2

Some Functional Datasets and Associated Statistical Problematics

This chapter could have been entitled "Some statistical problematics and associated functional data". In fact, there are many nonparametrical statistical problems which occur in the functional setting. Sometimes, they appear purely in terms of statistical modelling or, on the contrary, they can be drawn directly from some specific functional datasets. But, in any case, the proposed solutions should look at both points of view. This chapter describes various functional data with their associated statistical problems. Of course, some statistical processing (such for instance unsupervised classification) concerns all the following datasets but additional informative data (such for instance the knowledge of some response variable) can lead to particular problems. As we will see, these data have been choosen to cover different applied statistics fields, different shapes of curves (smooth, unsmooth), various grids of discretization (fine, sparse) and also different types of statistical problems (regression, discrimination, prediction and classification). Obviously, the methods presented later on will concern many other functional datasets while at the same time, these datasets can motivate other statistical problems not investigated here. Although these functional datasets are available on various websites, we give them in the companion website (*http://www.lsp.ups-tlse.fr/staph/npfda*) in which they are presented in a appropriate format, directly usable by readers both for familiarizing themselves (by reproducing the examples) with the functional nonparametric methods described in this monograph and eventually to compare them with their own alternative approaches.

2.1 Functional Chemometric Data

Spectrometry is a modern and useful tool for analyzing the chemical composition of any substance. As pointed out by [FF93], in order to analyze such kind of data, "chemometricians have invented their own techniques based on heuristic reasoning and intuitive ideas". The two most popular methods are partial least squares (see [W75] and [MNa89] for more details), and principal

component regression ([M65]). As chemometrics was a starting point for developing the functional nonparametric methodology, they play a major role in this book, and it is natural to begin by the presentation of such dataset.

2.1.1 Description of Spectrometric Data

The original data come from a quality control problem in the food industry and can be found at *http://lib.stat.cmu.edu/datasets/tecator*. Note that they were first studied by [BT92] using a neural networks approach. This dataset concerns a sample of finely chopped meat. Figure 2.1 displays some units among the original spectrometric data.

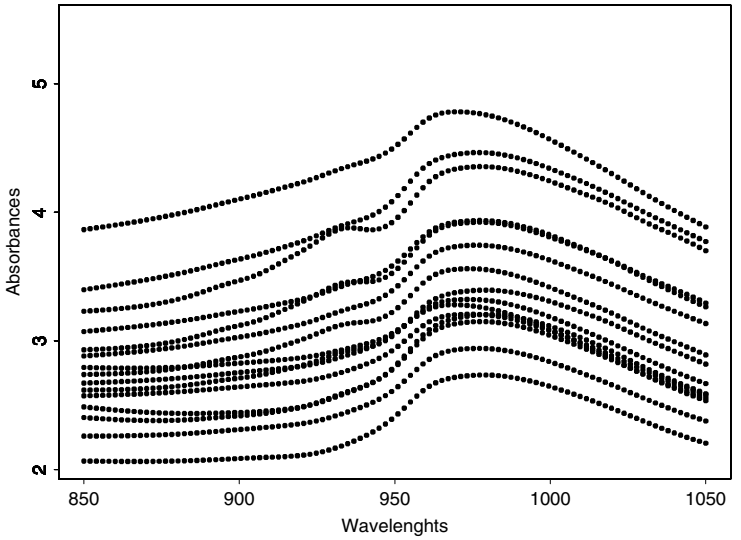

Fig. 2.1. Original Chemometric Data Concerning 15 Subjects

This figure plots absorbance versus wavelength for 15 randomly selected pieces of meat. More precisely, for each meat sample the data consists of a 100 channel spectrum of absorbances. Absorbance is the $-log_{10}$ of the transmittance measured by the spectrometer, and the data were recorded on a Tecator Infractec Food and Feed Analyzer working in the wavelength range 850-1050 nm by the near-infrared (NIR) transmission principle. One unit appears clearly as a discretized curve. Because of the fineness of the grid (spaning the discretization), we can consider each subject as a continuous curve. This was