4	477	33							
Total		718								

Table 7. Predictors for discrimination in the 3 subgroups

Subgroup	Number of predictors	Selected predictors for discrimination
PM1	33	6 : MP0607, RMP0405, RMP2401, RMP2407, RQP2407, RCP2407 8 : MM0607, MN3, RMN2, RMN1, RMN0, RMN4, RMN5, RMN6, RMN7 4 : MT1518, DT1907, RMT2207, RMT1921 4 : MG2402, RMG2407, RG20, RDG0723 8 : DH2319, DH0723, MH1416, H20, RMH2407, RLH2407, RQH2407, RMH0307 3 : V7, RMV0407, RQV0407
PM2	29	7 : MP0607, PMAX0, RMP0405, RMP2401, RMP2407, RLP2407, RQP2407 5 : MM0607, MN3, RMN0, RMN4, RMN7 4 : DT1907, T20, RMT2207, RMT1921 6 : MG2402, RMG2407, RQG2407, RG20, RDG0723, RMG1416 6 : DH0723, H20, RMH2407, RLH2407, RQH2407, RMH0307 1 : RMV0407
PM3	22	4 : MP0607, PMAX0, RMP2401, RMP2407 3 : MM0607, MN3, RMN2 3 : MT1518, T21, RMT1921 6 : MG2402, G21, RMG2407, RLG2407, RDG2319, RDG0723 5 : DH2319, DH0723, RMH2407, RLH2407, RMH0307 1 : RMV0407

Some variables contain more information than others in discrimination, but all are useful in this step. Some appear in the three subgroups, as last available pollution before the forecast (MP0607, MM0607), some residual effects for O3 (RMP2407) or for temperature (RMT1921), gradients (MG2402, RMG2407, RDG0723, DH0723, RMH2407, RLH2407, RMH0307) and wind (RMV0407). Some are specific to subgroups as $P_{\max 0}$ for PM2 and PM3.

To assign a mean-day in a class we use the classical Mahalanobis distance D^2; this distance is naturally the Euclidean distance in the space defined by discriminant factors:

$$D_t^2(\mathbf{x}) = (\mathbf{x} - \mathbf{m}_t)^T \mathbf{V}_t^{-1} (\mathbf{x} - \mathbf{m}_t), \tag{3}$$

where $\mathbf{x}$ is the vector containing the chosen predictors, $\mathbf{m}_t$ is the mean vector for class C_t with the same predictors and $\mathbf{V}_t$ is the within-class covariance matrix.

With the assumption of MultiNormality, the *a posteriori* probability of pertaining to C_t may be computed as

$$p_t(\mathbf{x}) = \exp\{-0.5 D_t^2(\mathbf{x})\} / \sum_{t=1,4} \exp\{-0.5 D_t^2(\mathbf{x})\}. \tag{4}$$

The classification results are furnished in three subtables (Table 8), with the numbers of mean-days for each class; note that these numbers are less than those indicated in Table 6, due to missing data in the predictors (the DFA may be performed only on complete data).Diagonal elements of these three tables correspond to well classified mean-days, the lower parts show *omitted assignments* (the mean-day is assigned to a lower class), and the upper parts – *false assignments* (the mean-day is assigned to an upper class).

Table 8. Misclassification results for the 3 subgroups (assignment by resubstitution)

Subgroups	PM1					PM2					PM3				
Assigned class:	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4	Total
Observed class															
1	12	3	1	0	16	13	1	0	0	14	28	1	0	2	31
2	5	47	10	4	66	12	64	18	8	102	15	109	9	34	167
3	5	20	87	32	144	5	24	85	35	149	3	0	32	4	39
4	2	8	57	245	312	2	19	59	346	426	3	63	24	368	458
Total	24	78	155	281	538	32	108	162	389	691	49	173	65	408	695
Well assigned	391/538 = 72.7%					508/691 = 73.5%					537/695 = 77.3%				
False assignment	50/538 = 9.3%					62/691 = 9.0%					50/695 = 7.2%				
Omitted assignment	97/538 = 18.0%					121/691 = 17.5%					108/695 = 15.5%				

Different results are obtained by jackknifing, slightly worse: for PM1, the three percentages are respectively 65.2%, 13.4% and 21.4%. In fact, the percentages computed in Table 8 are of no great interest. It would be interesting to assign costs to each misclassification, as some errors are more serious than others. Our aim is to forecast an O_3 value, and an assignment to a class is not the final result; it is only a *step to choose a regression model.* Broadly speaking, when the assignment is not good, it is difficult to find the exact reasons; but we may argue that often we are on the borders of two classes.

