

Comparing Curves Using Additive Models

Esteban Walker and S. Paul Wright
Dept. of Statistics, University of Tennessee
Knoxville, TN 37996

Abstract

Advances in technology have increased dramatically the amount of data measured in industrial processes. Thousands of measurements are available nowadays in operations where previously only a single measurement, at a given point in time or space, was taken. These measurements allow the reconstruction of the whole profile or "signature" of the operation over time or space. Examples are the tonnage applied in a stamping press during a stroke and the density profile of particleboard. Many of these signatures have complicated forms that are not well modeled with parametric models. In this paper, a relatively new class of models called additive models is used to assess the sources of variation active on these signatures. The model contains a nonparametric or smooth portion to model the form of the signature, and a parametric portion to evaluate other sources of variation. An analysis of variance table is developed to test the magnitude of sources of variation. These techniques are illustrated using density profiles of engineered wood boards. Instructions on the implementation of these techniques in S-Plus and SAS are given.

1. Introduction

In many situations, the response of interest is not a single point, but a whole profile. In biostatistics, curves arise naturally as the reaction of an organism to a treatment over time. Techniques like repeated measures and profile analysis have been very popular for a long time (see e.g. Morrison, 1990). In industry, one of the first to deal with curves as responses was Taguchi (1987), who called them "dynamic characteristics." More recently, Miller and Wu (1996) and Bisgaard and Steinberg (1997) have looked at methods to design and analyze experiments involving dynamic characteristics. In these papers the curves of interest consist of few points and thus are adequately modeled with low order polynomials. In many potential applications, however, the curves consist of

hundreds of points and have complicated forms that are not well modeled even by high order polynomial models.

An area that has seen great growth recently has been that of nonparametric regression or smoothing. The idea of smoothing is to use flexible functions, which may themselves be combinations of simpler functions, to fit a curve to the data. The main advantage of smoothing is that the functional form of the curve is not determined a priori. Hastie and Tibshirani (1990, p. 9) say:

"A smoother [or smoothed curve] is a tool for summarizing the trend of a response measurement Y as a function of one or more predictor measurements X_1, \dots, X_p . It produces an estimate that is less variable than Y itself; hence the name smoother. An important property of a smoother is its *nonparametric* nature: it doesn't assume a rigid form for the dependence between Y and X_1, \dots, X_p ."

In many cases, the variation in the data is naturally described by models that contain combinations of parametric and nonparametric terms. Recognizing this fact, Hastie and Tibshirani (1990) introduced a new and powerful class of models called generalized additive models (GAM). In this article, a GAM that combines smoothers and parametric terms is used to compare a series of curves. The landmark book by Hastie and Tibshirani (1990) is an excellent introduction to GAM.

This article illustrates the use of GAM to analyze dynamic responses. In Section 2, smoothing techniques are discussed and illustrated with an example. GAM are presented in Section 3; an ANOVA decomposition is proposed and illustrated using a real data set. Section 4 expands the ideas of Section 3 for multiple sources of variation and illustrates them using a larger data set. Conclusions are included in Section 5. An appendix includes S-Plus and SAS code for the examples presented.

2. Smoothing

2.1 Basics

One of the basic objectives of data analysis is to identify the "signals" or systematic effects in a data set. This task is complicated by the presence of "noise" or random variation. In other words, the goal is to "filter" the data in order to separate the signal from the noise. A common situation is when data are collected to investigate the relationship between two variables. The general model is

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) + \varepsilon,$$

where \mathbf{y} is a response that depends on a regressor \mathbf{x} through a, usually unknown, function \mathbf{g} . ε represents the noise which, without loss of generality, is assumed to have mean zero and variance σ^2 and be independent of \mathbf{x} . Two basic objectives of the analysis in this context are to estimate the form of \mathbf{g} and the magnitude of σ^2 .

An undesirable aspect of parametric techniques is that they are based on *a priori* models that often are selected rather arbitrarily. To ameliorate this problem, many diagnostic tools (i.e. influence analysis and collinearity diagnostics) have been developed to supplement the traditional parametric approach to detect problems and, if necessary, modify the model in light of the observed data. However, the use of these diagnostics requires a certain level of sophistication on the part of the user. On the other hand, the very fact that the diagnostics are based on a flawed model could hamper the detection of inadequacies and lead to erroneous conclusions.

The idea of smoothing is to fit flexible functions whose final form is determined by the data and by the chosen level of smoothness of the curve. Many smoothing methods exist and several excellent books on the subject have appeared recently (e. g. Härdle 1990, and Green and Silverman, 1994). Figure 1 illustrates smoothing splines at different levels of smoothness.

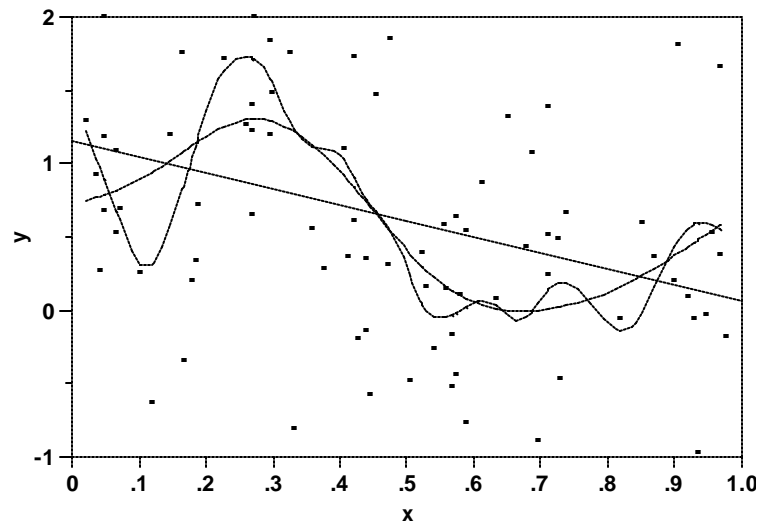


Figure 1. The solid line is a "less smooth" spline than that the broken line. As a reference, the least squares line (smoothest line possible) is also plotted.

