- ***Communication*** : ASME Fluids Engineering Division Spring Meeting "Quantification of Uncertainty in Computational FLuid Dynamics"
  Titre : The adjoint sensitivity method, a contribution to the code uncertainty evaluation
  Auteurs : A. OUNSY, F. DE CRECY, B. BRUN

- ***Communication*** : NURETH 6
  Titre : Pseudo-Cubic Thin-Plate Type Spline Method for Analyzing Experimental Data
  Auteurs : F. DE CRECY

- ***Communication*** : NURETH 6
  Titre : The efffect of grid assembly mixing vanes on critical heat flux values and azimuthal location in fuel assemblies
  Auteurs : F. DE CRECY

- ***Communication*** : NURETH 6
  Titre : A better approach to the MDNBR concept
  Auteurs : F. DE CRECY

- ***Communication*** : NURETH 6
  Titre On component, volume heataed, boiling pool thermalhydraulics
  Auteurs : M. BEDE, C. PERRET, H. PRETEL, J.M. SEILER

- ***Communication*** : NURETH 6
  Titre : Reflood modelling under oscillatory flow conditions with Cathare
  Auteurs : J.M. KELLY, J. BARTAK, A. JANICOT

- ***Communication*** : NURETH 6
  Titre : The adjoint sensitivity method. A contribution to the code uncertainty evaluation
  Auteurs : A. OUNSY, B. BRUN

- ***Communication*** : NURETH 6
  Titre : Analysis of the UPTF downcomer tests with the Cathare multi-dimensional model
  Auteurs : I. DOR

- ***Communication*** : NURETH 6
  Titre : The top down reflooding model in the Cathare code
  Auteurs : J. BARTAK, D. BESTION, T. HAAPALEHTO

- ***Communication*** : NURETH 6
  Titre : New developments in Cathare 2 code
  Auteurs : F. BARRE, P. BOUDIER, M. PARENT, V. FIGEROU, M. FLORES, D. KADRI

- ***Communication*** : NURETH 6
  Titre : Interfacial friction in low flowrate vertical annular flow
  Auteurs : J.M. KELLY, R.L. FREITAS

PSEUDO-CUBIC THIN-PLATE TYPE SPLINE METHOD FOR ANALYZING EXPERIMENTAL DATA

F. de Crécy
Commissariat à l'Energie Atomique
Centre d'Etudes Nucléaires de Grenoble
Service de Thermohydraulique des Réacteurs
85 X - 38041 GRENOBLE CEDEX (France)
Tél. 76.88 32.18 Fax : 76.88.51 77

ABSTRACT

A mathematical tool, using pseudo-cubic thin-plate type spline, has been developed for analysis of experimental data points. The main purpose is to obtain, without any a priori given model, a mathematical predictor with related uncertainties, usable at any point in the multidimensional parameter space. The smoothing parameter is determined by a generalized cross validation method. The residual standard deviation obtained is significantly smaller than that of a least square regression. An example of use is given with critical heat flux data, showing a significant decrease of the conception criterion (minimum allowable value of the DNB ratio).

INTRODUCTION

In many scientific fields, a large number of experimental data points can be obtained, each one with unknown uncertainty, but there is no reliable physical model to describe the studied phenomenon. We need a mathematical tool to establish both the best estimate value and the confidence interval at any point. Usually, least square regression methods are used to solve this problem. However, quite restrictive assumptions underpin the use of such methods : the physical phenomenon has to obey an a priori given form model, or, in the best cases, a limited choice between several a priori given models. Only the adjustable coefficients depend directly on the experimental data points. Moreover, the residual standard deviation may seem to the experimentors rather high with respect to the expected experimental uncertainties.

An alternative method is proposed in this paper, based on the use of pseudo-cubic thin-plate type spline. The only underlying hypothesis is that the studied phenomenon is rather smooth (possessing continuous derivatives). One of the main advantages of such a method is that it results in a strongly validated residual standard deviation which is usually significantly smaller than that obtained with the classical least square regression method.

Although this mathematical tool was originally developed for thermohydraulic studies and especially critical heat flux, it may be used in many other scientific areas.