3.1.2.3. How to forecast maximum O_3 pollutions

Now we are going to choose the model. As we have done until now, to choose a model we must select the best predictors (the regressors) in each class. This choice must lead to the smallest number of regressors because in the most important classes, more specifically in C_1, we have few mean-days at our disposal and they are the most important for control strategy. The basic idea is to:

(1) choose a model for each mean-day "j" in each class C_t in a subgroup, by a stepwise regression (STEPWISE procedure in SAS), selecting optimal subsets of independent regressors in a multiple regression analysis by maximum R_2 improvement, with a model such as:

$$Y_j = \mu + a_1 X_{1j} + a_2 X_{2j} + \varepsilon_j,\ j = 1, ..., n, \tag{5}$$

where Y_j is the maximal pollution for mean-day "j", X_{1j} and X_{2j} are the selected regressors in the total set $\mathbf{x}$ and n is the number of mean-days for the class

(2) make a station adjustment by the model

$$Y_{jh} = \mu_h + a_{1h} X_{1jh} + a_{2h} X_{2jh} + \varepsilon_{jh},\ j = 1, ..., n; h = 1, ..., H, \tag{6}$$

where h is the index for station and $H = 3$ for PM1 and PM2, and 2 for PM3.

Having estimated a_{1h} and a_{2h}, it is possible to obtain an estimated value for Y_{jh}. The different predictors selected in each subgroup class are given in Table 9. The last lines of this table are good indicators of differences between stations in subgroups: quite different R^2 values mean that station effects are important. The residual standard deviations are around 20; this value is similar to the sensor error.

The major problem here is due to the selection of predictors. In this situation, to select them we are sometimes obliged to come back to the original ones; the information contained in residuals may not be sufficient when only a subset of predictors is selected. So, the selection is made on both, original d residual. Even if the quality of models is not generally good, even if all predictors are not used, we must not forget that these regressions follow the discrimination step in which more predictors were used. These regressions are only local ones.

Table 9. Selected predictors in models; R^2: determination coefficient, s_R: residual standard deviation, n: sample size; * : residual is used; @ : original predictor is used.

Subgroups	PM1				PM2				PM3			
Class:	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
Predictors												
Ozone												
PMAX0								*		*		*
RMP0405						*			*			
RMP2401			@	@								
RMP2407										@		@
RQP2407							*					
RCP2407				*	@							
Nitrogen												
MM0607							*		*			
RMN2				@				@				
RMN1	*										*	
RMN4							*					
Temperature												
MT1518		*								*		*
QT1421			*									
T20							*					
T21			*								*	
RMT2207				*				@	@	@		*
RMT1921		@				@						
Gradient 40m												
MG2402								*				
G21							*					
RG20		*		@								
RMG1416		*		*								
Gradient 100m												
H21			*									
RMH2407	@							*				@
RLH2407						*						
RQH2407					*					*		
RMH0307										@		
Wind velocity												
V8								*				
RMV0407						@	*				@	
RQV0407						*						
Statistics on mean day												
R^2	0.44	0.55	0.43	0.25	0.41	0.34	0.29	0.25	0.51	0.48	0.58	0.27
s_R	20	13	13	13	14	15	11	15	16	13	14	13
n	19	66	145	312	15	107	153	429	36	175	43	464
Statistics on mean day corrected for station effect												
R^2	0.24	0.33	0.25	0.24	0.35	0.27	0.24	0.26	0.37	0.35	0.48	0.26
s_R	32	20	19	16	20	21	16	18	21	19	19	14
n	51	175	385	835	46	259	333	907	70	321	82	718

3.1.3. Probability for a day to be assigned to an alarm class

The forecasts are only estimates and, as such, they are not exact values; this statement is linked with all statistical problems, even if we must come to a decision. But the decision is based on dubiousness. Some precision measures may be given: for the estimate $\hat{Y}_{jht}$, for each class C_t, we may estimate the standard deviation s_{jht}. Without weighting, we assume that the true, and unknown, pollution value y_{jh} (computed in C_t for station h) belongs to a confidence interval defined as

$$y_{jh} \in [\hat{Y}_{jht} - t_{a/2;n-p} s_{jht}, \hat{Y}_{jht} + t_{a/2;n-p} s_{jht}], \tag{7}$$

where $t_{a/2;n-p}$ is the Student t critical value with $n-p$ degrees of freedom at the significance level α (n being the number of observations in C_t, p the number of regressors used in the forecast model). We may also estimate the probabilities for an observed value to be greater than the different AIRPARIF levels (130 or 180): $Pr(Y_{jh} > 130)$ and $Pr(Y_{jh} > 180)$. If we use the weighting model, as it will be defined in 4.1 for forecasting, we obtain an estimate and a weighted standard deviation

$$\hat{Y}_{jh} = \sum_t p_t \hat{Y}_{jht}, \quad \sum_t p_t = 1, \tag{8}$$

$$s_{jh} = [\sum_t p_t s_{jht}^2]^{\frac{1}{2}}, \tag{8'}$$

and we may obtain a confidence interval similar to (7). But the knowledge of these 8 intervals is not directly linked to the probability of pertaining to an AIRPARIF alarm class. The study is in progress.

3.1.4. Application to a new day

Having done the analysis described so far, we are able to forecast maximum O_3 pollution for a new day. At 8am, we must have all the predictors for the eight stations (pollutant for each station, meteorological values for Saclay station). The strategy is indicated in Table 10.

For 1996, the results will be presented in Section 4.

