Technical report

Comment on « Representative volume emission profiles from rocket photometer data, by D. P. Murtagh *et al.* »

François THOUVENOT

Laboratoire de Géophysique Interne et Tectonophysique, I.R.I.G.M., Université de Grenoble, 38000 *Grenoble, France*

Received 12/11/84, revised 23/01/85, accepted 18/02/85.

Key words : rocket profiles, cross-validation cubic spline.

Annales Geophysicae 1986, **4,** A, 1, 87-94.

INTRODUCTION

In a recent paper, Murtagh *et al.* (1984) applied various techniques to derive volume emission profiles from rocket photometer results. One of the main problems is to preserve, in the smoothing and differentiation of the corrected flight data, any small-scale structure which might be aeronomically significant. Incremental straight line fitting proved most suitable in most cases but the altitude of peak emission can be ill-defined. For very noisy data, the Fourier filtering method gives a good indication of the height of peak emission but can lead to some distorsion of the layer profile. Finally, Murtagh *et al.* stressed that « cubic spline and polynomial fits should only be used in conjunction with some other method or when there are gaps or omissions in the data ».

It is the objective of this comment to discuss this assertion and to present results obtained with a smoothing using a cross-validation cubic spline. This technique seems adapted to difficult data sets and needs no apriorism to be introduced during the profile processing, which is not the case for conventional spline smoothings. The proposed method suffers from drawbacks which will be briefly analysed.

METHOD

Using cubic splines to smooth a data set usually proves perplexing. Denote indeed by (t_i, z_i) the n data points defined on interval $[a, b]$ with $a < t_1 < t_2 < \cdots < t_n < b$. Let $H^2[a, b]$ be the set of cubic splines defined on $[a, b]$ with nodes in each t_i and let α_i be the normalized weight affected to each data point with the condition

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i^2 = 1\,.$$

The smoothing amounts to seek the solution $\sigma_{n,\tau}$ of the problem

$$\underset{g \in H^2[a,b]}{\text{Minimize}} \times \left\{ \tau \int_a^b [g''(t)]^2 \, dt + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^2 \cdot [g(t_i) - z_i]^2 \right\}.$$

There remains to choose the smoothing parameter τ, i.e. the balance between the *smoothness* of $\sigma_{n,\tau}$ measured by the integral in the above equation and the *fidelity* to the data measured by the summation. In many situations this step is carried out hit or miss, with a trial-and-error approach which relies mainly on the observer's experience and on his own eye perception of what the smoothing should be like.

Wahba and Wold proposed in 1975 the so-called *cross-validation method*, grounded on original ideas by Allen (1974) and Stone (1974), to choose the smoothing parameter in certain cases. Several variations of this method were presented later on, but all of them suffered from an excessive computing time and from convergence problems when data points are too numerous (more than 100 points). Utreras (1979) succeeded in obtaining an algorithm which allows to get over these difficulties. He also formalized certain theoretical properties which had been surmised independently by Craven and Wahba (1979) and his notations will be used in the following.

Let $\sigma_{n,\tau,k}$ be the unique solution of the problem

$$\underset{g \in H^2[a,b]}{\text{Minimize}} \times \left\{ \tau \int_a^b [g''(t)]^2 \, dt + \frac{1}{n} \sum_{\substack{i=1 \\ i \neq k}}^{n} \alpha_i^2 [g(t_i) - z_i]^2 \right\}.$$