1. MAIN HYPOTHESIS

The physical phenomenon is assumed to be "rather smooth", that is to say possessing continuous derivatives with respect to all explaining (independant) variables, without vertical derivates. In practice, this phenomenon can be measured only with uncertainties which have to be added to the true value. We assume that the phenomenon may be modelled by :

$$y = f(\bar{t}) + \varepsilon(\bar{t}) \qquad (1)$$

where $\bar{t}$ represents all the explaining (independant) variables, $f(\bar{t})$ is a "true" deterministic smooth real function, and $\varepsilon(\bar{t})$ is a random function, of mean 0 and variance σ^2.

Both $f(\bar{t})$ and σ^2 are unknown, and we will look for estimations $\hat{f}(\bar{t})$ of $f(\bar{t})$ and $\hat{\sigma}^2$ of σ^2.

No hypothesis is made on the distribution function for $\varepsilon(\bar{t})$, but the mean is zero. In fact, the use of the results in most of the classical statistical tests will require that $\varepsilon(\bar{t})$ follow a normal (Gaussian) distribution.

2. AN ILLUSTRATION TO UNDERSTAND WHAT WILL BE DONE

Let us assume a (d+1) dimensional space of all independant and dependant variables. In this space, let us imagine a weightless thin elastic plate, connected to every experimental data point by a spring of length zero when not being pulled The elastic plate has energy or stress only in flexion or curvature, but can stretch parallel to itself without any stress or energy.

The equilibrium position of the plate is the one which minimizes the total energy, equal to the sum of the curvature energy of the plate, integrated over the whole parameter space, and the total pull energy of all the springs.

This equilibrium position will be considered as the solution to our problem.

So, it is obvious that the relative flexibility of the plate and springs is critically important for the equilibrium position.

Increasing steadily the relative stiffness of the springs from zero to infinity will lead to continuous deformation of the resulting plate, from a plane plate (which is the regression plane obtained with a multiple linear least square regression) to a plate which goes precisely through all the experimental data points (it is called an interpolating plate).

The choice of this relative flexibility is of primary importance and will be discussed later.

3. MATHEMATICAL PRINCIPLES

3.a. Basic solution

We have N experimental data points :

$$(y_i, \bar{t}_i)\ ; \quad i = 1,N$$

where y_i is the variable to be explained (the dependant variable), and $\bar{t}_i$ is a d-component vector containing the d explaining (independant) standardized variables, some of them may be binary to identify specific parameters.

The thin-plate we are looking for is the mathematical function which minimizes the total energy E_t :

$$E_t = E_c + \rho E_s \qquad (2)$$

where

E_c is the curvature energy of the plate and E_s is the energy of the springs.

ρ is the smoothing parameter ($\rho \geq 0$). It represents the relative flexibility of the plate and the springs.

For E_c, we use (as suggested by DUCHON [1]):

$$E_c = \int_{R^d} \sum_{k=1}^{d} \sum_{l=1}^{d} \left[\mathcal{F}\left[\frac{\partial^2 \hat{f}}{\partial t^k \, \partial t^l} \right](u) \right]^2 \|u\|^{d-1} du_1 \, du_2 \cdots du_d \qquad (3)$$

where $\|u\|$ is the euclidian norm in R^d.

$\mathcal{F}$ is the Fourier transform

and E_s is the energy of the springs

$$E_s = \sum_{i=1}^{N} (y_i - \hat{f}(\bar{t}_i))^2 \qquad (4)$$

That is to say the residual variance

Duchon [1] and Paihua [2] demonstrate that the solution to this problem is

$$\hat{f}(\bar{t}) = \sum_{i=1}^{N} \lambda_i \|\bar{t} - \bar{t}_i\|^3 + \sum_{j=1}^{d} \alpha_j t^j + \alpha_{d+1} \qquad (5)$$

(t^j is the j-th component of $\bar{t}$).