Table 10. Forecasting strategy for a new day

1. For each station	1a. Assign, using the predictors, a probability $p_t(\mathbf{x})$ of pertaining to a pollution class C_t, for each station. 1b. Estimate, by weighting or not, O_3 value.
2. Associate the 8 estimations	Decide in which AIRPARIF alarm level the day is assigned. The probabilities are, until now, used to decide if the day is close to the border of two classes.

3.2. Non-parametric method

3.2.1. Introduction

As a complementary point of view to the linear case we develop a non linear approach which introduces new possibilities of modelling and is easy to interpret.

The non-linear method we use is a kernel nonparametric forecasting method with exogenous variables. This method is a very general one and can be used in many situations. It can be interpreted heuristically by means of similarity of days in terms of favourable climatic conditions in producing pollution. This will permit to analyze the origins of bad predictions and to detect atypical days. The main drawback of this method is that it needs a large data base, and in our case high pollution peaks are not very frequent, the history of the data base being too short. However, in this paper we confine ourselves to this method [an approach using the additive separable models (Hastie, 1986) in some sense intermediate between linear and non-linear approaches, is now being developed].

3.2.2. Nonparametric model with exogenous variables

3.2.2.1. Non-linear model

The model is a *non linear autoregressive* one of *Markovian type*, with exogenous variables. Exogenous variables are defined as variables that influence the endogenous variable but are not themselves influenced by the endogenous variable. Let $(X_n)_{n\in N}$, $X_n \in R$ be the process we intent to forecast and $(Z_n)_{n\in N}$, $Z_n \in R^p$ the process of exogenous variables. The model is written in the form

$$\begin{cases} X_{n+1} = F(X_n, Z_n) + \varepsilon_n^1 \\ Z_{n+1} = G(Z_n) + \varepsilon_n^2 \end{cases}, \tag{9}$$

with (ε_n^1), (ε_n^2) being independent white noises which are also independent of X and Z.

We are interested in predicting X_{n+1} using the history $\{(X_k, Z_k), k \leq n\}$. Usually, the best predictor in a quadratic sense is the conditional expectation. For the Markovian structure the predictor is provided by the autoregressive function:

$$E(X_{n+1} \mid (X_n, Z_n) = (x, z)) = F(x, z) \text{ for } x \in R \text{ and } z \in R^p. \tag{10}$$

3.2.2.2. Estimation

Let K be a $p+1$-dimensional kernel, (h_n) a real positive decreasing sequence, named *window* or *bandwidth*, and

$$K_{h_n}(u) = \frac{1}{h_n^{p+1}} K(\frac{u}{h_n}),\ u \in R^{p+1}.$$

We estimate F by the Nadaraya-Watson kernel estimator of the autoregression function (assuming $0/0 = 0$):

$$\hat{F}(x,z) = \frac{\Sigma_{t=0}^{n} K_{h_n}((x,z) - (X_t, Z_t))X_{t+1}}{\Sigma_{t=0}^{n} K_{h_n}((x,z) - (X_t, Z_t))}. \tag{11}$$

Typical kernels are the Epanechnikov and the Gaussian kernels.

The kernel predictor is:

$$\hat{X}_{n+1|n} = \hat{F}(X_n, Z_n). \tag{12}$$

The prediction interval is $[q_{\text{inf}}; q_{\text{sup}}]$ defined by the two empirical quantiles

$$P(\hat{X}_{n+1|n} < q_{\text{inf}}) = 0.1 \text{ and } P(\hat{X}_{n+1|n} < q_{\text{sup}}) = 0.9, \tag{13}$$

that is to say

$$P(q_{\text{inf}} \leq \hat{X}_{n+1|n} \leq q_{\text{sup}}) = 0.8. \tag{14}$$

3.2.2.3. Convergence

The literature on convergence of estimators (11) and (12) in the autoregressive case

$$X_{n+1} = F(X_n, X_{n-1}, ..., X_{n-r}) + \varepsilon_n$$

is extensive. In particular, results of almost sure pointwise convergence and results of joint asymptotic normality of the estimated regression at a finite number of distinct points are available. With nice assumptions on F and ε_n we can refer to Duflo (1990) and Yakowitz (1989). If moreover the sequence is mixing we can, for example, quote Ango Nze and Portier (1994), Ango Nze and Doukhan (1993), Bosq and Lecoutre (1992), Robinson (1983), Truong and Stone (1992) for the almost sure convergence, and Schuster (1972), Roussas and Tran (1992) for the convergence in distribution.

With exogenous variables there are very few results of convergence.

3.2.2.4. Interpretation

The estimator $\hat{F}$ defined by (11) can be reformulated as

$$\hat{F}(x_n, z_n) = \sum_{t=0}^{n-1} \omega_{n,t} x_{t+1}, \tag{15}$$

where $\omega_{n,t}$ is defined by

$$\omega_{n,t} = \frac{K_{h_n}((x,z) - (x_t, z_t))}{\sum_{t=0}^{n} K_{h_n}((x,z) - (x_t, z_t))}. \tag{16}$$

This means that the forecast value $x_{n+1|n}$ is the barycentre of the realizations associated to the coefficients $\omega_{n,t}$.

