Atmospheric Environment Vol. 26A, No. 5, pp. 759–764, 1992.
Printed in Great Britain.

0004-6981/92 $5.00+0.00
© 1992 Pergamon Press plc

A FOG FORECASTING METHOD IN A DEEPLY EMBANKED VALLEY

J. J. BOREUX* and J. GUIOT†

* Observatoire de Physique de Globe, Université Blaise Pascal, 12 Avenue des Landais, F-63000 Clermont-Ferrand, France and Fondation Universitaire Luxembourgeoise, Rue de Déportés, B-6700 Arlon, Belgium and † Laboratoire de Botanique Historique et Palynologie, Faculté de St-Jérôme, case 451, F-13397 Marseille cedex, France

(*First received* 25 *April* 1991 *and in final form* 10 *June* 1991)

Abstract—This paper presents a statistical model used to forecast fog in the Meuse Valley in Belgium. The method is a bootstrap discriminant analysis using eight predictors: river surface temperature, air pressure, air temperature at two elevations, wind speed and relative humidity at the same two locations. These data are measured from November 1989 to April 1990. Tests are done to determine the number of resampling needed for this data set and the optimum projection delay for prediction from the meteorological data. The best results are obtained for the prediction at 0700 h UT using meteorological data at 0400 h UT. The reliability of the model is given by a probability $\alpha=0.16$ of clear weather forecasting when it is foggy and a probability $\beta=0.26$ of fog forecasting when there is clear weather. These results are finally checked on 27 new observations in November 1990: the 6 foggy days are perfectly predicted and 24% of the clear days are badly predicted.

Key word index: Fog, Meuse Valley, discriminant analysis, forecasting, bootstrap.

1. INTRODUCTION

The nuisances caused by fog are numerous. Among these we shall mention the air and road traffic perturbations. Especially in the Meuse Valley (Belgium) every year fog causes serious road accidents.

For a few decades a lot of research work has been carried out to increase visibility by means of forced evaporation (Appelman and Coons, 1970) or artificial precipitation (Silverman and Kunkel, 1970) of droplets. It is generally agreed that the best results have been obtained with supercooled fogs (Serpolay, 1961) which are uncommon in the Meuse Valley (Boreux, 1988).

Another approach to reduce accident risks is to forecast fog formation a few hours before its appearance. As a matter of fact a method enabling its short-term forecasting would obviously be of valuable assistance to road as well as air traffic supervisors. So far no reliable method of forecasting fog a few hours in advance has been developed and tested with success.

Indeed the saturation of an air mass in contact with the ground is far more difficult to understand than the saturation of an air mass higher up in altitude. In addition to the usual factors air saturation is also influenced by landscape and nature of the terrain, vegetation and the presence or absence of a stream. Human activities can also play a leading part in fog formation by increasing the content of condensation nuclei (Boreux and Serpolay, 1990) and the water vapor in the atmosphere (Gorbinet and Serpolay, 1985). We can see that man–ground–atmosphere interactions considerably increase the complexity of the standard problem of air mass saturation–condensation, which extremely complicates any attempt at short-term fog forecasting by means of a semi-empirical model.

Consequently we have chosen to develop a statistical short-term fog forecasting model.

As fog occurrence is a discontinuous variable that meteorologists try to explain from continuous meteorological variables, the appropriate technique is the discriminant analysis which has been used for the snow-slide forecasting (Der Megreditchian *et al.*, 1975) and in meteorology (Murphy and Katz, 1985).

Since bootstrapping, introduced in statistic analysis by Efron (1979), is particularly efficient in determining the forecasting performance, it has been adapted to discriminant analysis.

2. DATA

Every 30 min a certain number of meteorological readings were taken in the Meuse Valley (altitude $z=82$ m) and over the surrounding plateau (altitude $z=200$ m):

for the screen air temperature :T_a^z
for air humidity :U_a^z
for the temperature of the soil surface :T_g^z
for the average speed and associated wind direction :V_m^z and D_m^z
for the maximum speed and associated wind direction :V_x^z and D_x^z

Two more parameters also interact in fog forecasting: the river surface temperature T_w and atmospheric pressure P_a. As their measurement is not automatic, they are unfortunately not available at the same frequency as the previous ones. Consequently we have been obliged to estimate them as being constant over 24 and 12 h periods, respectively.