The same authors demonstrate also than the N+d+1 coefficients λ_i and α_j may be found as the solution of the (N+d+1) linear symetric system

$$\left.\begin{array}{ll} \sum_{k=1}^{N} \lambda_k \|\bar{t}_k - \bar{t}_i\|^3 + \frac{\lambda_i}{\rho} + \sum_{j=1}^{d} \alpha_j t_i^j + \alpha_{d+1} = y_i & i = 1, \ldots n \\ \sum_{i=1}^{N} \lambda_i \, t_i^j = 0 & j = 1, \ldots d \\ \sum_{i=1}^{N} \lambda_i = 0 & \end{array}\right\} \qquad (6)$$

where t_i^j is the j^{th} component of $\bar{t}_i$

From this system (6), it is clear that the solution (i.e. the coefficient λ_i and α_j) depends strongly on the smoothing parameter.

3.b. Remark : number of degrees freedom

Equation (5) is the characterisation of the solution of the minimization of E_t (equation (2)).

It may seem surprising to see that this formula has N coefficients λ_i and d+1 coefficient α_j, i.e. a total of N+d+1 coefficients for only N experimental data points.

This is true whatever the plate is : a very smooth plate, quite far from the data points but very close to the regression plane obtained by a multiple linear least square regression, has exactly the same number of coefficients as the interpolating plate, going precisely through all the experimental data points and then becoming far more curved and complicated.

This means that we must not confuse the number of coefficients with the classical statistical notion of number of degrees of freedom used by the formula.

The number of degrees of freedom used can be considered as a measure of the "complexity" of the plate.

For least square regression methods, the number of degrees of freedom used is equal to the number of adjusted coefficients, but this particular case is not true for splines.

It is well known that any type of smoothing spline has more coefficients than experimental data points. This is also true for multidimensional pseudo-cubic thin-plate type splines.

Wahba [5] and Silvermann [6] have introduced a "smoothing matrix"* and they have demonstrated that, when the smoothing parameter is determined by generalized cross validation (see § IV-c), the trace of this "smoothing matrix" matches with the number of degrees of freedom used.

The number of available degrees of freedom is the number N of data points minus the number of degrees of freedom used.

This number of available degrees of freedom is widely used in various statistical applications, i.e. to obtain an unbiased estimate of the residual standard deviation and confidence or tolerance intervals.

* *What is a smoothing matrix ? For fixed values of ρ and of the $\vec{t}_i$, $i = 1,N$, one can demonstrate that there is a linear dependancy between the N-dimensional vector $Y = (y_i \,;\, i = 1,N)$ of the raw values of the data and the N-dimensional vector $Z = (z_i \,;\, i = 1,N)$ of the smoothed values. So, there is a $N \times N$ matrix, named the smoothing matrix A_ρ such that : $Z = A_\rho Y$.*

This smoothing matrix is of primary importance for statistical calculations.

The trace of this smoothing matrix varies between d+1 (for a spline equal to the least square regression plane) and N (for the interpolating spline).

3. c. The cross validation method

The choice of the smoothing parameter ρ, that is to say the relative flexibility of the thin plate and of the springs is fundamental.

For $\rho = 0$, we obtain the multiple linear least square regression plane.

For $\rho = \infty$, we obtain an interpolating plate with

$$\forall_i \in \{1,\ldots,N\} \;,\; y_i = z_i \tag{7}$$

The solution $\hat{f}$ varies continuously with ρ between these two extreme cases.

What we consider as the best value of the smoothing parameter ρ is chosen by an iterating process called "generalized cross validation" (GCV).

For a given value of ρ, we eliminate the first experimental data point and build a thin-plate type spline with the N-1 remaining points. We only use this thin-plate to compute the predicted value at the first and previously eliminated point. Let us note e_1 the difference between the predicted and measured value at this data point. After reinstating the first point, we treat the following point in the same way and, step by step, all the data points.

So, we obtain N values e_i, $i = 1, N$ which are the differences between predicted and measured values, each predicted value based on the N-1 other experimental data points.

Once completed, we compute

$$V(\rho) = \frac{\sum_{i=1}^{N} e_i^2}{N} \tag{8}$$

and we iterate on the value of ρ until the minimum of $V(\rho)$ is found.

The value of ρ which minimizes $V(\rho)$ is called the cross validated value of the smoothing parameter, and will be used in all subsequent calculations.