Let's choose $\tilde{K}$, a one-dimensional positive kernel decreasing on R^+, D, a distance on R^{p+1}, and set

$$K(u-v) = \tilde{K}(D(u,v)). \tag{17}$$

Then the coefficients $\omega_{n,t}$ can be interpreted as *similarity indices* between the vectors (x_t, z_t) and (x_n, z_n), the former designing the process at time t, and the latter the present of the process. Thus, because of the kernel properties, the greater the distance between the past and present vectors, the smaller the value of the kernel. Consequently,

– in order to obtain a high similarity index between a past instant t and the present instant n, the similarity must be strong for all the variables;

– if the distance between (x_t, z_t) and (x_n, z_n) is great, the kernel value will be insignificant and will set the similarity index almost to zero.

The value of the window parameter h is highly related to the choice of the distance D. If the point (X_n, Z_n) is isolated in the space R^{p+1}, h has to be large enough to ensure that there are some points in a ball of size related to h centered at the point (X_n, Z_n). But if h is too large the prediction will loose accuracy for the points which are not isolated. To avoid this drawback, we will use an alternative method, that is, a nearest-neighbour method: we will choose a number L and for each point compute the bandwidth h so that exactly L points are in the support of the kernel.

This computing choice adapts the kernel locally to the density: if a great concentration of points occurs, then h is small. The density of similar points will be very high around the point representing present n. If the points are sparse, h will be large, but in fact, the same number of points will interfere in the prediction.

The main idea arising from this interpretation is that "the same circumstances lead to the same futures".

3.2.3. O_3 air pollution

3.2.3.1. The choices

In the context of forecasting O_3 air-pollution, this idea is expressed under the following terms: the days with the same precursor emissions and the same meteorological conditions will be followed by the same maximum O_3 air pollution.

The variables we use are:

– X_t, the O_3 concentration daily maximum, denoted by PMAX0 beforehand,

– Z_t, defined by the precursor emissions (nitrogen oxides) and the meteorological conditions (temperature, gradient and wind velocity).

After studying the relationship between the different components occurring in the process of O_3 production (see for example Toupance, 1988), and after numerous attempts with many variables, we retain the following ones:

– T_n, the maximum temperature of the day n,

– W_n, the mean wind velocity at 58m during the afternoon of the day n,

– G_n, the sum of the positive values of the gradient at 40m in the night between the day $n-1$ and the day n,

– MN_n, the value of the nitrogen oxides molar sum, at 3am on the day n.

In order to compute the similarity indices $\omega_{n,t}$ we have to choose a kernel $\tilde{K}$, a number L of nearest-neighbours and a distance. The kernel $\tilde{K}$ is the Gaussian kernel, the distance $D_{n,t}$ between days t and n is a weighted Euclidean distance :

$$D_{n,t} = \sqrt{\sum_{i=1}^{p+1} \alpha_i (\zeta_i^n - \zeta_i^t)^2}, \quad \zeta^n, \zeta^t \in R^{p+1}. \tag{18}$$

In order to determine the number of nearest neighbours and the α_i coefficients, an optimization is conducted over a set of specific days, named N, which contains several typically polluted days. Let MAE be the mean absolute error defined by

$$MAE_{(\alpha,L)} = \frac{1}{\#(N)} \sum_{n \in N} \mid \hat{X}_{n+1|n} - X_{n+1} \mid . \tag{19}$$

Each variable has its own scale and in order to indicate which one has a strong effect (high α_i coefficient) it is useful to rescale the variables by dividing them by their one-hour standard deviations.

To obtain the optimum values of the coefficients and the bandwidth, we search on a grid of values of α and $L(\alpha_{\text{opt}}, L_{\text{opt}}) = \arg\min(MAE_{\alpha,L})$. Finally, the standard deviations are incorporated into the coefficients.

3.2.3.2. Results and comments

Optimization gives:

– $L = 10$. Actually, in most cases the number of significant (with high similarity index) days is only 2 or 3; this is due to the fact that there are few polluted days in the data base and the optimization was driven on those days.

– $\alpha_1 = 1$ (for O_3 maximum on the previous day, PMAX0). High value of this variable indicates favourable climatic conditions and its coefficient is significant.

– $\alpha_2 = 5$ (for T_n). The value of the coefficient is high so temperature seems to be a very important variable. It is not surprising because of the well known relationship to the air pollution formation during summer.

– $\alpha_3 = 1$ (for W_n). If the wind velocity is high the polluted air is dispersed, which explains the presence of this variable.

– $\alpha_4 = 0$ (for G_n). This variable is perhaps not convenient to measure the mixing layer height. In addition, it may be too large in Saclay compared to the Paris area.

– $\alpha_5 = 0$ (for MN_n). There is no influence of this variable.