The most common smoothing techniques, like smoothing splines, regression splines, kernel, and lowess belong to the class of linear smoothers, so called because they

can be written as $\hat{\mathbf{g}} = \mathbf{H}\mathbf{y}$. In this expression $\hat{\mathbf{g}}$ is the value of the fitted curve at the observed points, \mathbf{y} is the n vector of responses, and \mathbf{H} is an $n \times n$ "hat" matrix that depends on the smoothing method, the observed values of \mathbf{x} , and the selected level of smoothness, but not on \mathbf{y} . Consequently, well known results from linear models apply to these models. For example, the number of "equivalent degrees of freedom" (edf) of a smoother is computed as $\text{trace}(\mathbf{H})$, whereas an estimate of the error variance is $s^2 = \text{SSE}/\text{trace}(\mathbf{I} - \mathbf{H})$, where SSE is the residual sum of squares.

As a rule, the higher the level of smoothing, the fewer the edf used by the fit, and the higher the error edf. For example, the "smoothest" curve is a straight line that uses only one edf, whereas less smooth curves are "wigglier" and use more edf. In this respect, smoothing is akin to fitting different order polynomials to the data. Deciding the amount of smoothness is critical and, in some way, equivalent to choosing a model in a parametric analysis.

The amount of smoothing can be determined either subjectively or objectively. Graphically, one could plot some characteristic of the model (e.g. SSE) against the value of the smoothing parameter and select a value for which the characteristic "stabilizes." Another possibility is to select the amount of smoothing that eliminates trends from residual plots or that exposes interesting features of the curve. Even though subjective methods are simple and reasonable, objective or "automatic" methods are needed in some situations such as when there are a large number of curves, or when the aim is to make inferences from the curve. A prime example of a situation that calls for an automatic method is when the smoothed curve is used to test for lack of fit of a parametric model (see Hart, 1997). Ultimately, the "optimal" level of smoothness depends on the amount of noise and the form of the underlying curve. Among the most popular automatic methods to choose the smoothing parameter are generalized cross-validation (GCV), generalized maximum likelihood (GML) and the plug-in method (see e.g. Härdle, 1990, p. 147, Green and Silverman, 1994, p. 29, Hart, 1997, Chap. 4, Wabha, 1990).

Often times in industrial applications, the curves are fairly well defined, except for a small amount of measurement error. In these cases, the smoothed curve is used, not to find structure, but as a parsimonious way to describe a series of measurements. For well defined curves, an adequate amount of smoothing can be easily determined by

plotting the residuals. The largest amount of smoothing (smallest model edf) that yields a residual plot free of trends should be chosen.

The use of smoothers has been hampered by the fact that they do not have an explicit form since they are defined locally and therefore can't be expressed globally. Smoothers are very powerful when the objective is to find structure in the data but not to describe them analytically. For the comparison of structures, particularly complicated ones, smoothers are invaluable.

Another barrier to the use of smoothers has been that, in order to perform well, relatively large samples are needed. This is because, due to their flexibility, smoothers absorb many more degrees of freedom than parametric models to fit the model. This has become a lesser limitation as instruments capture more and more information and large amounts of data are available.

2.2 Example

Manufacturers of engineered wood boards, which include particleboard and medium density fiberboard, are very concerned about the density properties of the board produced. The density is measured using a profilometer which uses a laser device to take a series of measurements across the thickness of the board. A profilometer takes multiple measurements on a sample (usually a 2x2 inch piece) to form the vertical density profile (VDP) of the board. Figure 2 illustrates a typical VDP of a board after the sanding operation. It consists of 314 measurements taken 0.002 inches apart.

The overlaid curve is a regression spline (B-spline) fitted to the data with 16 edf. This amount of smoothing was chosen based on the residual plot. More smoothing (less edf) generated clear patterns in the residuals, whereas less smoothing (more edf) did not improve on the appearance of the plot. (see Appendix 1 for residual plots for 10, 16 and 20 edf).

Using notation borrowed from S-Plus (1997), this curve can be described as $VDP = bs(\text{Depth}, df = 16)$.

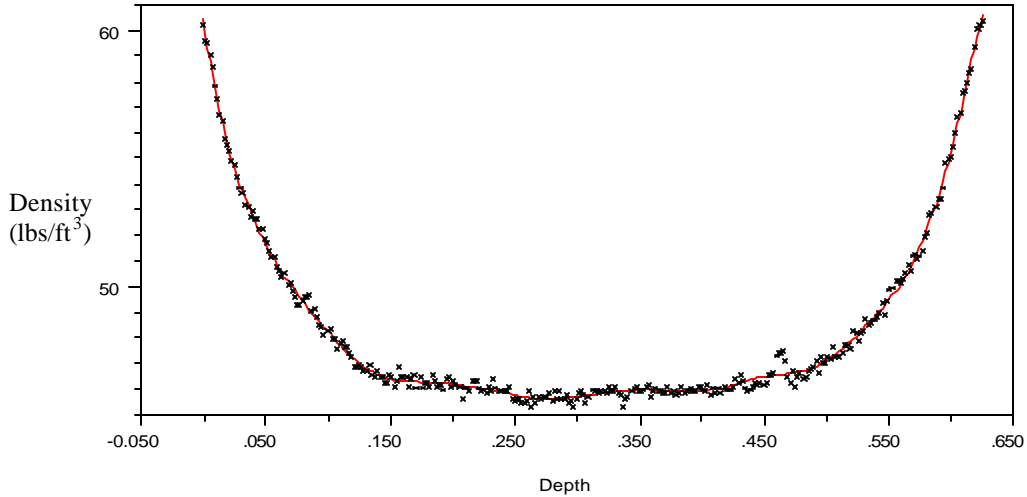


Figure 2. Example of the typical VDP of a sanded board along with a B-spline

3. Comparison of curves

3.1 Generalized Additive Models

Linear models belong to a larger class called generalized additive models (GAM) in which the effect of multiple regressors enter the model in an additive way. In a GAM, however, these effects do not have to be parametrically defined. In its most general form, a GAM is written as

$$\mathbf{h}(\mathbf{y}) = \sum_{j=1}^p \mathbf{g}_j(\mathbf{x}_j) + \varepsilon,$$

where \mathbf{h} is the link function, the \mathbf{g}_j 's are univariate functions, one for each predictor and ε represents noise with the same properties as before. In this paper, the responses will be assumed to have an approximate conditional normal distribution, so the link function will be the identity. The \mathbf{g}_j 's can be parametric (e.g., $\mathbf{g}_j(\mathbf{x}_j) = \mathbf{x}_j\beta_j$) or nonparametric (e.g. a smoothed curve). For example, if the response depends on two regressors, one of which enters linearly and the other as a smooth function, the model can be expressed as

$$\mathbf{y} = \mathbf{x}_1\beta + \mathbf{g}(\mathbf{x}_2) + \varepsilon.$$

Due to their nature these types of models are also called semiparametric or partial smoothers (see Green and Silverman, 1994, Ch. 4).