As for observations on the dependent variable with regard to fog duration we do not have any automatic visibility measuring device and therefore generally do not know what time fog occurs because it is usually (during autumn and winter) a nocturnal phenomenon. On the other hand, the time when fog clears can be established with relative accuracy.

All variables make up the primary variables.

For fog forecasting two primary variables have been disregarded when two others have been combined to form a secondary variable:

- we excluded the temperature of the soil surface which is closely connected with the air temperature and the wind direction which is always badly defined when the air is very stable (i.e wind speed ≈ 0), that is to say during periods that are favorable to fog formation;
- over a 30 min period the wind speed changes and its maximum power (V_x) can be very different from its average value (V_m). As the dynamics of fluid flow around aerosol particles are important in fog formation and fog dispersion we make a combination between these values which we will call 'wind speed' calculated as follows: $WS=\ln(\alpha V_m+\beta V_x+\gamma+1)$.

As confirmed by Fig. 1 the wind is clearly responsible for fog only when its speed remains at very low values; after a certain threshold (here 0.7 m s^{-1}) there is no possibility of fog formation. The relationship between fog and wind speed is not linear and is consequently better represented by a log-transform which emphasizes low speeds. The natural logarithm of speeds added to 1 is used to avoid zero arguments.

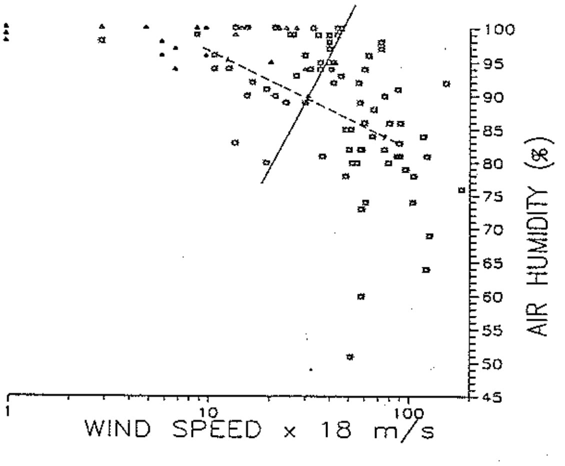

Fig. 1. Distribution of the foggy days (triangle) and clear days (small suns) as a function of the wind speed at 200 m elevation (in hm per half an hour or in m s^{-1} multiplied by 18) and as a function of the air relative humidity (in %). The wind speed axis is in logarithmic scale.

Finally eight explanatory variables have been retained: $TW=T_w$, $PRES=P_{air}$, $T82=T^{82}_{air}$, $WS82=WS^{82}$, $RH82=U^{82}_{air}$, $T200=T^{200}_{air}$, $WS200=WS^{200}$, $RH200=U^{200}_{air}$.

3. METHOD

Fog forecasting is a problem of two-group discrimination. The first group is defined by the occurrence of fog and the second one by the occurrence of clear weather. The discriminant analysis (due to Fisher, 1936) enables one to classify observations characterized by a set of meteorological variables called also predictors. The n observations on the p predictors give the data matrix $\mathbf{X}=(x_j)$ with $i=1$ to n and $j=1$ to p.

In fact we want to test the hypothesis of fog against the alternative hypothesis of clear weather. This requires one to compute a function to help make a decision. In factorial discriminant analysis this function is defined as a linear combination of the predictors. For any given observation (among the n observations which are available for the analysis) of the p predictors, we have:

$$U_i=\sum_{j=1}^{p} u_j(x_{ij}-\bar{x}_j), \qquad (1)$$

where $\bar{x}_j$ is the mean of the predictor j computed on n observations.

The computation of the coefficients u_j is based

- on the maximization of the distance between the gravity centers of the two groups, say $(\bar{x}_{11}, \bar{x}_{12} \ldots \bar{x}_{1p})$ and $(\bar{x}_{21}, \bar{x}_{22} \ldots \bar{x}_{2p})$, where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ are, respectively, the mean of the predictor j computed on the n' observations from class 1 and the $n-n'$ observations in class 2, and
- the minimization of the inertia of each group (groups as dense as possible).