3.2.3.3. Forecasting

The predictions are made for each air quality measurement site, with the same meteorological data for all the sites. On the morning of day n, the previous day maximum O_3 concentration in each site is available, but the variables T_n and W_n are unknown and must be replaced by predictions. These predictions can either be the National Meteorology Office ones (available the day before the evening) or predictions computed by the nonparametric forecasting method using the available data. The models we use are the following:

Temperature forecasting:

$$\hat{T}_n = f_1(t_1^n, ..., t_r^n, G_n) + \eta_n \tag{20}$$

with t_1^n denoting the temperature at 10am of the previous day, ... , t_r^n – the temperature at 6am of the day ($r = 21$). The kernel is the Gaussian one, the distance $D_{n,t} = \sqrt{\sum_{i=1}^{r} \beta(t_i^n - t_i^t)^2 + \gamma(G_n - G_t)^2}$ and the optimization process gives $L = 10$, $\beta = 3$, $\gamma = 11$.

Wind velocity forecasting:

$$W_n = f_2(\omega_n) + \nu_n, \quad \text{where} \quad \nu_n = \frac{1}{4}\sum_{i=1}^{4} \omega_i^n, \tag{21}$$

with ω_1^n denoting wind velocity at 3am of the day n, ... , ω_4^n – wind velocity at 6am of the day n.

The kernel is the Gaussian one, the distance $D_{n,t} = \sqrt{(\omega_n - \omega_t)^2}$ and the optimization process gives $L = 50$.

4. Results and discussion

Computations by the two methods have been done for two sets of data: summers 94 and 95 and summer 96. For the first period the models are fitted with the knowledge of all the data until 30th September 1995; for the second period forecasts are computed in real situation, that is, the data known are those until the day before the one forecasted. It is known that summers 94-95 were very hot and more polluted than previous years, summer 96 was less hot, more windy and less polluted: none level 2 alarm was detected.

4.1. Linear models

An important problem occurs when a mean-day is at the border of two classes; each model may give different results. The simplest solution is to estimate Y_{jh} for each class, giving $\hat{Y}_{jht}$, and to weight these values by a posteriori probabilities $p_t(\mathbf{x})$. This

Table 11. Results in assignment following AIRPARIF alarm rules depending on weighting or not by *a posteriori* probabilities

	Without weighting				With weighting			
Predicted alarm: Observed alarm	0	1	2	Total	0	1	2	Total
0	173	12	**4**	189	172	16	**1**	189
1	6	19	**2**	27	7	19	**1**	27
2	0	*4*	4	8	0	*6*	2	8
Total	179	35	10	224	179	41	4	224

was applied to 1994-95 data, where all stations furnished observations; the results are given in Table 11.

From Table 11, it appears that there is a tendency to increase the number of false alarms (level 2) without weighting ($6 = 4 + 2$ compared to $2 = 1 + 1$) and, on the contrary, to increase the number of slightly missed alarms with weighting (6 compared to 4). By scrutiny of results it was noticed that false alarms were due to an over-estimation for a subset of stations, whereas the others were correctly estimated. A look at the observed data shows that when a station is going beyond the alarm level, a subset of others is not far from this level. So we may argue that some constraints exist between stations: a high value for one station is not compatible with too low values for others. Henceforth, a strategy may be to choose for the basic estimation the model without weighting. If a station presents too high estimates with respect to the other two (PM1 and PM2), we may choose the model with weighting. The rules found after some trials are presented in Table 12.

Using these rules, the assignments to AIRPARIF alarm rules are given in Table 13a. To take account of the continuity of the estimation scale, we introduce a new alarm level between 0 and 1 (coded 0.5) defined as: "at least two stations have an observed maximum greater than $110\mu g/cm^3$", and we obtain the results in Table 13b.

From the 13 days with no observed alarm, 10 may be considered as being at the limit of alarm 1. This may justify a *softer approach*, by introducing a probability of pertaining to an alarm class defined by AIRPARIF.

We may note that such a modelling is not an object *per se*. If the environmental object appears clear, the relevant methodology to reach it is strongly influenced by the past practice. Our modelling would surely not have been strictly the same if we had been free to choose our procedure for defining alarm rules. The rule "*at least 2 stations with O_3 concentrations exceeding a fixed level*" has an evident advantage: its simplicity. But we are not sure that it is the best rule. It is always difficult, when one is facing such a problem, to propose the best solution. The European rule is to use a moving average over 8 hours, which seems inapplicable in forecasting situation.

Table 12. Rules to choose weighting or not in estimation; m_g is the maximum of means in subgroup PM_g $(g = 1, 2, 3)$

if	$\{m_1 \in [130,150]$ *and* $\{m_2 < 110\}$ *and* $\{m_3 < 90\}\}$
	or
	$\{m_1 \in [150,180]$ *and* $\{m_2 < 125\}$ *and* $\{m_3 < 135\}\}$
	or
	$\{m_1 > 180\}$ *and* $\{m_2 < 165\}$ *and* $\{m_3 < 80\}$
then	choose weighting model for PM1
if	$\{m_2 \in [130,150]$ and $\{m_1 < 100\}$ and $\{m_3 < 95\}\}$
	or
	$\{m_2 \in [150,180]$ and $\{m_1 < 120\}$ and $\{m_3 < 100\}\}$
	or
	$\{m_2 > 180\}$ and $\{m_1 < 130\}$ and $\{m_3 < 150\}$
then	choose weighting model for PM2
if	$\{m_3 \in [130,150]$ and $\{m_1 < 100\}$ and $\{m_2 < 105\}\}$
	or
	$\{m_3 \in [150,180]$ and $\{m_1 < 130\}$ and $\{m_2 < 130\}\}$
	or
	$\{m_3 > 180\}$ and $\{m_1 < 160\}$ and $\{m_2 < 160\}$
then	choose weighting model for PM3