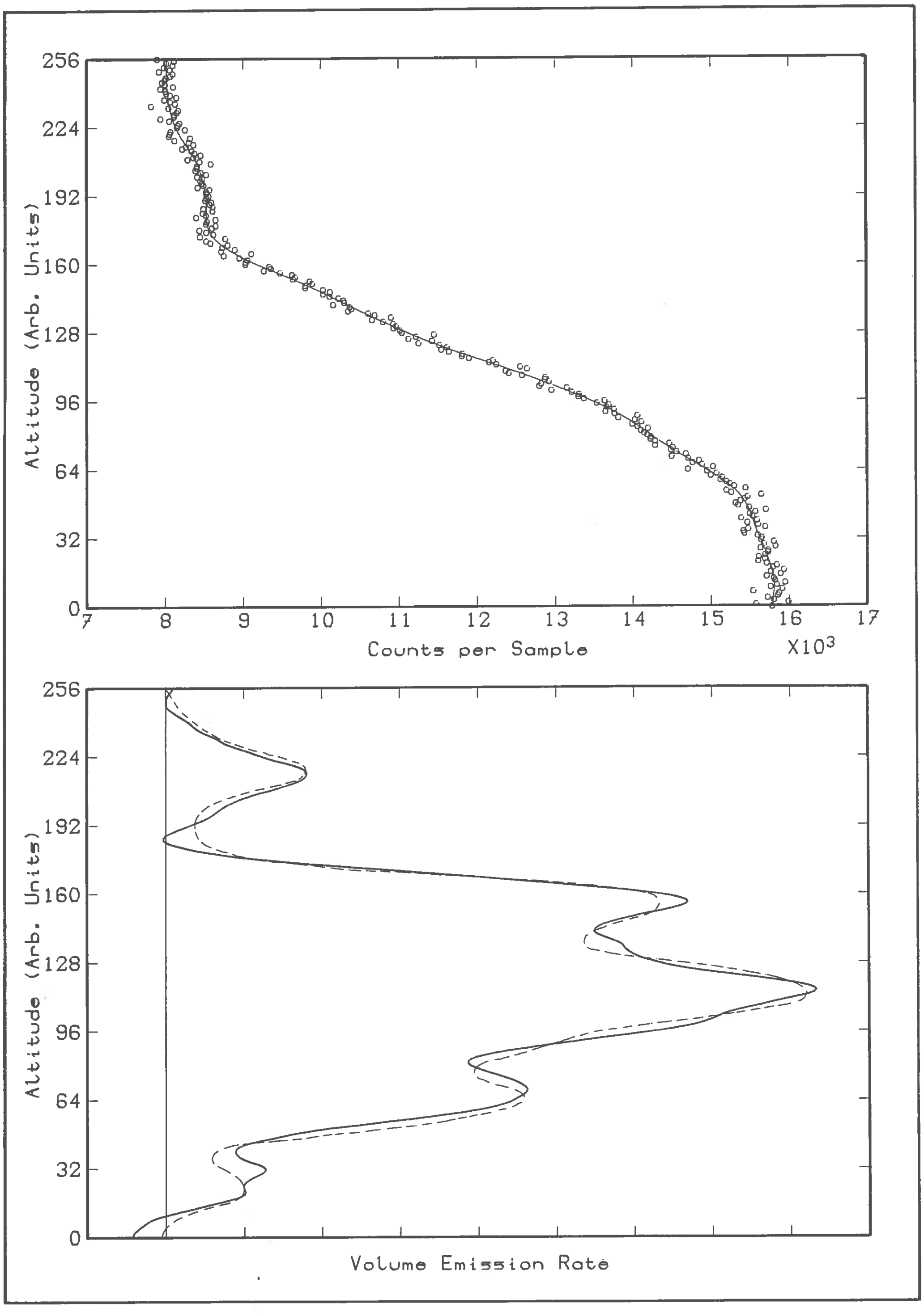

Figure 1
Test profile with 5 peaks. Integrated profile with noise added (upper part). Result of the cross-validation method (lower part, solid line), with the original profile shown in dashed line.