A GAM is usually fitted using a relatively simple, but computer intensive, method called backfitting. This method consists of sequentially fitting individual portions of the model and using the partial residuals to fit the remaining parts of the model. This is done for each term in the model and iterated until convergence. For more details on all aspects

of additive models see the seminal book *Generalized Additive Models* by Hastie and Tibshirani (1990).

Using indicator variables in a GAM provides an intuitive and elegant way to compare curves that are measured at the same points. For example, a model suitable for the comparison of two curves is

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) + \mathbf{z}\beta + \mathbf{g}(\mathbf{z},\mathbf{x}) + \varepsilon,$$

where \mathbf{z} is either 0 or 1, indicating the different curves and \mathbf{x} denotes the values at which the response is observed. In this model β is an estimate of the difference in average levels of the curves and the interaction term, $\mathbf{g}(\mathbf{z},\mathbf{x})$, allows the curves to have a different form, i.e. not to be equidistant. Notice that both nonparametric terms are based on the same smoother, \mathbf{g} . This is a small limitation and greatly simplifies the interpretation and the computation of the fitted curves. The GAM's discussed in this article use the same smoother for all the nonparametric terms.

The contribution of each term can be assessed using F-tests similar to those in linear models using sums of squares and edf. The test for the equality of the two curves is equivalent to testing the nullity of the last two terms in the above model. Such a test is proposed later and is an alternative to that proposed by King, Hart and Wehrly (1991) used in the context of VDP by Winistorfer, Young and Walker (1996).

The model above can be generalized to compare k curves, by including $k-1$ indicator variables, $z_1 \dots z_k$, and corresponding interaction terms. The model can be expressed as

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) + \sum_{i=1}^{k-1} z_i \beta_i + \sum_{i=1}^{k-1} \mathbf{g}(z_i, \mathbf{x}) + \varepsilon. \quad (1)$$

The appropriate interpretation of the terms in the model would depend on the parameterization used for the indicator variables. Regardless of the parameterization, the joint test of the effects of indicator variables and interactions is equivalent to a test of equality of the curves. One of the advantages of this approach is that the differences among curves can be separated into differences in level (β 's) and differences in "form" (interactions). Fitting model (1) is equivalent to fitting separate smoothers to each curve.

Table 1 presents the ANOVA for model (1) for k curves with n common points with each smoother in the model accounting for p degrees of freedom. This analysis is similar to analysis of covariance in the sense that the effects of interest are adjusted by

the covariate \mathbf{x} . In fact, since all the curves are observed at common values of \mathbf{x} , the sum of squares for level is the usual between-curve sum of squares, regardless of the level of smoothing. This is because the adjustment is the same for every curve.

Table 1. ANOVA table for analyzing the differences among k curves using a GAM

Source	Edf	SS
Covariate (x)	p	SS(x)
Level (β 's)	k-1	SS(level)
Form (interaction)	p(k-1)	SS(form)
Residual	k(n-p-1)	SSE
C Total	nk-1	

The sums of squares are computed by fitting sequentially the terms in the model. Using some of the ideas of Hastie and Tibshirani (1990, p. 155), an asymptotic test for the equality of the curves can be constructed as

$$F = \frac{\{SS(\text{level}) + SS(\text{form})\}/(p+1)(k-1)}{SSE/k(n-p-1)}.$$

This test can be divided into two subtests to assess the effect of differences in level and form. If the k curves are assumed to be a random sample from a population of curves, then the subtests are

$$F_L = \frac{SS(\text{level})/(k-1)}{SS(\text{form})/p(k-1)} \quad \text{and} \quad F_F = \frac{SS(\text{form})/p(k-1)}{SSE/k(n-p-1)}.$$

If the curves are fixed then the denominator of F_L would also be $SSE/k(n-p-1)$.

Approximate tests can be derived by comparing these statistics against F-distributions with corresponding degrees of freedom.

3.2 Fitting interactions

The backfitting algorithm is relatively simple and works well when the regressors are continuous. However, the smoothing of interactions with indicator variables poses computational problems. The problem is that these type of interactions can not be generated simply by smoothing a product like in linear models. Hastie and Tibshirani (1990, p. 264) propose several ways of fitting these interactions. Even though software is

available (e.g., Wang, 1998), currently the only way of fitting interactions of this type with commercial software is by using either B-splines or natural splines in S-Plus or in SAS's TRANSREG procedure. It is important for the practitioner to realize that even though S-Plus allows the use of these interactions with other smoothers in the `gam` command, the results will not be correct. The results reported in this article were generated using B-splines with S-Plus. The results using natural splines were similar. The code used is included in Appendix 2.

3.3 Example

The VDP's of nine boards selected in the same eight-hour shift are illustrated in Figure 3a. Differences in level are evident, whereas differences in form are not nearly as clear. A suitable model to analyze differences between profiles is

$$\mathbf{VDP} = \mathbf{g}(\mathbf{Depth}) + \mathbf{Board} + \mathbf{g}(\mathbf{Depth}, \mathbf{Board}) + \varepsilon.$$

In this model \mathbf{g} is a smoother and \mathbf{Board} is a categorical factor taking on values from 1 to 9 depending on the board (represented in the model by eight indicator variables generated automatically by the software). All the smoothing terms in the model are B-splines based on 16 edf because lower values yielded clear trends in the residual plot and higher values did not have a noticeable effect on either the plot or the SSE.

Fitting the above model with only the first term fits a single "average" curve (not shown) to the nine curves. The fitted values for the no-interaction model (only the first two terms) are plotted in Figure 3b. This plot shows curves with different levels that are equidistant, in the sense that the vertical difference between any two curves is constant. In other words, the model allows for differences in level, but not in form.