Some algebraic manipulations (Stone [3]) and a fast Monte-Carlo method (Girard [4]) lead to a fast and easy way to compute $V(\rho)$, which requires the solution of only one N+d+1 linear symmetric system with several (typically 2 to 5) right hand terms.

3.d. Uncertainties

As for the very classical linear regressions, there are two types of uncertainty

residual uncertainty : an estimation $\hat{\sigma}$ of σ, the standard deviation of the random function of equation (1)

predictor's uncertainty : the most probable difference between the true function f and its estimation $\hat{f}$ ($\hat{f}$ is a random variable (because it depends on the y_i, which are random variables) of mean f. The standard deviation of $\hat{f}$ is called the residual uncertainty).

Silverman [6] gives an unbiased estimate of the uncertainty of the predictor, reliable only in the vicinity of each experimental data point

$$\sigma_{i_{pred}} = \sqrt{a_{ii}}\ \hat{\sigma} \qquad (9)$$

where a_{ii} is a diagonal term of the smoothing matrix A_ρ.

These values are arbitrarily extended by an interpolating plate to cover the experimental data range.

4. COMPARISON WITH LEAST-SQUARE REGRESSION METHOD

The main advantages of the thin-plate type spline method with respect to least square regression methods are :

it does not require any given physical model.

it usually obtains a significantly smaller standard deviation (typically, it is reduced by a factor 1.2 to 1.7 for critical heat flux experimental data).

The main drawbacks are :

. it is not a well known method.

. it requires more computational power : the memory size required varies as $(N+d+1)^2$ and the CPU time as $(N+d+1)^3$. Typically, it takes around 3 to 5 minutes CPU of a HP715/50 work station and around 4 MBytes of memory for a 700 data point calculation.

. it is less easy to present in a paper and to use for physical or mathematical extensions : in the absence of any given physical model, physical analysis or extension has to be graphic.

Remark *: The thin-plate spline method may be considered as a non-parametric least square method, but it is not a classical, kernel smoothing, non-parametric least-square method.*

As is the case for the classical, kernel smoothing, non-parametric least-square method, the number of available degrees of freedom is a real function varying continuously with the strength of the smoothing.

5. EXAMPLE TO 8 EPRI CHF TESTS

As an example, we put this methodology into practice with 8 rod bundle CHF tests developed for EPRI in Columbia University by Fighetti and Reddy [7]. All the tests numbered 156 to 164, except 159, have been used. These CHF tests have been performed with a 5 × 5 rod bundle with Westinghouse mixing grids and the main characteristics of these tests are presented in fig. 1.

For each test and in many cases for the entire set of tests, we build both a stepwise regression and a pseudo-cubic multidimensional thin-plate type spline with the smoothing parameter determined by GCV, on exactly the same data set.

When the whole set of tests is taken into account, we can use some specific binary variables :

A = -1 for typical cell,
A = +1 for guide thimble cell.

B = -1 for a 14 ft heating length,
B = +1 for a 8 ft heating length.

C = -1 for a 26" distance between mixing grids,
C = +1 for a 22" distance between mixing grids.

D = -1 for an axially uniform heat flux,
D = +1 for a non-axially uniform heat flux.

The expressions obtained by a step-by-step multilinear least square regression is in the form of :

$$\phi = \sum_{i,j,k,\ell,m,n,p} \lambda_{i,j,k,\ell,m,n,p}\, P^i G^j x^k A^\ell B^m C^n D^p$$

with $(i,j,k) \in \{0,1,2\}^3$ and $(\ell,m,n,p) \in \{0,1\}^4$

We used a very limited number (12 as a maximum) of non-zero coefficients $\lambda_{ijk\ell mnp}$.

The use of each binary variable A, B, C, D is of course optional.

Table 1 shows a comparison between regression and spline, mainly in terms of the number of degrees of freedom (Ndof) used by the predictor, the residual standard deviation (unbiased estimation) and $\mathcal{R}^2$, the rate of total variance explained by the predictor The presence of one or several letters, A, B, C, D, in the name of the calculation means that the corresponding effect has been taken into account

For example, calculation W_ _BC_ takes into account the effects of the heating length (letter B) and the distance between mixing grids (letter C) but neither the effects of the nature of the cell (letter A) nor the axial variation of the heat flux (letter D).