$\sigma_{n,\tau,k}$ is therefore the smoothing spline with parameter τ for the whole data set except data point (t_k, z_k). If the difference between z_k and $\sigma_{n,\tau,k}(t_k)$ is measured by $[z_k - \sigma_{n,\tau,k}(t_k)]^2$, τ could be sought as the value minimizing the quantity

$$V_0(\tau) = \frac{1}{n} \sum_{k=1}^{n} [z_k - \sigma_{n,\tau,k}(t_k)]^2 \, .$$

But it seems rather arbitrary to choose an equal weight for all deviations $[z_k - \sigma_{n,\tau,k}(t_k)]^2$. This can be explained first because data points do not necessarily contribute with the same weight in the computation of the spline function, and second because the estimation should be worse for edge points than for middle points.

To make clear the quantity $V(\tau)$ which should be minimized instead, further notations have to be introduced. Let $\mathbf{y}$ and $\mathbf{z}$ be the vectors

$$\mathbf{y} = (y_1, ..., y_n) \quad \text{with} \quad y_i = \sigma_{n,\tau}(t_i)$$
$$\mathbf{z} = (z_1, ..., z_n) \, ,$$

let $\mathbf{A}(\tau)$ be the matrix transforming $\mathbf{z}$ into $\mathbf{y}$ and denote by $\mathbf{D}$ the diagonal matrix with elements $d_{ii} = \alpha_i^2$. Utreras defines $V(\tau)$ as