Figure 3c shows the fit from the full model, which is equivalent to fitting separate B-splines to each board. As is apparent, the curves now vary both in level and form.

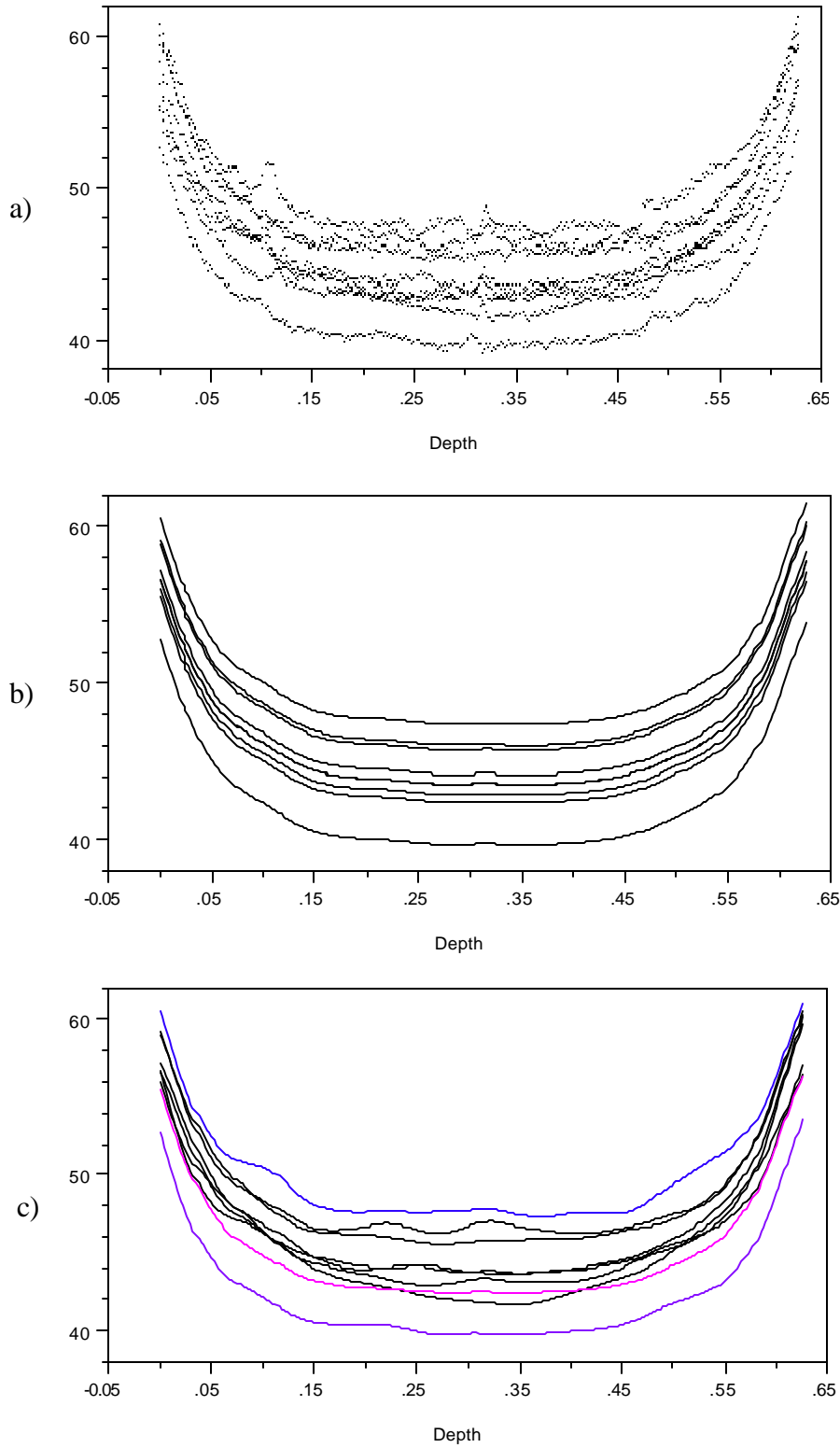


Figure 3. a) Observed VDP of nine boards. b) Fit without interaction, c) Fit with interaction. The vertical axis in all plots is density in lbs/ft³.

Table 2 shows the results of the ANOVA decomposition for these data. The overall test for equality of the curves is

$$F = \frac{(13,180.52 + 672.37) / 136}{0.07} = 1,470.55, \text{ with 136 and 2673 df,}$$

with $p < 0.0001$. This result suggests that there are significant differences among the curves.

Table 2. ANOVA table for the comparison of nine VDP's.

Source	Edf	SS	MS
Covariate (Depth)	16	33,797.17	
Level (Board)	8	13,180.52	1,647.57
Form (interactions)	128	672.37	5.25
Residual	2,673	187.99	0.07
C Total	2,825	47,838.06	

Assuming that the nine VDP's are a random sample from the shift, the nature of the differences can be further explored using the statistics:

$$F_L = \frac{1,647.57}{5.25} = 313.82 \quad \text{and} \quad F_F = \frac{5.25}{0.07} = 75.00,$$

with 8, 128 and 128, 2673 degrees of freedom, respectively. The observed significance levels are both $p < 0.0001$, suggesting that the curves differ both in level and form. The significance of these results is not unexpected since the level of noise is very low compared with the differences among the curves. In such cases, it is insightful to complement the analysis by computing the amount of variation due to the different components. Since the majority of the variation is due to the curve itself (or the covariate) it seems reasonable to remove this variation before computing the proportions. Using 14,040.89 ($= 47,838.06 - 33,797.17$) as total variation, 93.87% ($= 13,180.52/14,040.89$) is due to differences in curve levels and 4.79% ($= 672.37/14,040.89$) to differences in the form of the curves.

The manager of the process can use the results of this analysis to guide the efforts to improve the uniformity of the VDP. In this case, resources should be focused on reducing the variation in the level of the profile from board to board. Of course, process

knowledge is needed to know what factors are responsible for controlling the different aspects of the VDP.