Figures 2 and 3 show two comparisons between correlations and pseudo-cubic splines for both 1 - $\mathcal{R}^2$, the rate of total variance not explained by the predictor, and $\hat{\sigma}$, an unbiaised estimation of the residual standard deviation. We can see that the use of a spline usually leads to a significant decrease of the unbiased estimation of the residual standard deviation.

Pseudo-cubic thin-plate type spline may also be used for the analysis of variance. Indeed, to test the existence of a specific effect, we have to build two pseudo-cubic splines on the same experimental data set, one with a binary variable coding this specific effect, the other without. A Fischer-Snedecor test shows whether if there is actually a significant decrease of the residual standard deviation due to the binary variable, and if so, whether this specific effect is significant.

An analysis of variance of this sort is presented in figure 4 and leads to the following conclusions :

According to these 8 CHF tests :

- there is no significative difference in CHF between a typical cell and a guide thimble cell.
- the effects of the heating length and of the distance between the mixing grids are highly significant.
- the effect of axially non uniform heating is significant.

It is also possible to obtain many graphics (cartesian curves, contour map, 3D views,...), confidence and tolerance intervals for the CHF, and to build many look-up tables. This idea in CHF studies is becoming accepted for predicting non-standard conditions. It is currently used in various safety codes, e.g. CATHARE (France), RELAP5/MOD3 (USA), THERMOHYDRAULIK (Germany), CATHENA and ASSERT (Canada).

	Nb of points	Spline			Correlation		
		Ndof used	$\hat{\sigma}$	$\mathcal{R}^2$	Nb of coeff	$\hat{\sigma}$	$\mathcal{R}^2$
W156	113	21 1	.1367	.7543	7	1458	6776
W157	79	17 2	.1066	943	6	1075	9319
W158	68	24 4	1376	.9378	5	1676	8666
W160	63	19 7	.0997	9778	7	127	9534
W161	65	19 7	0519	97	7	0605	9479
W162	73	13.5	.1335	.8615	6	133	8453
W163	39	7	.138	.8107	4	.1348	.8026
W164	96	32.9	103	.9494	7	.1295	.8873
W____	596	32.5	.2333	.7726	8	.2433	.7418
W_A__	596	44.9	227	.7895	9	.2388	.7517
W__B__	596	55.5	.1667	.8886	9	1831	854
W___C_	596	60.2	.2152	.816	9	236	7574
W____D	596	50.1	.208	.8249	8	2239	.7813
W__BC_	596	99.3	.1349	.9329	11	.17	.8745
W__B_D	596	70.1	.1609	.899	10	.1837	.8533
W___CD	596	74.4	.1673	.8918	11	.1947	.8355
W__BCD	596	125.1	.122	.948	12	.1601	.889
W_A_CD	596	98.1	.1563	.9097			
W_AB_D	596	80.8	.1543	.909	12	.1788	.8615
W_ABC_	596	123.8	.1256	.9447	12	.1646	.8826
W_ABCD	596	131.6	.1195	.9508	12	.1601	.8890

Table 1 : Comparison of the residual standard deviation and the rate of explained variance between regressions and pseudo-cubic splines built on the same data set : the first part of the table corresponds to test by test calculations, and the second part corresponds to calculations on the whole set of data.

CONCLUSION

We have presented a mathematical method for analyzing experimental data using pseudo-cubic thin-plate type-spline.

As long as the physical phenomenon may be considered as "smooth" and there is no reliable physical model available to describe the phenomenon, the use of this method must be advised, especially if the residual standard deviation is particularly important.

This is particulary the case for the burn-out studies, where the critical heat flux is one of the limiting phenomena for the nuclear power plant. In this case, the use of the present method may lead to a significant decrease of the residual standard deviation.