We define $\mathbf{T}$ as the total covariance matrix between the p predictors (then calculated on the n observations) and $\mathbf{c}$ as a vector of p elements c_j proportional to the distance between the two gravity centers:

$$c_j=\sqrt{n'(n-n')/n^2}\,(\bar{x}_{1j}-\bar{x}_{2j}), \qquad (2)$$

where n' is the number of fog observations. It can be demonstrated that vector $\mathbf{u}=(u_1, \ldots, u_p)$ is given by {see for instance Lebart *et al.* (1982), Murphy and Katz (1985) or Dagnelie (1974)}:

$$\mathbf{u}=\mathbf{T}^{-1}\mathbf{c}. \qquad (3)$$

The discriminant power D or the Mahalanobis distance measures the capability of the m predictors to separate the two groups:

$$D=\mathbf{c}'\mathbf{T}^{-1}\mathbf{c}. \qquad (4)$$

An observation i will be classified in group 1 if $U_i>0$. We may refine the technique to ensure that the discriminant function forecasts most cases of fog occurrence because the risk is at its greatest when a fog is not predicted and to define a threshold S to ensure that at least 95% of the actual fog observations are in group 1. A fog forecast will be expressed for observation i if $U_i>S$. This can be compared to the first type error α in the hypothesis testing theory: $\alpha=$Prob (Fog|Clear)$=5\%$. The proportion of clear observations predicted as fog can be compared to the second type error $\beta=$Prob (Clear|Fog). It is expected to be higher than 5%, but it is related to a less important risk than α.

Figure 1 illustrates the method. It shows the distribution of foggy days and clear days according to wind speed in Loyers ($z=200$ m) and air humidity in the Meuse Valley ($z=82$ m). The criterion of 95% of accurate fog prediction means that the limit between the two groups should be set as shown in the figure: one fog occurrence appears among the 23 observations in group 2. The error α is then about 5%. The consequence is that in 31 cases out of 94 in group 1 the weather turns out to be clear. The error β is then about 33%, which reflect the existence of atmospheric conditions potentially favorable to fog formation in cases where the condensation nuclei are not sufficient to trigger the fog. If the two-dimensional representation is enlarged to eight dimensions β can be decreased.

Discriminant analysis can be completed when the discriminant function is computed. However, the use of that function in a probabilistic way is of prime importance in prediction. This may be solved using standard statistics, using specified probability functions. The introduction of bootstrapping by Efron (1979) enables us to be free from all assumptions. The idea is to resample the original observations in a suitable way to construct pseudo data sets from which the estimates are performed. The variability of these estimates from one pseudo data set to another is the key for their reliability appreciation.

The resampling is done by random extraction with replacement. A pseudo data set contains as many observations as the original one and consequently some observations are present several times and others are not. In discriminant analysis there are two ways of resampling the original data: either to resample inside each group so that the size of each group is fixed, or to resample in the whole data set so that the size of each group can vary. We have chosen the second approach, which is less conservative. For each pseudo data set we calibrate a discriminant analysis that is afterwards applied to the remaining observations. The use of the observations not randomly withdrawn provides the basis for an independent verification of the reliability of the predictions.

For each kth replication (from 1 to K) we extract n observations distributed in n'_k fog observations and $n-n'_k$ clear observations. The discriminant function is given by the vector $\mathbf{u}_k=(u_{1k}, \ldots, u_{mk})$, the discriminant power is denoted D_k. The threshold S_k is defined so that 95% of fog observations are well predicted. According to this discriminant function and this threshold clear observations from the pseudo data set assigned to the fog group are counted to provide β_{Dk} (D meaning dependent sample). The observations not randomly extracted from the pseudo data set are used to provide α_{Ik} and β_{Ik} (I means independent sample). These last two probabilities give a truer prediction efficiency.