Table 13. Results of assignment following AIRPARIF alarm rules depending on rules defined in Table 11

	a				b			
Predicted alarm: Observed alarm	0	1	2	Total	0	1	2	Total
0	175	13	1	189	168	3	1	172
0.5					7	10	0	17
1	6	20	1	27	6	20	1	27
2	0	4	4	8	0	4	4	8
Total	181	37	6	224	181	37	6	224

4.2. Non-parametric forecasting

First we compare our temperature and wind forecasts with the National Meteorology Office ones and with the forecasts based on the persistency assumption (the forecast for the day j is the realization on the day $j - 1$), on different sets of days:

- A : all days of summers 94 and 95,
- T : days of the same period with temperature variation greater than 5 degrees,
- W : days of the same period with variation of wind greater than 2 m/s.

Table 14. Comparison of temperature and wind velocity forecasts

	Mean absolute error for A	Mean absolute error for T
Temperature forecast		
Non parametric forecast	2.5	4.1
Meteorology forecast	1.7	2.6
Persistancy forecast	2.8	6.8
Wind velocity forecast		
Non parametric forecast	1.1	1.6
Meteorology forecast	1.0	1.5
Persistancy forecast	1.4	3.0

Results given in Table 14 show that our forecasts are better than the persistency ones (especially for the days which have large changes in comparison to the day before) and worse than the National Meteorology Office ones, but they are easier to use in an automatic procedure since they utilize only the information included in the data base.

Maximum ozone forecasting results with these two sets of meteorological forecasts give are summarized in Table 15. Without the Meteorological Office forecasting, ozone forecasts are underestimated. The principal reason for this is that the non-linear model is not well suited to forecast at very high temperatures: too few such points are in the data base.

Table 15. Results of assignment following AIRPARIF alarm rules depending on meteorological forecasting

	Without Meteo Office forecasts				With Meteo Office forecasts			
Predicted alarm: Observed alarm	0	1	2	Total	0	1	2	Total
0	176	12	0	188	162	17	0	179
1	20	18	0	38	14	17	7	38
2	4	5	0	9	2	5	2	9
Total	200	35	0	235	178	39	9	226

With the Meteorological Office forecasts there are too much level 1 alarms predicted at level 2, but no level 0 alarm is predicted at level 2 (Table 15). Level 2 alarms are not predicted as well as with the linear method.

4.3. Specific days

Now lets examine carefully some specific days to understand the differences between the two methods.

1 Aug 95 (Table 16A)

This day was very hot (33.1°C) and a bit windy (3,5m/s), the gradient in the night took high values, and the day before was also very hot. This is a typically polluted day and all the methods forecasted it well. In particular, the temperature and wind forecasts are well predicted.

7 Jul 95 (Table 16B)

The weather changed this day: the day before temperature was 22.8°C instead of 26.3°C, the gradient in the night was low. The models did not expected pollution in these circumstances and all the predictions are too low, missing the level 1 alarm.

Table 16A. August 1st 1995

Station	Realization	Linear without weighting	Linear with weighting	Non-linear	Non-linear with Meteo forecasts
1f92	159	159	159	171	129
13f75	162	156	156	175	152
71f75	162	156	156	145	149
1f93	162	169	168	162	187
1f94	133	166	163	151	153
4f91	163	168	166	128	142
1f77	137	142	141	139	105
4f75	136	130	129	153	150
Level alarm	1	1	1	1	1

Table 16B. July 7th 1995

Station	Realization	Linear without weighting	Linear with weighting	Non-linear	Non-linear with Meteo forecasts
1f92	108	79	79	71	95
13f75	132	79	81	62	86
71f75	109	79	79	102	90
1f93	107	116	106	105	107
1f94	136	114	105	88	125
4f91	161	123	113	104	112
1f77	89	83	78	72	125
4f75	104	71	67	?	?
Level alarm	1	0	0	0	0

4 Aug 94 (Table 16C)

The gradient in the night was high, and the day before was hot. Temperature on this day was very high (34.5°C) but badly predicted by the non linear model. The Meteorological Office forecast was better and so was the ozone forecast but it did not reach the level 2. Both linear methods predicted well the level 2.

13 Aug 95 (Table 16D)

The temperature was not very high (24.4°C), well predicted by the non-parametric method but badly predicted by the Meteorology Office. The wind velocity is large (5m/s) and in consequence the day was not polluted. The day before was hot (31.1°C) but the gradient in the night was not very high. There occurred a change of weather that neither the variables in the group PM2 of the linear methods or the Meteorological Office could anticipate, which led to overestimation of the ozone maximum on that day.