REFERENCES

1. Duchon J., "Splines minimizing rotation-invariant semi-norm in Sobolev spaces", *Constructive Theory of Functions of Several Variables, Springer-Verlag, Berlin*, 85-100 (1977)
2. Paihua M.L., "Quelques méthodes pour le calcul des fonctions splines à une ou plusieurs variables"; *Thèse, Univ. of Grenoble* (1978)
3. Stone M., "Cross-validatory choice and assessment of statistical prediction (with discussion)", *J. Roy. Statistic. Soc.*, Ser. B, 36, 111-147 (1974)
4. Girard D., "Asymptotic optimality of the fast randomized versions of GVC and CL in ridge regression and regularization", *The Annals of Statistics*, for parution in 1991.
5. Wahba G., "Bayesian confidence intervals for the cross validated smoothing spline", *Technical Report n° 645, Department of Statistics, Univ. of Wisconsin*, Madison (1981)
6. Silvermann B.W., "Some aspects of the spline smoothing approach to non-parametric regression curve fitting", *J. Roy. Statistic. Soc.*, 47, 1, 1-52 (1985)
7. Fighetti C.F., Reddy D.G., "Parametric study of CHF data", *EPRI NP 2609 RP 813, Columbia Univ.*, New-York 10027, sept. (1982)

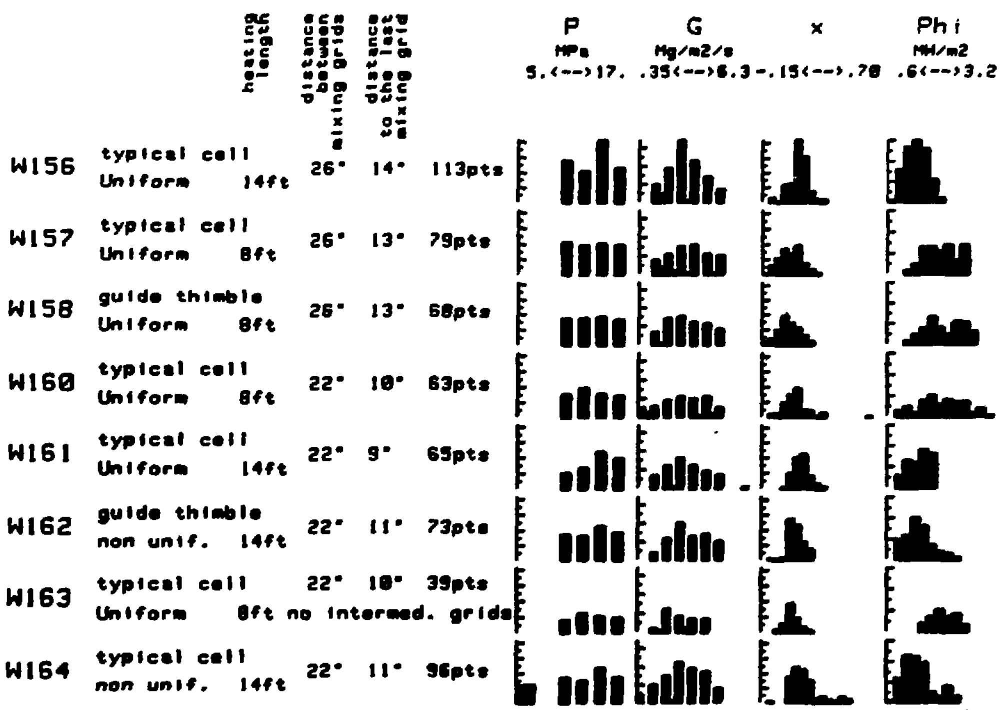


FIGURE 1 : The main characteristics and the distribution of the parameters of the 8 EPRI CHF tests used in this example