When the K classifications are done we compute the means and the standard deviations of all these statistics over the K replications. This gives the vectors $\bar{\mathbf{u}}=(\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_m)$ and $\mathbf{Su}=(Su_1, Su_2, \ldots, Su_m)$, the mean discriminant power $\bar{D}$ with its standard deviation SD, the mean threshold $\bar{S}$ with its standard deviation SS. These average values are only used in the result interpretation while as for the fog forecasting it is better to take each equation and set up K predictions whose variability provide an appreciation of their reliability.

The error terms β_D, α_I and β_I are the means of these terms obtained from each replication.

4. RESULTS

The method is illustrated with the prediction of fog at 0700 h UT from meteorological data observed at 0400 h UT from November 1989 to April 1990. Calculations are based on 117 observations including 23 foggy days (data processing has been made by the program package PPPHALOS: Guiot, 1990).

Figure 2 shows the evolution of the probabilities β_D, α_I and β_I when they are averaged from the first replication to the first 500 ones. It clearly appears that these parameters have a great variability in the first 50 replications. They reach an equilibrium value after 300 replications. The risk of an unpredicted fog tends to 16% on the independent samples. This relative low value is compensated by a risk greater than 30% for a predicted fog when the weather has been clear (31% on the dependent data and 34% on the independent

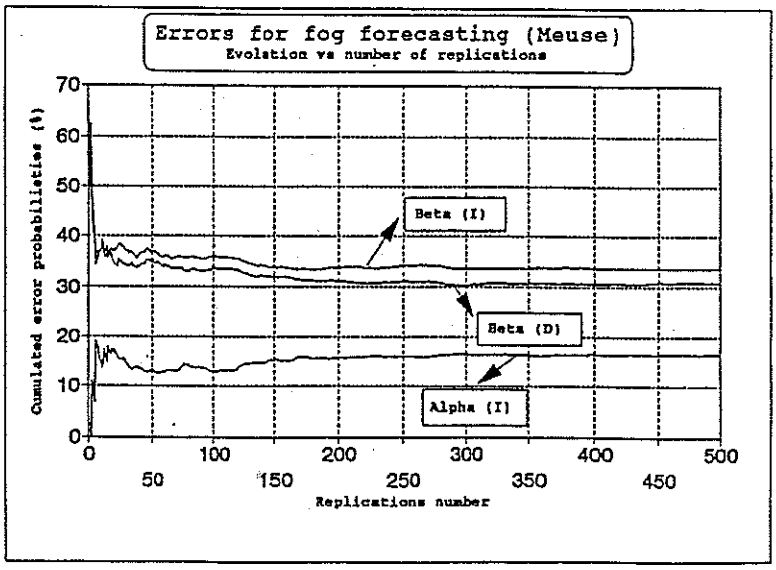

Fig. 2. Evolution of the error probabilities β_D, α_I and β_I as a function of the number of replications. The error probability at step k is the mean computed on the first k pseudo data sets.

ones). For the following analyses, we decided to use a value of 300 for K.

Short-range fog forecasts at 0700 UT are shown in Fig. 3 for prediction periods varying from 0 to 6 h. In fact these limits tend to represent upper limits. If meteorological data at 0400 h UT tell us that there will be fog at 0700 h, that means that the fog will appear between 0400 h and 0700 h (a period of less than 3 h). We could interpose that sometimes fog is previously formed at the time of fog prediction. This possibility is unlikely because the time of fog formation is usually within 2 h preceding sunrise and 30 min after sunrise (Pilie *et al.*, 1974). On the other hand forecasting clear weather at 0400 h UT, is truly 3h predictions. The best forecast is obtained for a 3 h projection. On the one hand this time-lag is sufficient