Table 16C. August 4th 1994

Station	Realization	Linear without weighting	Linear with weighting	Non-linear	Linear with Meteo forecasts
1f92	215	198	172	103	166
13f75	193	190	174	153	161
71f75	187	169	158	156	142
1f93	186	206	195	109	102
1f94	144	192	180	117	147
4f91	123	173	165	116	163
1f77	133	160	150	114	126
4f75	137	152	141	111	101
Level alarm	2	2	2	1	1

Table 16D. August 13th 1995

Station	Realization	Linear without weighting	Linear with weighting	Non-linear	Linear with Meteo forecasts
1f92	64	100	97	97	125
13f75	?	106	97	77	104
71f75	49	97	94	68	117
1f93	68	188	153	95	162
1f94	61	180	146	102	124
4f91	73	172	143	124	136
1f77	70	85	102	90	116
4f75	55	83	99	57	123
Level alarm	0	2	1	0	1

4.4. Discordance/concordance in forecasting for 1996

The two model families were applied to 1996 situations. This year seems to be atypical because the pollution was not high (no alarm of level more than 1 observed).

During the 107 days covering the period (from June 1st to September 15th) we have used 7 models:

– 5 linear models:

- MOD1: all significant variables without weighting,
- MOD2: all significant variables with weighting,
- MOD5: an intermediate model corresponding to the rules defined in Table 12,
- MOD3: a substitute to MOD1 (only temperature and O_3) without weighting,
- MOD4: a substitute to MOD2 (only temperature and O_3) with weighting,

– 2 nonparametric models :

- MOD6: the basic one,
- MOD7: same as MOD6, but using forecasting for T_n and W_n given by the National Meteorology Office.

The results are given in Tables 17a to 17c; "Obs" means observed alarms, "?" means that prediction was not possible due to the lack of a some variables, "s" is the mean square error for prediction. They suggest that

– for linear models weighting is always better; the substitute model is perhaps good enough, specially for a year as 1996, where false alarms were not detected,

– for nonparametric models meteorological forecasts are surely useful.

Table 17a. Results for models with all significant variables

		MOD1				MOD2				MOD5			
		Predicted alarms				Predicted alarms				Predicted alarms			
Obs	Total	0	1	2	?	0	1	2	?	0	1	2	?
0	90	61	17	3	9	65	15	1	9	61	18	1	9
1	17	3	11	1	2	3	11	1	2	4	10	1	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0
		s = 33.7				s =29.2				s =32.2			

Table17b. Results for substitute models

		MOD3				MOD4			
		Predicted alarms				Predicted alarms			
Obs	Total	0	1	2	?	0	1	2	?
0	90	66	10	5	9	70	11	0	9
1	17	7	8	0	2	6	9	0	2
2	0	0	0	0	0	0	0	0	0
		s = 35.1				s =27.7			

Table 17c. Results for nonparametric models

		MOD6 Predicted alarms				MOD7 Predicted alarms			
Obs	Total	0	1	2	?	0	1	2	?
0	90	76	5	0	9	69	7	0	14
1	17	10	5	0	2	7	9	0	1
2	0	0	0	0	0	0	0	0	0
		$s = 29.3$				$s = 28.3$			

5 Final remarks

Until now we have presented separate results for both approaches. From a mathematical point of view, linear methods and nonparametric ones use different statistical tools, and apparently may give, for high pollution level, some slightly different results: overestimation for the former, underestimation for the later. Nevertheless, differences are perhaps more formal than real : the fundamental idea underlying both is that it is impossible to have one model for all situations. If one existed, it would be a nonlinear model able to describe a great range of meteorological conditions for, say, predictors all varying on a specific continuous scale corresponding to the great variety of Ile-de-France climate. But it would be an intractable one!

So we must find a framework in which models can be simplified. We have written beforehand for the non parametric model: the same circumstances lead to the same futures. Fundamentally, it is the same for the linear approach: to isolate O_3 episodes pollution in order to find coherent classes, in which the assumption of linearity is reasonable. CART algorithms (Classification and Regression Tree, Breiman et al., 1984), often used in American Literature, have the same aim (California State Software, 1991; Burrows, 1991; Horie, 1987; Seinfeld, 1988; EPA, 1991; National Research Council, 1991).

It is clear that the approach we have developed is not yet entirely satisfactory; we have results which may be applied immediately, but improvements are necessary, specially by introducing statistical reasoning as an essential part of guidance in air pollution advisories. As always, *far more than the development and application of methods, statistical science is a way of thinking.*

REFERENCES

Andenberg M.R. (1973). *Cluster Analysis for Applications*, New York, Academic Press

Ango Nze P., Doukhan P. (1993). Functional estimation for time series: a general approach, (*to appear*). Preprint Orsay, 93-43 (France)

Ango Nze P., Portier B. (1994). Estimation of the density and the regression functions of an absolute regular stationary process. *Pub. Ins. Uni. Paris* **38**, 2, 59-87.

Bosq D., Lecoutre J-P. (1987). *Théorie de lanalyse fonctionnelle*, Economie et statistiques avancées, Economica, Paris.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and Regression Tree.* Wadsworth, Belmont, CA.

Burrows W.R. (1991). Objective guidance for 0-24 hour and 24-48 hour mesoscale forecasts of lake-effect snow using *CART*, *Wea. Forecasting* **6**, 357-378.