4. Comparing Groups of Curves

4.1 The Model

The ideas presented can be generalized when more sources of variation are present. Suppose that there is interest in comparing m groups of curves observed at n common points. Each group consists of m_i curves ($i = 1, \dots, m$) for a total of $M = \sum_i m_i$ curves. A suitable model to describe the data would be:

$$y = g(x) + \mathbf{Group} + g(x, \mathbf{Group}) + \mathbf{Curve}(\mathbf{Group}) + g(x, \mathbf{Curve}(\mathbf{Group})) + \varepsilon. \quad (2)$$

The first term represents the overall effect the covariate, **Group** is the differential effect the group, and $g(x, \mathbf{Group})$ allows the form of the group profiles to vary. The term **Curve(Group)** represents a nested effect that allows the levels of the curves to vary within each group and $g(x, \mathbf{Curve}(\mathbf{Group}))$ is a term that account for differences in form within groups. Table 3 presents the ANOVA decomposition of model (2) which is similar to the analysis of covariance of a split-plot design with **Group** being the whole plot.

Additionally, the sum of squares for **Curve(Group)** can be divided into m portions with $m_i - 1$ edf, each portion due to differences in levels within a group. Similarly, the sum of squares for $g(x, \mathbf{Curve}(\mathbf{Group}))$ can be subdivided into m portions with $p(m_i - 1)$ edf assessing the differences in form within each group. This part of the analysis is identical to the one described in the previous section carried out within each group of curves.

Table 3. ANOVA table to compare groups of curves

Source	Edf	SS
Covariate (x)	p	
Group (level between group)	m-1	SS(Group)
x, Group (form between group)	p(m-1)	SS(x, Group)
Curve(Group) (level within group)	$\sum_{i=1}^m (m_i - 1)$	SS(Curve(Group))
x, Curve(Group) (form within group)	$\sum_{i=1}^m p(m_i - 1)$	SS(x, Curve(Group))
Residual	M(n-p-1)	SSE
C Total	Mn-1	

Approximate tests of hypotheses can be constructed using ideas from mixed models. For example, if the curves within groups are considered random, then the appropriate tests for differences in average group level and differences in form per group would be

$$F_{\text{Group}} = \frac{MS(\text{Group})}{MS(\text{Curve}(\text{Group}))} \quad \text{and} \quad F_{x, \text{Group}} = \frac{MS(x, \text{Group})}{MS(x, \text{Curve}(\text{Group}))},$$

respectively.

In order to conduct tests within groups, SS(Curve(Group)) and SS(x, Curve(Group)) need to be divided into components for each group. For that purpose, an analysis like the one presented in Section 3.2 needs to be carried out for each group. If the SSE's for each group are similar, then the pooled SSE from Table 3 can be used to construct the tests for all groups.

4.2 Example

The VDP data used in Section 3.3 is part of a larger set that includes data on two additional shifts ("Groups"). The additional two shifts consist of eleven and four boards ("Curves"), respectively. Thus, the full data set consists of 24 profiles. Using only the first three terms in model (2) produces a smoothed profile for each of the three shifts. These appear in Figure 4. The similarity of these profiles, both in form and level, suggests that the shift to shift variation is small.

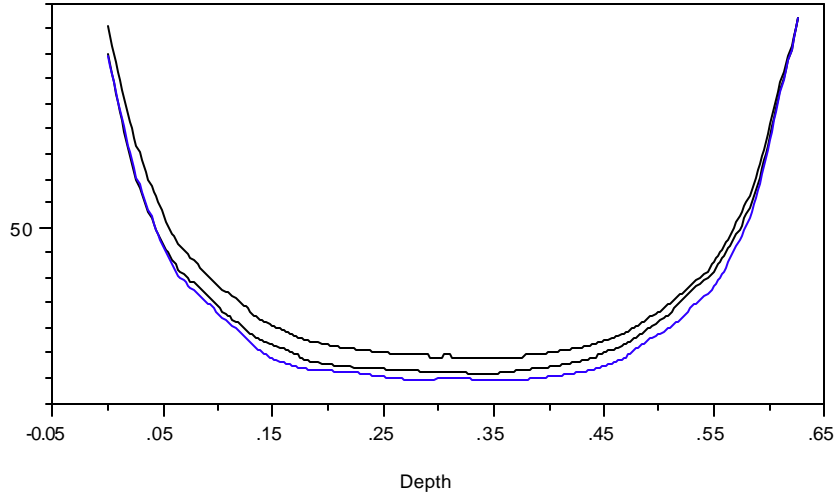


Figure 4. Smoothed VDP's for the three shifts. Vertical axis is density in lbs/ft³.

Fitting the full model (2) is equivalent to fitting smoothers individually to each of the 24 profiles. Figure 5 shows the fitted profiles using the full model and Table 4 presents the associated ANOVA table. The statistics

$$F_{\text{Shift}} = \frac{420.42}{939.66} = 0.45 \quad \text{and} \quad F_{\text{Depth,Shift}} = \frac{5.20}{10.40} = 0.50,$$

test the differences among shift levels and shift forms, respectively. Neither is significant, suggesting that, on the average, the profiles are the same for the three shifts. The conclusion is that shifts are not a significant source of variation.

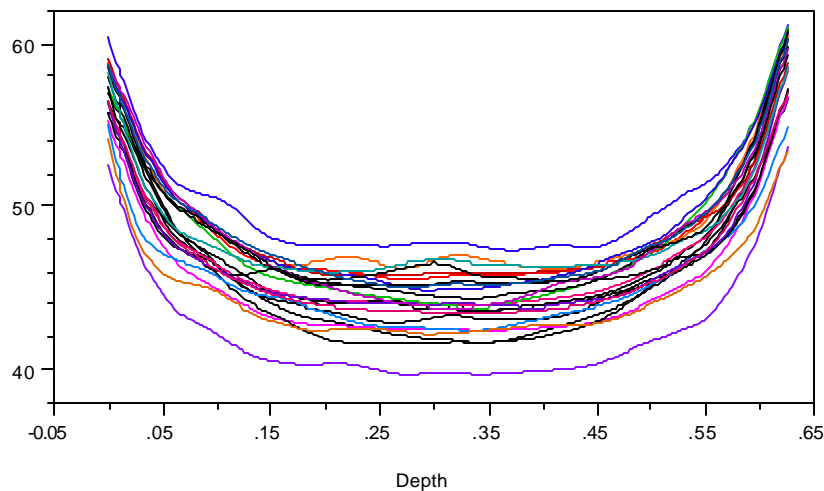


Figure 5. B-splines fitted individually to 24 profiles. Each curve has 16 df. Vertical axis is density in lbs/ft³.