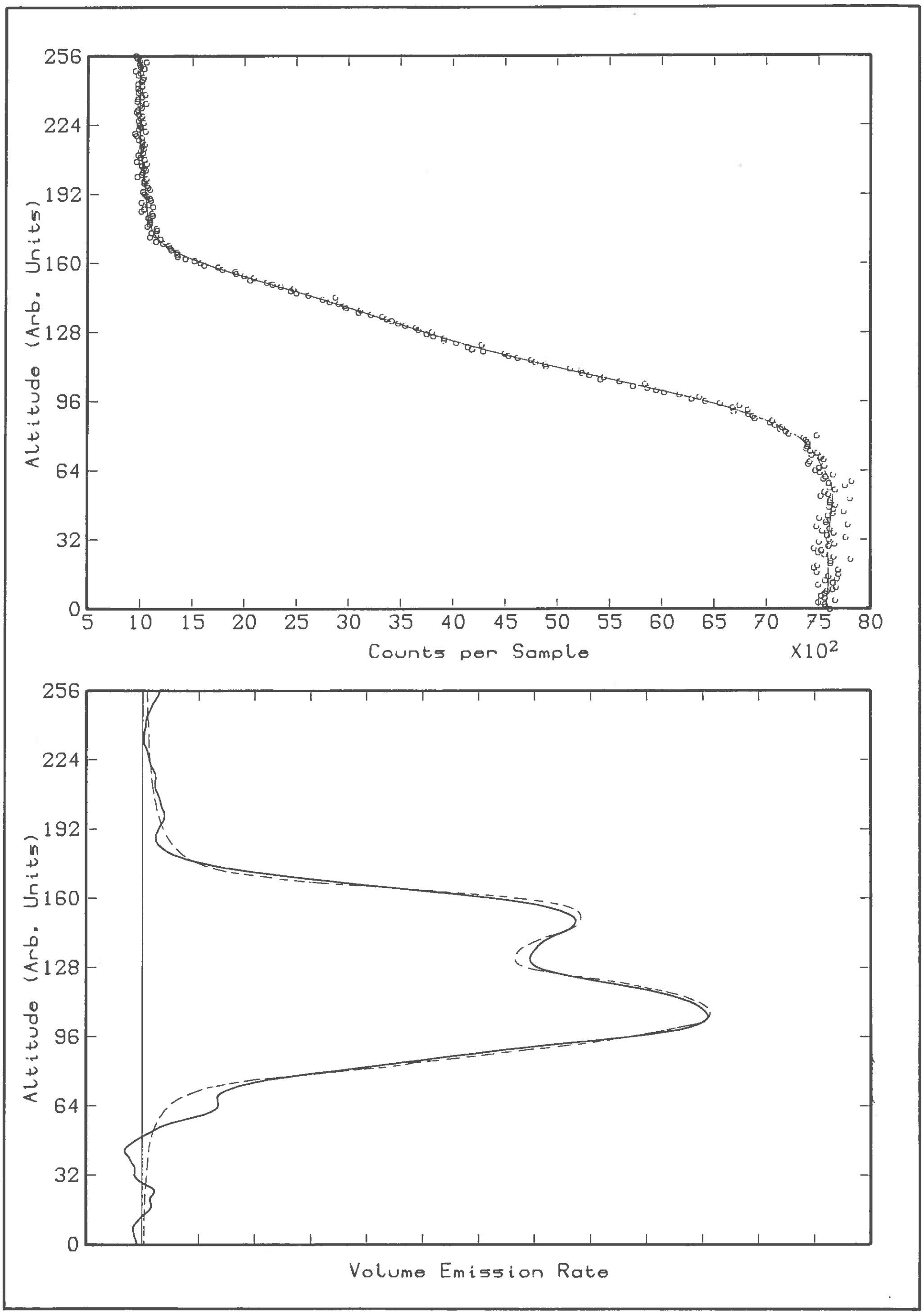

Figure 2
Test profile with 2 peaks. As for figure 1.

$$V(\tau) = \frac{1}{n} \frac{\| \mathbf{D}^{1/2}[\mathbf{A}(\tau) \cdot \mathbf{z} - \mathbf{z}] \|^2}{\left\{1 - \frac{1}{n}\operatorname{Tr}[\mathbf{A}(\tau)]\right\}^2}.$$

To minimize V on the set of positive real numbers (*generalized cross-validation*), it is a must to dispose of a quick algorithm for the computation of $A(\tau)$. If we consider the set of cubic splines which have the property to be a polynomial of the first degree on both intervals $[a, t_1]$ and $[t_n, b]$, the canonical basis of this vectorial space is composed of elements σ_i $i = 1, \dots, n$ such as

$$\sigma_i(t_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j . \end{cases}$$

Denote by $\mathbf{\Omega}$ the matrix with elements

$$\omega_{ij} = \int_a^b \sigma_i''(t) \cdot \sigma_j''(t) \, \mathrm{d}t .$$

It can be shown that eigenvalues β_{in} of $\mathbf{A}(\tau)$ are related to eigenvalues λ_{in} of $\mathbf{D}^{-1}\,\mathbf{\Omega}$ by