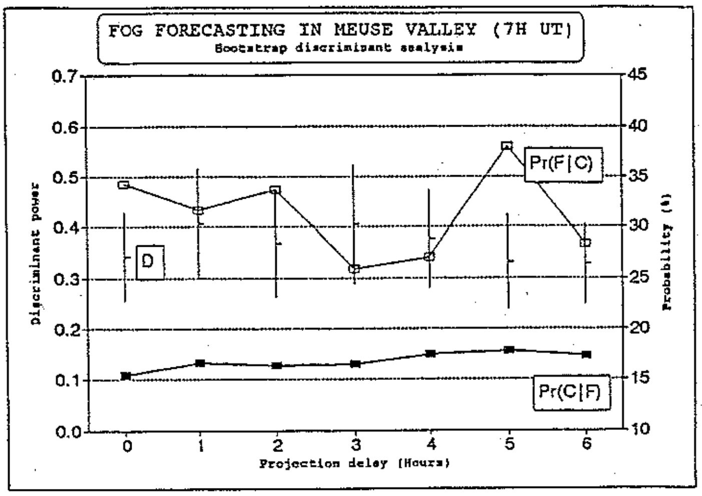

Fig. 3. Fog forecasting in Meuse Valley at 0700 h UT for different steps. Projection delay j means a forecasting of fog at 0700 h using meteorological data at $7-j$ h. The vertical bars represent the discriminant power D with ± the bootstrapped standard deviation. $Pr(F|C)$ is the error probability β_t computed on independent data (F is fog and C is clear) and $Pr(C|F)$ is the error probability α_t.

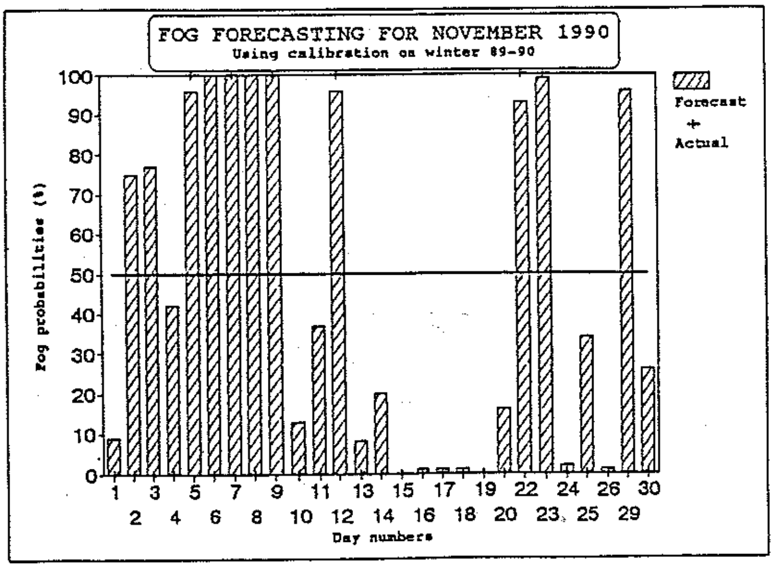

Fig. 4. The statistical short-term fog forecasting model applied on data in November 1990. The cross on the point of intersection between the axis '100' and the day concerned shows the days when fog has been observed while the cross at the bottom of the chart indicates clear days. The bar chart gives the corresponding probabilities computed by the model.

to allow the authority to decide on a suitable strategy, on the other hand the probabilities α_I and β_I are minimal and the discriminant power is maximal.

Table 1 shows the discriminant function for the eight meteorological variables. The significance of the coefficients must be appreciated on their distribution around their average value. For instance, for the river surface temperature T_w 263 coefficients on 300 are positive. Therefore, the coefficient 0.141 is significant at 88%. Accordingly wind speed has the most significant influence on fog formation. Nevertheless the lack of significance of the other coefficients does not mean that they have no effect, but that in the available data they do not provide any additional information according the other ones. Table 1 also shows that the fog appearance is well predicted in 84% of the cases $(1-\alpha_I)$ with a 26% risk of a mistaken prediction of clear weather (β_I). Since these values are based on independent data they are really reliable.

Table 1. Coefficients of the discriminant function, their bootstrap standard deviation and their confidence level, for the prediction at 0700 h UT from meteorological data at 0400 h UT (26 foggy days and 92 clear days). The abbreviations of the predictors are given in the text. *Coeff.* means coefficient of the discriminant function, *Std.dev.* means its bootstrap standard deviation and *Con. level* means its confidence level empirically computed on the 300 bootstrapped functions