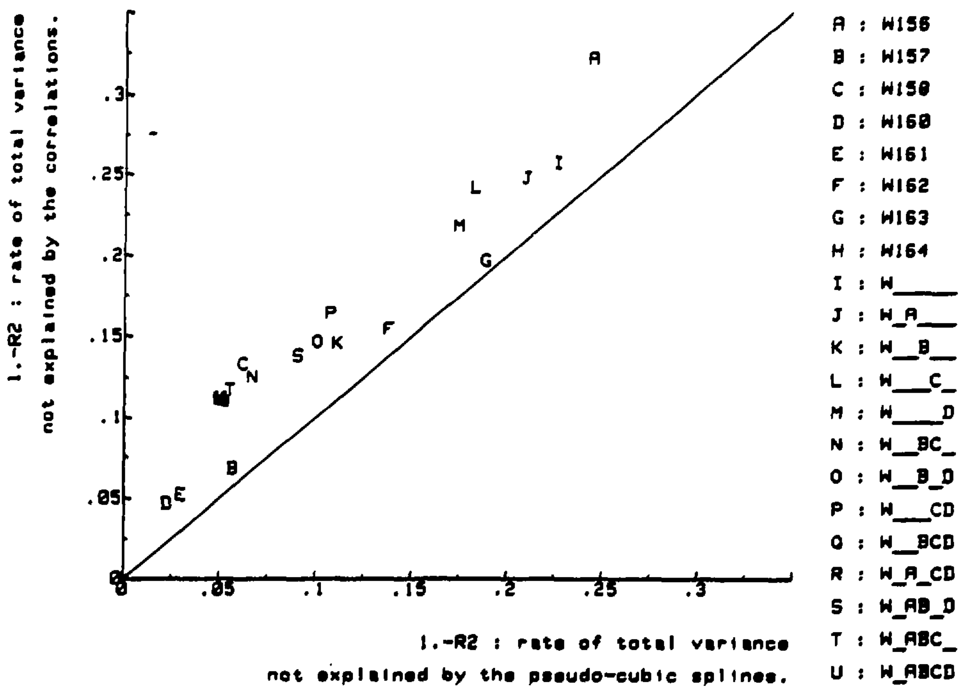


FIGURE 2 : Comparison, for various experimental data sets (one per letter), of the rate of residual variance between pseudo-cubic thin-plate type spline (X-axis) and least square regression (Y-axis)