$$\beta_{in} = \frac{1}{1 + n\tau\lambda_{in}}.$$

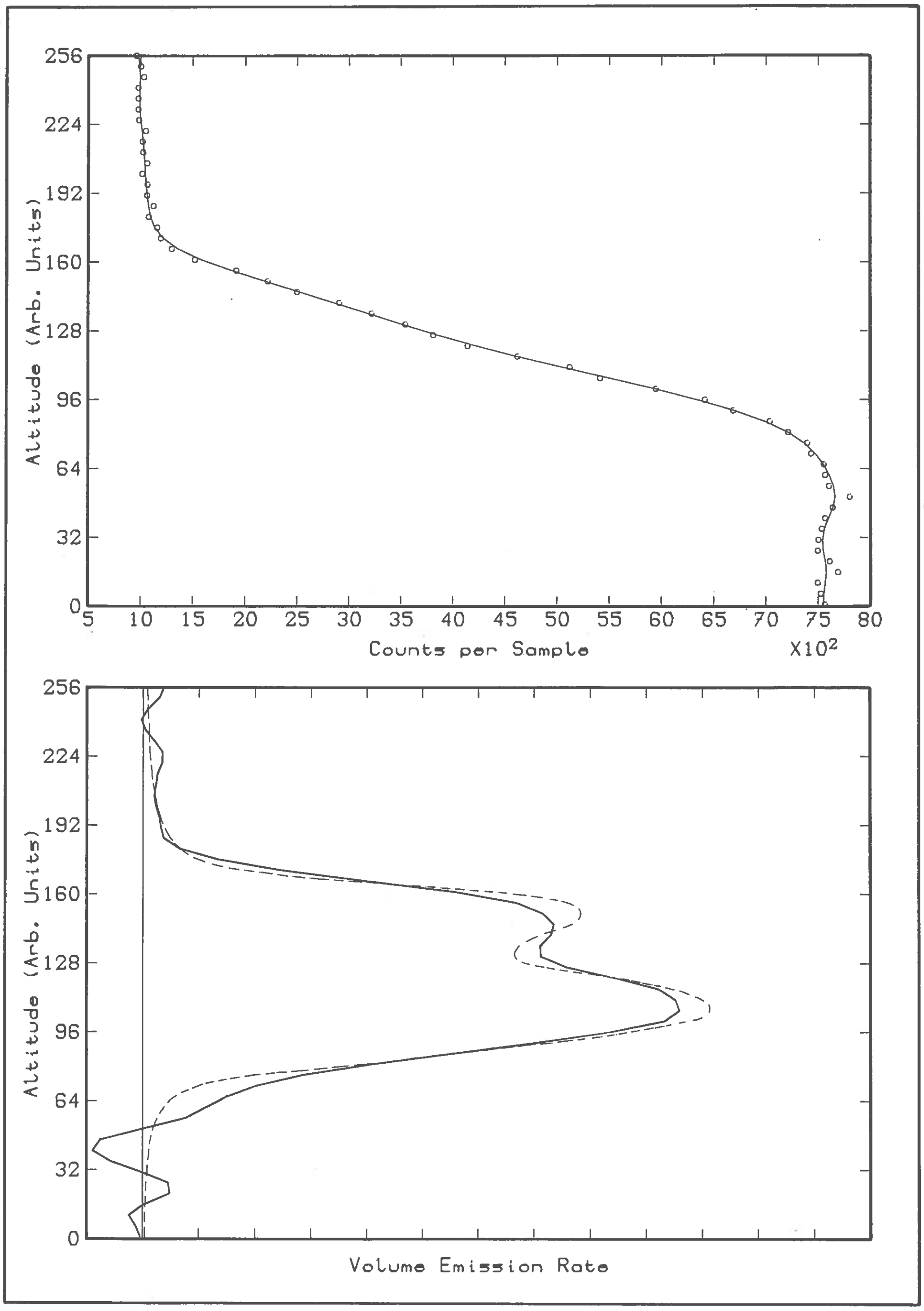

Figure 3

Test profile with 2 peaks and data reduced to 1/5th. As for figure 1.

The interest of the routines written by Utreras lies in the evaluation of λ_{in}. Compared to an exact computation which would require a computing time varying as n^3, Utreras' method keeps to n^2. Moreover, if the data set consists of equidistant points, the computing time varies as $2n$ only.

RESULTS

This smoothing technique involving cross-validation cubic splines was already successfully used in Solid Earth geophysics to define the amplitude decrease of seismic waves along deep seismic sounding profiles (Thouvenot, 1983). However, in the quoted study, the smoothing was not critical, as no differentiation of the fitting curve was necessary. Before applying the cross-validation technique to the data in Murtagh *et al.*, it was therefore useful to test it on synthetic data sets as Murtagh *et al.* did in their paper.