	TW	*PRES*	*T82*	*WS82*	*RH82*	*T200*	*WS200*	*RH200*	*D*	*S*	β_D	α_I	β_I
Coeff.	0.141	−0.023	−0.001	−0.195	0.031	0.001	−0.414	0.065	0.406	0.175	24.4	16.5	25.9
Std. dev.	0.116	0.080	0.080	0.074	0.084	0.080	0.173	0.073	0.117	0.158			
Con. level	88%	57%	51%	99%	30%	53%	99%	84%					

5. VALIDATION AND PROSPECTS

The validation is done on data sampled 6 months later in November 1990 (Fig. 4). It contains 6 foggy days (represented by a cross on the point of intersection between the axis '100' and the day concerned) and 21 clear days. The 300 discriminant functions calculated for the 1989–1990 winter are applied to these 27 observations. We have decided to retain the forecasting given by more than 50% of the discriminant function. The 6 foggy days are then predicted with success. For the clear days the model fails for days 2, 3, 6, 8 and 29. In conclusion, the α-value of 16% calculated for 1989–1990 winter is confirmed by a value of 0% and the β-value of 26% is confirmed by a value of 24% in November 1990.

In conclusion, this statistical short-term fog forecasting model is reliable since we are able to forecast fog in the Meuse Valley with a good rate of success $(1-\alpha=84\%)$. Nevertheless, the β value does not suit us. As there is a close agreement between numbers of droplets in a fog and the condensation nuclei in atmosphere we will attempt to improve this model by adding a new predictor: the abundance of centers of water vapor condensation per unit volume.

Acknowledgements—This work has been financed by the Belgian ministry of the "Travaux Publics pour la Région Wallonne" and by a doctoral research fellowship (to J. J. Boreux) of the European Community Comission in the frame of the EPOCH program.

REFERENCES

Appleman H. S. and Coons F. G. (1970) The use of jet aircraft engines to dissipate warm fog. *J. appl. Met.* **9**, 464–467.

Boreux J. J. (1988) Etude des brouillards locaux denses dans la vallée de la Meuse. Rapport d'activité 1987–1988, FUL 1988.

Boreux J. J. and Serpolay R. (1990) Comparative CCN concentration measurements between two sites in the Meuse valley and occurrence of localized dense fogs in relation with industrial activities. *Proc. of the* 3rd *Int. Aerosol Conf.*, Vol. 1, pp. 515–518.

Dagnélie P. (1975) *Analyse Statistique à Plusieurs Variables.* Vander, Bruxelles.

Der Megreditchian G., Sinolecka C., Veysseire J. M., Boiret P., Yessaian A. and Belzane J. (1975) Approche statistique du problème d'évaluation des risques d'avalanches. *La Météorologie* **3**, 121–143.

Efron B. (1979) Bootstrap methods: another look at the jacknife. *Ann. Statist.* **7**, 1–26.

Fisher R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

Gorbinet G. and Serpolay R. (1985). Influence of an industrial site on the characteristics of CCN supersaturation spectra associated with air masses of different origin. *J. Recherches Atmos.* **19**, 193–202.

Guiot J. (1990) Methods of statistics in paleoclimatology. PNEDC report, Université d'Aix-Marseille 3.

Kunkel B. A. and Silverman B. A. (1970). A comparison of the warm fog clearing capabilities of some hygroscopic materials. *J. appl. Met.* **9**, 634–638.

Lebart L., Morineau A. and Fénelon J. P. (editors) (1982) *Traitement des Données Statistiques: Méthodes et Programmes.* Dunod, Paris.

Mahalanobis P. C. (1936) On the generalized distance in statistics. *Proc. Nat. Inst. Science-India* **12**, 49–55.

Murphy A. H. and Katz R. W. (editors) (1985) *Probability, Statistics and Decision Making in the Atmospheric Sciences.* Westview Press, Boulder, CO.

Pilié R. J., Mack E. J., Kocmond W. C., Rogers C. W. and Eadie W. J. (1975) The life cycle of valley fog. Part I: micrometeorological characteristics. *J. appl. Met.* **14**, 347–363.

Serpolay R. (1961) Nouveaux résultats d'ensemencements de brouillards surfondus à l'aide de pulvérisations de propane liquide (en collaboration avec P. D. Cot). *Comptes-rendus de l'Académie des Sciences* **253**, 171–174.