The analysis within each shift is done similarly to that of shift A presented in Section 3.3. Fitting the three shifts individually, the mean square errors were 0.0703, 0.0601 and 0.0560, so the pooled MSE = 0.06 from the full model was used for the tests. Given the amount of degrees of freedom available, all the F-tests are highly significant. A breakdown of the percent overall variation (after adjusting for Depth) explained by the different components appears in the last column of Table 4.

Table 4. ANOVA table of 24 profiles in three shifts

Source	edf	SS	MS	%
Depth	16	86,871.14		
Shift	2	840.84	420.42	3.41
Depth*Shift	32	166.26	5.200	0.67
Board(Shift)	21	19,732.93	939.66	
Shift A	8	13,180.52	1,647.57	53.40
Shift B	10	5,711.60	571.16	23.14
Shift C	3	840.81	280.27	3.41
Depth*Board(Shift)	336	3,492.75	10.40	
Depth*Shift A	128	672.37	5.25	2.72
Depth*Shift B	160	2,381.15	14.88	9.65
Depth*Shift C	48	439.27	9.15	1.78
Residual	7128	450.97	0.06	1.83
C Total	7535	111,554.9		

The last column of Table 4 indicates that 53.4% of all the variation in the data (after adjusting for the covariate, Depth) is generated by differences in the levels of the profiles within shift A. The next largest source of variation (23.14%) is between levels of the profiles in Shift B. Figure 6 shows the fitted profiles for shifts B and C, which should be compared with those for shift A that appears in Figure 3c.

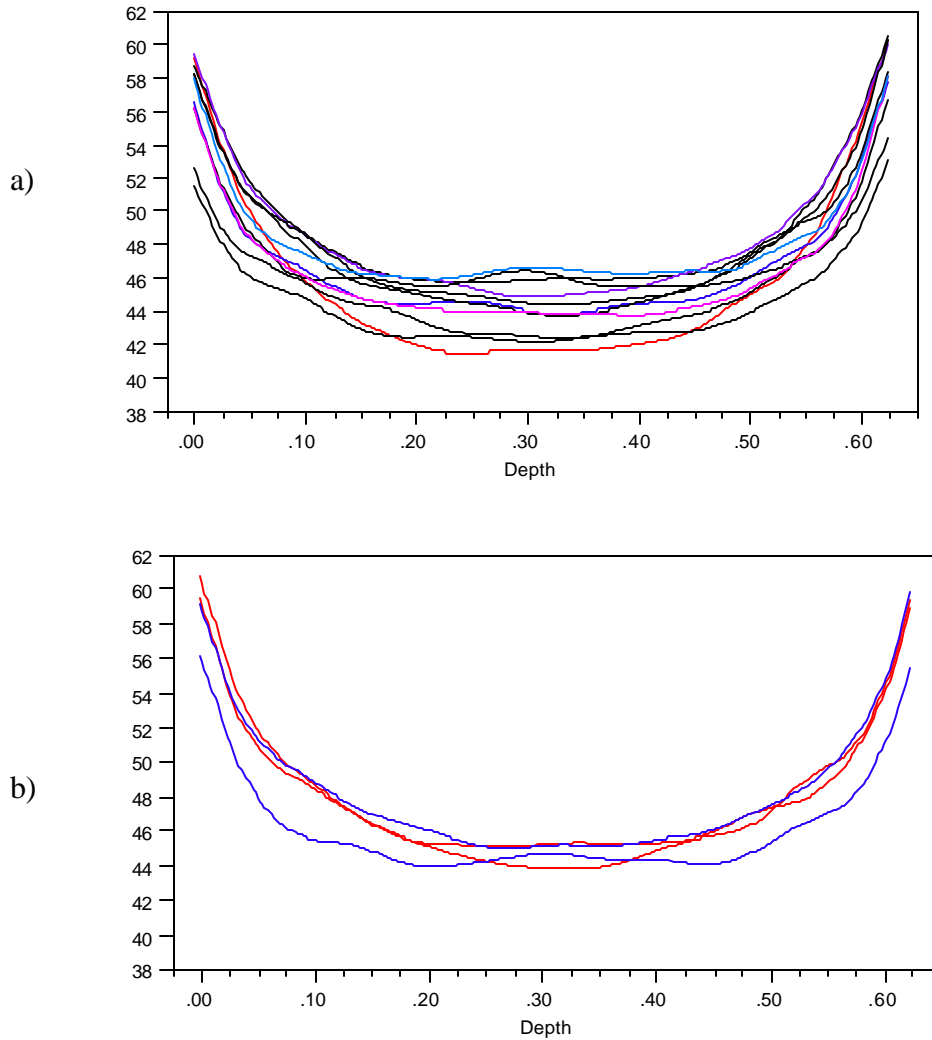


Figure 6. Fitted profiles. a) Shift B. b) Shift C. In both plots the vertical axis is density

Summary and Conclusions

This paper shows a relatively simple and intuitive way to analyze variation when the responses are curves. Generalized additive models are used to fit complicated curves in a nonparametric fashion and to assess the differences between curves. Since the terms enter in an additive way, these models are similar to linear models. Furthermore, the proposed technique can be carried out with available commercial software.

The procedure presented is applicable when all the curves are measured at the same locations and the level of smoothing is the same for all the curves. These restrictions are compatible with most data sets that consist of curves generated in industry.

The procedure does not take into consideration the presence of autocorrelation, which is almost always present in data that are observed within very small intervals of time or space. Indeed, the VDP data analyzed in this paper exhibited a fairly high amount of positive autocorrelation. Ignoring autocorrelation affects the selection of the "optimal" smoothing level of the curve. Specifically, in the presence of positive autocorrelation, the common selection criteria (e.g. GML and GCV) tend to underestimate the amount of smoothing, producing curves that are too "wiggly" (Wang, 1998a).