We first chose to test the capability of the method to extract stratified layers from data. A profile with 5 peaks was generated, then integrated, and random Poissonian noise was added (fig. 1). Count rates are similar to the ones adopted by Murtagh *et al.* in their synthetic data sets. The noisy data are smoothed by a

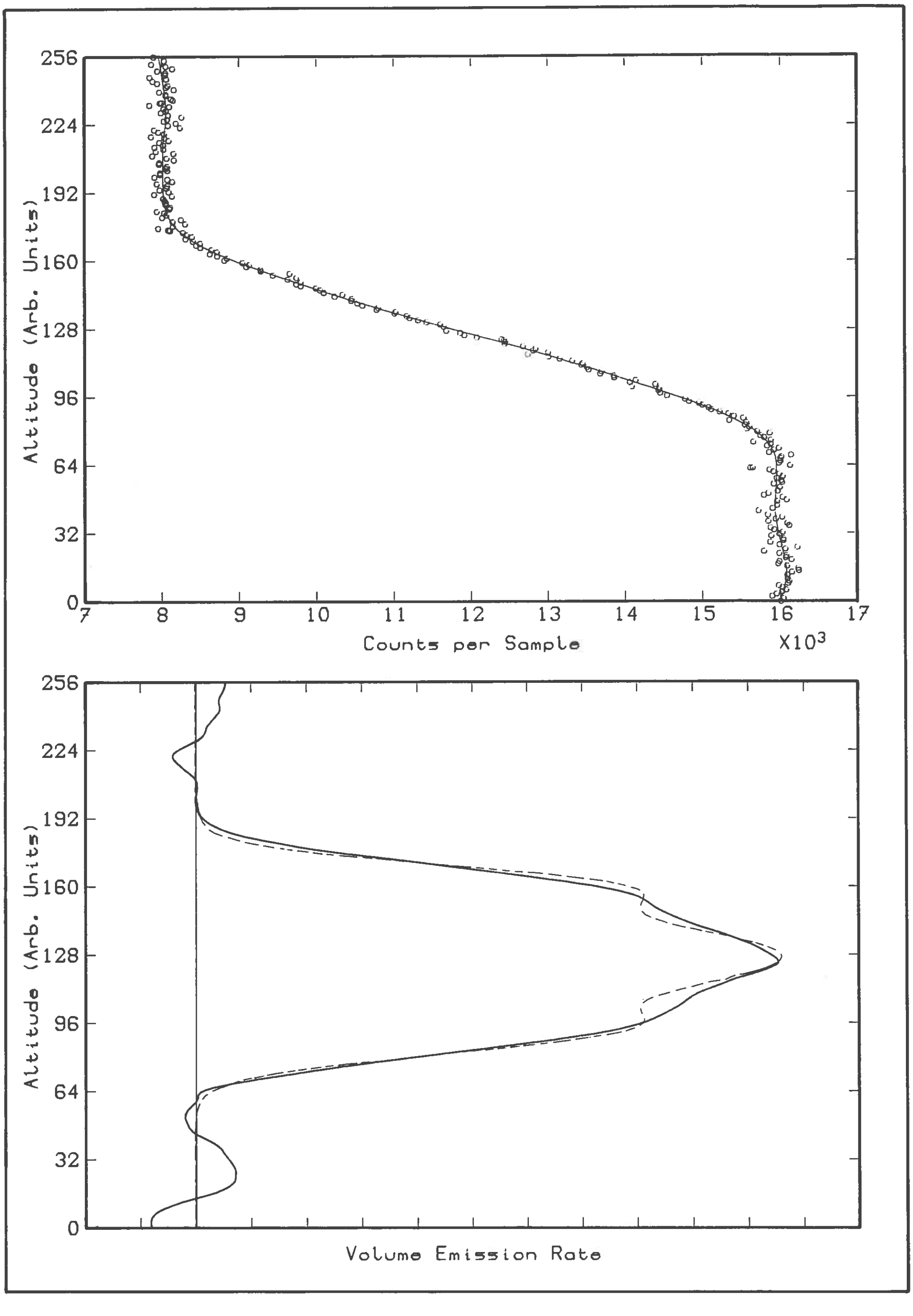

Figure 4
Test profile : data from Murtagh et al. (1984). *As for figure* 1.

cross-validation cubic spline which is then differentiated and compared to the original curve. Except for the lower peak which is not properly detected, the result seems satisfactory.