California State Software (1991). *CART*, Lafayette, California.

Clark T.L. (1982). Application of Prognostic Meteorological Variables to Forecasts of Daily Maximum One-Hour Ozone Concentrations in the Northeastern United States. *Journal of Applied Meteorology* **21**, 1662-1671

Duflo M. (1990). *Méthodes récursives aléatoires.* Masson, France.

Eder B.K, Davis J.M., Bloomfield P. (1994). An Automated Classification Scheme Designed to Better Elucidate the Dependence of Ozone on Meteorology *Journal of Applied Meteorology* **33**, 1182-1198.

EPA (1991). *Guidelines for Regulatory Application of the Urban Airshed Model.* EPA-450/4-91-013. U.S. Environmental Protection Agency, Office of Air Quality Policy and Standard, Research Triangle Park, North Carolina.

Granger C.W.J. (1969). Investigating causal relations by econometric models and cross spectral methods. *Econometrica* **37**, 424-438.

Györfi L., Härdle W., Sarda P. Vieu P. (1989). *Non parametric curve estimation from time series.* Lecture notes in Statistics, Springer-Verlag, Berlin.

Hastie T., Tibischriani R. (1986). Generalized additive models. *Statist. Science* **1** (1), 297-306.

Horie Y. (1987). *Episode Representativeness Study for the South Coast Air Basin*, Valley Research Corp. Prepared for the South Coast Air Quality Management District. El Monte, California.

National Research Council. (1991). *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, National Academy Press, Washington, D.C.

Poggi J.M. (1994). Prévision non paramétrique de la consommation électrique. *Revue de Statistiques Appliquées* **XLII** (4), 83-98.

Rhodes P., Miller-Gonzalez A. (1994). On predicting the Magnitudes of Ozone Concentrations and Occurrences of Ozone Episodes, Final Report, University of Texas Medical Branch at Galveston.

Robinson P.M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185-207.

Roussas G., Tran L.T. (1992). Asymptotic normality of the recursive kernel regression estimate under dependance conditions. *Annals of Statistics* **20**, 98-120.

Ryan W.F. (1995). Forecasting Severe Ozone Episodes in the Baltimore Metropolitan Area. *Atmospheric Environment* **29** (17), 2387-2398

SAS Institute Inc. (1985). *SAS users Guide: Statistics, Version 5 Edition.* Cary, NC.

Schuster E.F., (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *The Annals of Math. Stat.* **43**, 84-88.

Seinfeld J.H. (1988). Ozone air quality models: a critical review. *JAPCA* **38**, 616-645.

Tomassone R., Dervin C., Masson J-P. (1993). *Biométrie, modélisation de phénomènes biologiques*, Masson, Paris, 2ème éd..

Toupance G. (1988). *L'ozone dans la basse troposphère. Théorie et pratique.* Laboratoire de physico-chimie de lenvironnement, Université de Créteil.

Truong Y.K., Stone C.J. (1992). Nonparametric function estimation involving time series. *Annals of Statistics* **20**, 1, 77-97.

Vieu P. (1991). Smoothing techniques in time series analysis. In:*Non parametric functional estimation and related topics.* Roussas, G. (Ed) Kluwer Academic Publishers.

Wypij D., Sally Liu L.-J. (1994). *Prediction Models for Personal Ozone Exposure Assessment.* In: *Case Studies in Biometry.* Lange, N. And al. (Ed) Wiley, New York, 41-56.

Yakowitz S. (1989). Nonparametric density and regression. *J. Multivariate Anal.* **30**, 124-136.

Received 15 August 1998

O prognozowaniu zdarzeń ozonowych w rejonie Paryża

STRESZCZENIE

Artykuł prezentuje dwa podejścia do problemu wyjaśniania zależności stężenia ozonu od stężenia zanieczyszczeń pierwotnych (tlenków azotu) i od warunków meteorologicznych. Celem opracowanych metod jest umożliwienie, o godz. 8 rano, predykcji maksymalnej wartości ozonu, która wystąpi po południu, przy użyciu 6-letnich danych dotyczących zanieczyszczeń powietrza i danych meteorologicznych. Pierwsze z podejść, klasyczne, bazuje na zastosowaniu analizy skupień, analizy wariancji, analizy dyskryminacyjnej i regresji krokowej. Pozwoliło ono na identyfikację w rejonie Paryża trzech wewnętrznie jednorodnych grup stacji monitorowania. Wewnątrz tych grup, dzienne profile ozonu tworzą skupienia o malejącym poziomie; grupy są dobrze rozróżnialne na podstawie poprzednich wartości ozonu, stężeń zanieczyszczeń pierwotnych oraz warunków meteorologicznych. Podejście drugie oparte jest na metodach nieparametrycznych i używa jądrowego estymatora funkcji autoregresji ze zmiennymi zewnętrznymi. Prognoza jest ważoną sumą maksymalnych wartości ozonu obserwowanych w przeszłości. Metody są porównane na podstawie danych z roku 1996.

SŁOWA KLUCZOWE: zatrucie powietrza, stężenie ozonu, predykcja, model liniowy, nieparametryczne prognozowanie jądrowe.