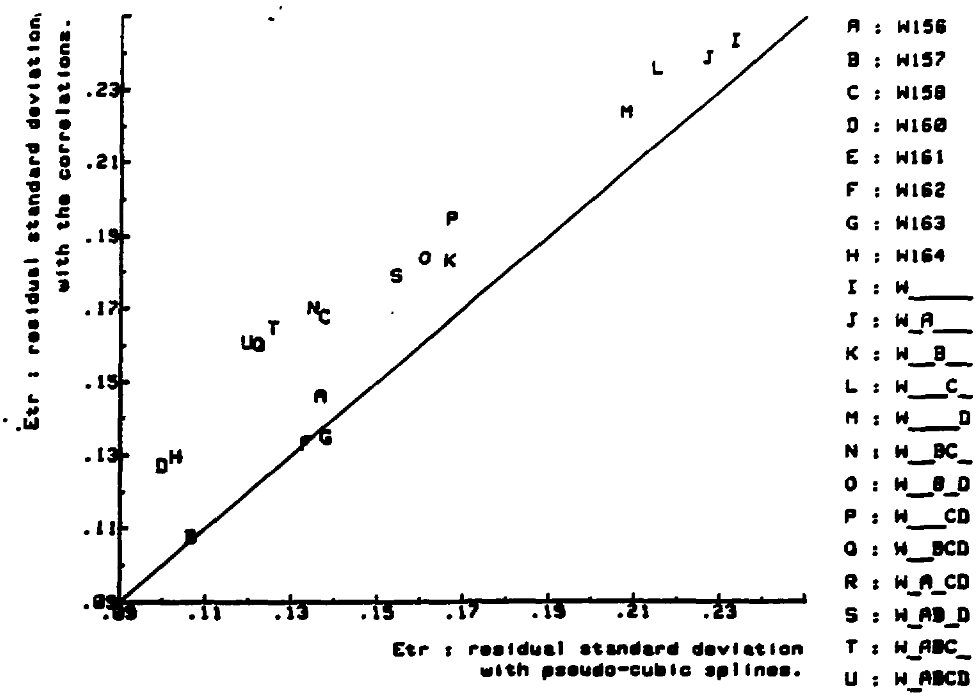
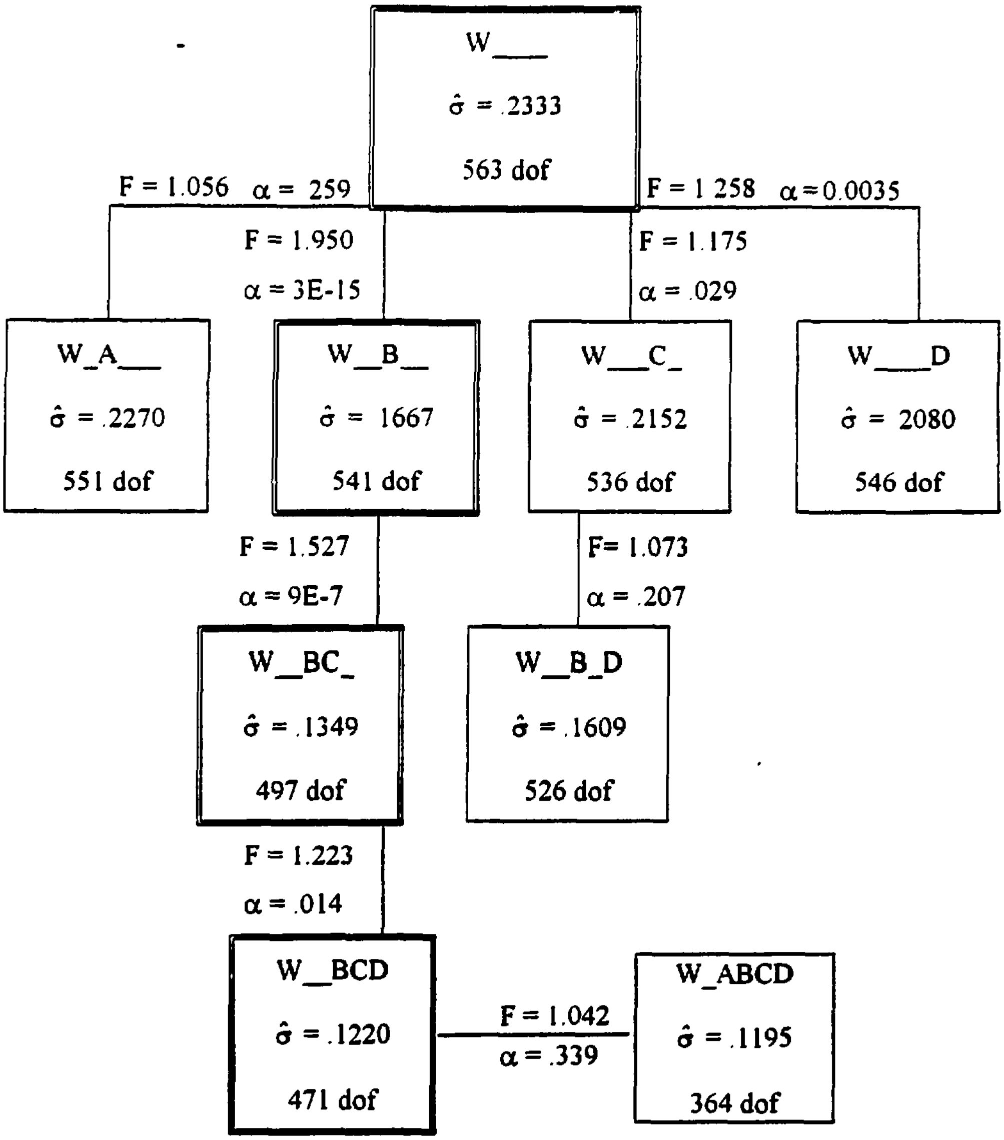


FIGURE 3 : Comparison, for various experimental data sets (one per letter), of the residual standard deviation pseudo-cubic thin-plate type spline (X-axis) and least square regression (Y-axis)

ANALYSIS OF VARIANCE USING SPLINES



F = Fischer-Snedecor's variable
α = probability that there is no effect

FIGURE 4

Each square represent a spline calculation.
The presence of letters A,B, C or D represent a specific effect, as explained in § VI