Of course real emission rates tend to be zero or near zero in some regions of the profile. We therefore constructed another profile to see how the smoothing behaved with near-zero slopes (fig. 2). We changed here to another count rate scale which is closer to the one shown by Murtagh *et al.* in their real data set.

An interesting advantage of the present technique seems to be the capability of extracting information from a reduced data set. For instance, figure 3 uses the same data set as figure 2, but only one point out of five is used. The relative maximum is of course degraded, but still present. Another consequence is the intensification of spurious oscillations at low altitude, their amplitudes being stronger than the one of the degraded peak.

The two next figures are comparisons with results obtained by Murtagh *et al.* in their figures 7 and 8. For the synthetic data set (fig. 4), the present technique, even if not able to discern the two relative maxima on both sides of the main peak, shows up two related inflexion points. If the relative maxima could not be properly attained, the sole reason is a lack of definition

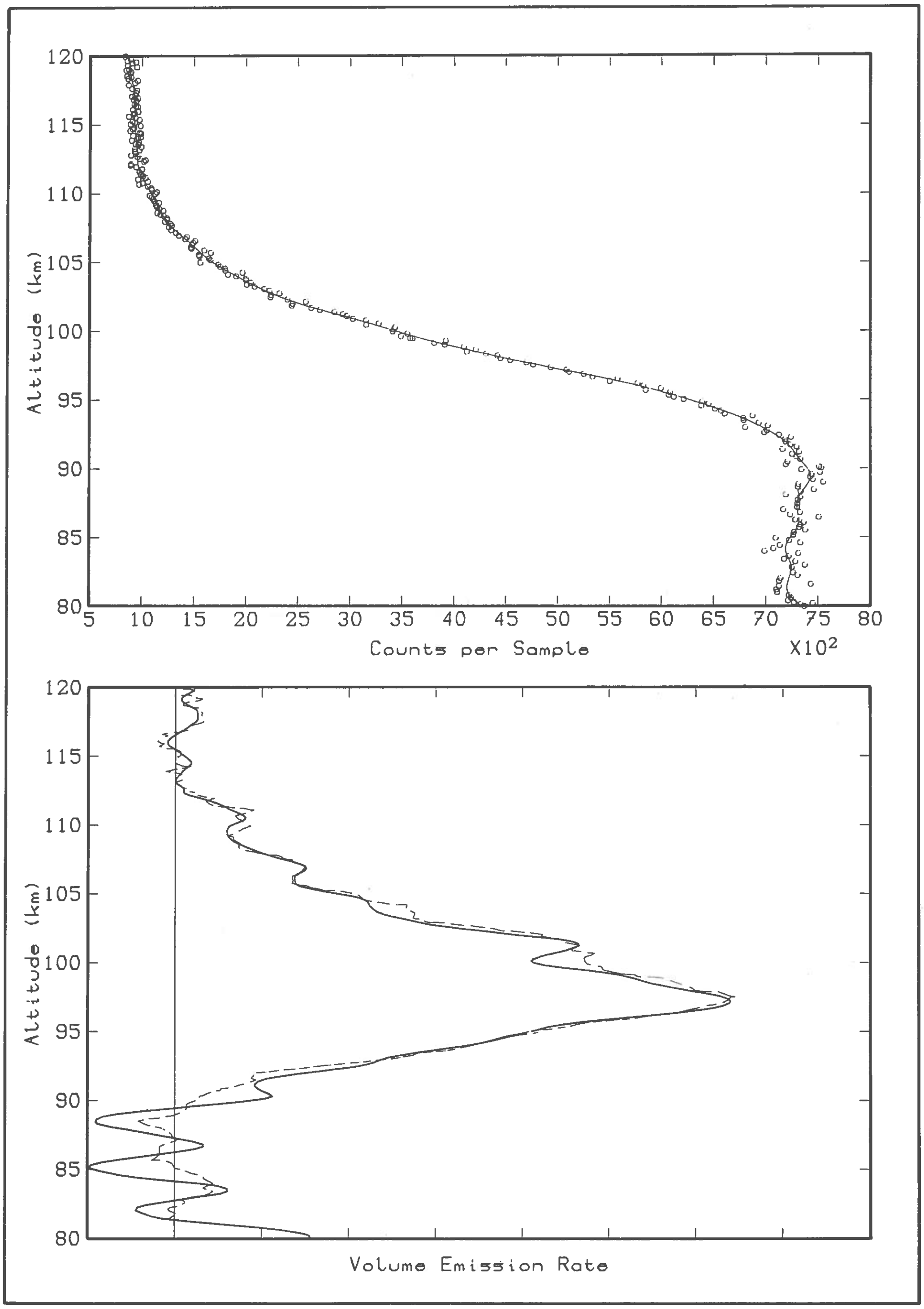

Figure 5
275 nm *photometer results : data from Murtagh* et al. (1984). *Corrected integrated profile (upper part). Result of the cross-validation method (lower part, solid line) compared to the result of the incremental straight line method of Murtagh* et al. *(lower part, dashed line).*

of the data : the best statistical analytical representation of the data set obviously blurs there some information.

Figure 5 shows the processing of a real data set. The principal peak is found at a 97.5 km altitude, in keeping with results by Murtagh *et al.* A relative maximum at a 102 km altitude is questionable, given the large amplitude of oscillations which occur at low altitude, where the volume emission rate should be zero or near zero. However, the experience brought by figure 3 shows that such a feature might be significative, the abovementioned oscillations being merely due to a much lower density of data points in the lower region. This small-scale structure of the layer is present in the results by Murtagh *et al.*, but not so conspicuously.