Wang (1998a, 1998b) suggests a method that estimates simultaneously the level of autocorrelation and the amount of smoothing. Wang (1998a) provides code for fitting this model with smoothing splines using SAS's MIXED procedure. Even for single profiles (each consisting of 314 points) this method required a large amount of CPU time and did not always converge. It also requires considerable knowledge of SAS programming.

Other methods have been proposed to accommodate anomalies in the data such as autocorrelation and heteroscedasticity. Some of these methods even allow the level of smoothing to vary for different curves (Brumback and Rice 1998, Verbyla, et. al., 1999, Lin and Zhang, 1999). For the most part, these models are fairly complex and require programming on the part of the user.

However, the very same fact that the points are taken so frequently, allows for the selection of adequate levels of smoothing by overlaying the smoothed curve on the observed points. Given the short distance between points, it is relatively easy to pick a curve (i.e. level of smoothing) that does not follow the local patterns of the data. Furthermore, since the main objective is to compare curves, the results tend to be fairly robust to changes in the amount of smoothing. The conclusions were the same when the analyses were run using 10 and 30 edf.

For the VDP data, the results suggest that there are small differences among the shifts and that most of the variation occurs in curves within shifts. In particular, more than half of the variation within shifts is due to differences in the level of the profiles in shift A. Computing proportions of variation after adjusting for the covariate give valuable insight into the magnitude of the sources of variation.

All the results were obtained by fitting models readily available in S-Plus and the TRANSREG procedure in SAS. B-splines were used as smoothers because the fit was

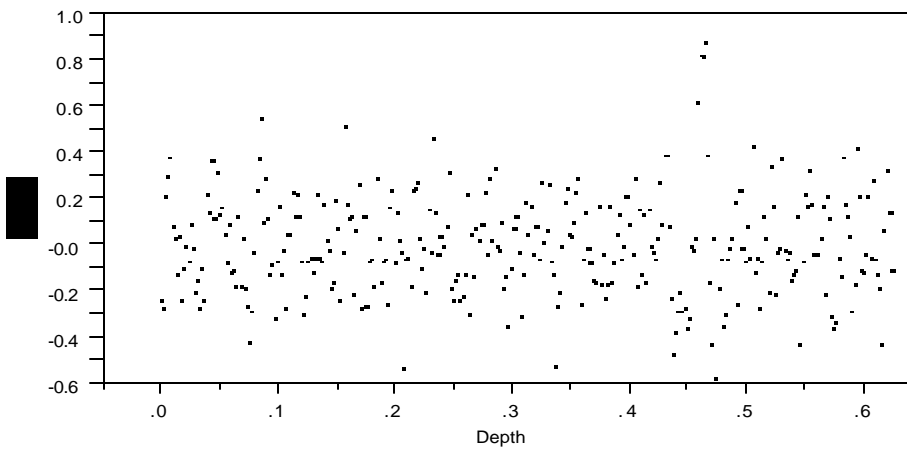
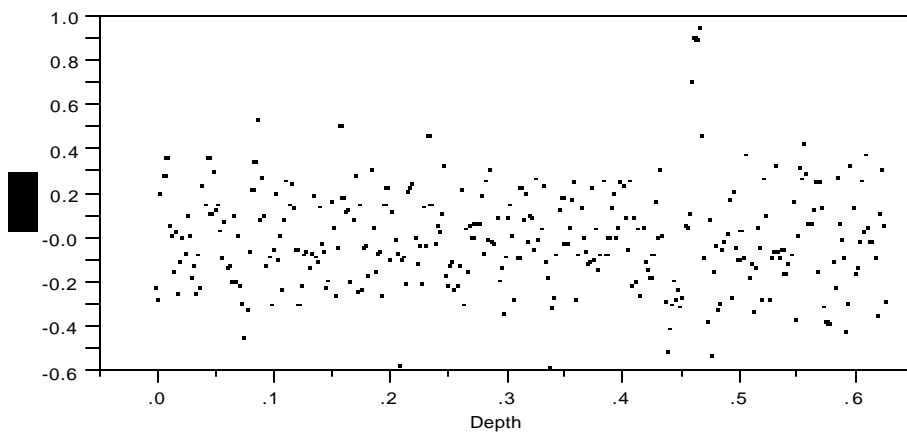
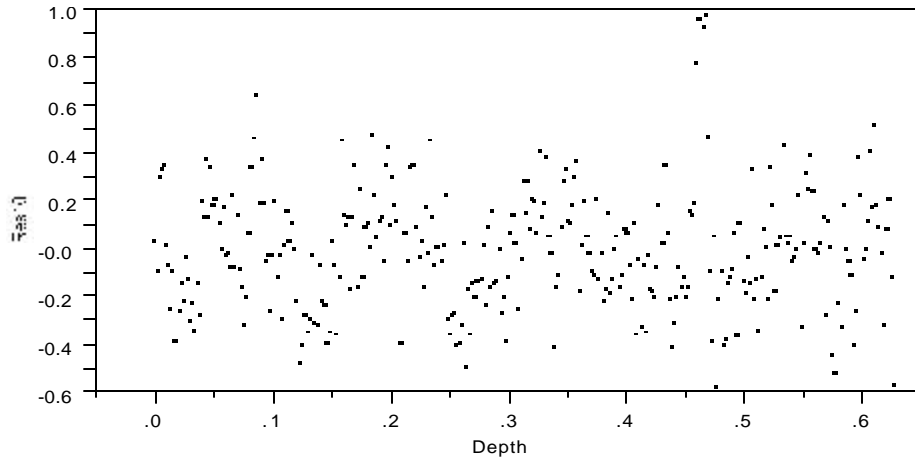
adequate with a relatively low number of edf and the results were correct when interactions were included. Theoretically, any type of smoother can be used as long as the software accommodates them, particularly with respect to interactions. Currently, the `gam` function in S-Plus only allows the use of B-splines and natural splines to perform analyses involving interactions with categorical regressors.

Acknowledgements

The authors want to thank Dr. Robert Mee for his help in the technical aspects of this paper and the support received from the Statistical Engineering Service, Center for Executive Education, University of Tennessee, Knoxville.

Appendix 1

Residual plots for the profile in Figure 2 for different degrees of freedom using B-Splines. From top to bottom: 10, 16, and 20 degrees of freedom.