CONCLUSIONS

Cross-validation cubic splines seem well-adapted to smooth rocket profiles to derive volume emission rates. We showed that stratified layers could be extracted and that spurious oscillations of the resulting curve at low altitude are probably due to insufficient data. When the profile involves ill-pronounced small-scale details, the present technique is however unable to detect them correctly but shows up related disturbances on the curve (inflexion points). Small-scale features, when revealed by the present technique, could possibly be used with independent corroborative data for an interpretation in aeronomical terms. Finally, as pointed out by Murtagh (personal communication), the main drawback of the method seems to be the lack of an estimate of the uncertainty in the derived profile.

Acknowledgments

The author benefited by a correspondence with D. P. Murtagh who drew his attention to disadvantages of the present method. Stimulating remarks from the reviewers were also welcome.

REFERENCES

Allen D. M., 1974. The relation between variable selection and data augmentation on a method for prediction. *Technometrics,* **16,**125-127.

Craven P., Wahba G., 1979. Smoothing noisy data with spline functions : estimating the correct degree of smoothing by the method of generalized cross-validation. *Num. Math.,* **31,** 377-403.

Murtagh D. P., Greer R. G. H., McDale I. C., Llewellyn E. J., Bantle M., 1984. Representative volume emission profiles from rocket photometer data. *Ann. Geophysicae,* **2,** 4, 467-474.

Stone M., 1974. Cross-validation choice and assessment of statistical prediction. *J. Roy. Statist. Soc.,* ser. B., **36,** 111-147.

Thouvenot F., 1983. Frequency dependence of the quality factor in the upper crust : a deep seismic sounding approach. *Geophys. J. Roy. Astron. Soc.,* **73,** 427-447.

Utreras F., 1979. Utilisation de la méthode de validation croisée pour le lissage par fonctions splines à une ou deux variables. Thèse Dr. Ing., Université de Grenoble, 192 pp.

Wahba G., Wold S., 1975. A completely automatic french curve : fitting spline functions by cross-validations. *Comm. statist.,* **4,** 1-7.