Appendix 2

The models used in this article were run using S-Plus 2000 (1999) using the `gam` function. They may also be fitted in SAS using the TRANSREG procedure. In both S-Plus and SAS, B-splines were used because interactions involving other nonparametric smoothers are not supported. The TPSPLINE procedure in SAS (experimental in version 7) fits additive models and computes confidence intervals, but does not support interactions or the use of categorical variables.

The following commands can be used after the data are already entered into the appropriate data structure (S-Plus data frame or SAS data set). The response variable is VDP; and the explanatory variables are Depth (values 0 to 0.626 in 0.002 inch increments), Shift (values A, B, C, defined as a "factor" in S-Plus), and Board (defined as a "factor" in S-Plus, with values A1-A9 for shift A, B1 -B11 for shift B and C1-C4 for shift C). The complete data table consists of $24 \times 314 = 7,536$ data points and is available at <http://web.utk.edu/~ewalker/VDP/VDP.TXT>. To construct Table 2, the following models were used using the data from shift A only:

```
model.1 <- gam(VDP ~ bs(Depth, 16))
model.2 <- gam(VDP ~ bs(Depth, 16) + Board)
model.3 <- gam(VDP ~ bs(Depth, 16) + Board + Board:bs(Depth, 16))
```

The fitted values from `model.2` and `model.3` are shown in Figures 3b and 3c, respectively.

Using S-Plus notation, `model.3` may be written more compactly as

```
model.3 <- gam(VDP ~ bs(Depth, 16) * Board).
```

The same results can be obtained in SAS by invoking the following commands:

```
proc transreg;    title "Model 1";
  model ide(VDP) = bspline(Depth / nknots=13)/test;

proc transreg;    title "Model 2";
  model ide(VDP) = bspline(Depth / nknots=13) class(Board)/test;

proc transreg;    title "Model 3";
  model ide(VDP) = bspline(Depth / nknots=13)|class(Board)/test;
```

The number of knots in the TRANSREG procedure corresponds to edf minus 3 in S-Plus.

The five primary models to create Table 4 are shown below. The submodels of models 4 and 5 may be fit by running each shift separately as in Table 2.

```
model.1 <- gam(VDP ~ bs(Depth, 16))
model.2 <- gam(VDP ~ bs(Depth, 16) + Shift)
model.3 <- gam(VDP ~ bs(Depth, 16) + Shift + Shift:bs(Depth, 16))
model.4 <- gam(VDP ~ bs(Depth, 16) + Shift +
               Shift:bs(Depth, 16) + Shift/Board)
model.5 <- gam(VDP ~ bs(Depth, 16) + Shift + Shift:bs(Depth, 16)
               + Board%in%Shift + bs(Depth, 16):Board%in%Shift)
```

Figures 4 and 5 are obtained by plotting the fitted values from model.3 and model.5. The last three models are written more compactly below. The compact notation uses standard S-Plus shorthand ("*" for main effects plus interaction), but it also takes advantage of the fact that each board has a unique identifier. In the case of model.5, the compact specification is not only easier to write, it also runs much faster due to the way S-Plus creates the model matrix.

```
model.3 <- gam(VDP ~ bs(Depth, 16) * Shift)
model.4 <- gam(VDP ~ bs(Depth, 16) * Shift + Board)
model.5 <- gam(VDP ~ bs(Depth, 16) * Board)
```

The following commands in SAS produce the same results:

```
proc transreg;    title "Model 1";
  model ide(VDP) = bspline(Depth / nknots=13)/test;

proc transreg;    title "Model 2";
  model ide(VDP) = bspline(Depth / nknots=13) class(Shift)/test;

proc transreg;    title "Model 3";
  model ide(VDP) = bspline(Depth / nknots=13)|class(Shift)/test;

proc transreg;    title "Model 4";
  model ide(VDP) = bspline(Depth / nknots=13)|class(Shift)
                  class(Board)/test;

proc transreg;    title "Model 5";
  model ide(VDP) = bspline(Depth / nknots=13)|class(Board)/test;
```

References

- Bisgaard, S. and Steinberg, D. M. (1997), The design and analysis of 2^{k-p} prototype experiments, *Technometrics*, 39, 52-62.
- Brumbeck, B. A. and Rice, J. A. (1998), Smoothing spline models for the analysis of nested and crossed samples of curves, *Journal of the American Statistical Association*, 93, 961-675.
- Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.
- Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer, New York.
- Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.
- Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.
- King, E. J., Hart, J. D. and Wehrly, T. E. (1991), Testing the equality of two regression curves using linear smoothers, *Statistics and Probability Letters*, 12, 239-247.
- Lin, X. and Zhang, D. (1999), Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society, B*, 62, Part 2, 381-400.
- Miller, A. and Wu, C. F. J. (1996), Parameter design for signal-response systems: A different look at Taguchi's dynamic parameter design, *Statistical Science*, 11, 122-136.
- Morrison, D. (1990), *Multivariate Statistical Methods, 3rd. Edition*, McGraw-Hill, New York.
- S-PLUS User's Guide, (1997), Data Analysis Products Division, MathSoft, Seattle, WA.
- Taguchi, G. (1987), *System of Experimental Design*, Kraus International; Publications, White Plains, New York.
- TPSPLINE Procedure, (1998), SAS Institute,
<http://www.sas.com/rnd/app/da/new/expdoc/chap8/index.htm>
- Verbyla, A. P., Cullin, B. R., Kenward, M. G., and Welham, S. J. (1999), The analysis of designed experiments and longitudinal data by using smoothing splines, *Applied Statistics*, 48, 269-311.

- Wabha, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia: SIAM
- Wang, Y. (1998a), Smoothing Spline Models With Correlated Random Errors, *Journal of the American Statistical Association*, 93, 341-348
- Wang, Y. (1998b), Mixed-Effects Smoothing Spline, *Journal of the Royal Statistical Society*, 60, 159-174.
- Winistorfer, P. W., Young, T. M. and Walker, E. (1996), Modeling and comparing vertical density profiles, *Wood and Fiber Science*, 28, 133-141.

Key Words: *Profiles, signatures, smoothing, nonparametric regression, splines, B-splines, ANOVA, semiparametric models.*