Data Acquisition, Processing and Analysis for Distributed Decision Support

Broad Contents

Prologue

This Monograph is a reflection of years of experience the authors had in their laboratory (Division for PIE and NDT Development, Indira Gandhi Center for Atomic Research, Kalpakkam, TamilNadu, India) analysing data obtained from various sources in different formats to arrive at crucial decisions.

Though the authors are predominantly concerned with data of non-destructive testing (NDT) origin, the methods and processes described herein can be applied to any domain of interest. The methods will be particularly suitable for solving problems in empirical domains of discourse, such as seismology, medicine, finance, biotechnology, geology, remote sensing, etc. and of course in NDT.

However, while the acquisition, processing and analysis methods will have a lot in common among various domains, while interpreting the results, the readers must be aware that the expertise of the problem-domain must be taken into account, in order to arrive at an acceptable decision.

A subject so diverse as data analysis and processing will have a number of facets and approach points, i.e., the "treatment" of the subject. The approach adopted by the authors of this monograph has its roots in "decision-making through classification." More emphasis has been given to the practical aspects of acquiring data, classifying data, and arriving at a suitable action or decision, on the basis of the classified data.

This is **not** a textbook on the Subject. Rather, this monograph should be viewed as practical field-notes on the Sections mentioned in the Contents page. Hence, normal academic topics pertaining to, say, *noise reduction* has not been dealt with in detail here. This is also true with topics such as *digital filter design*, *linear time-invariant systems*, *analog signal processing*, etc. Readers who wish to know more on these topics are guided to other excellent resources for these topics [1].

That being said, the topics listed are described comprehensively, with explanations, practical tips and figures, where applicable. Each Chapter also has a set of open questions at the end, aimed at kindling the thought-process of the reader on the topic discussed in the Chapter. While discussing the methods of data processing, analysis and decision-support, relevant code snippets are included to make the discussion complete. Readers are encouraged to verify the code first, and then directly apply and use the code in their projects, with appropriate reference to this Monograph. The authors assume no responsibility for any loss that may arise due to the application the code given here, without due diligence.

While describing the data analysis methods in this Monograph, examples and code-snippets are provided from the DESKPACK Software System (DSS). More information about the DSS is available at http://deskpack.tripod.com. Two-dimensional data analysis (image processing and analysis) has been carried out using the Image Tool software. This is a *free* and *powerful* software, which you can download and use from http://ddsdx.uthscsa.edu/dig/itdesc.html.

The Monograph also includes a glossary of technical terms used herein.

Happy Reading!

Chapter 1

Definition of Data

1.1 Introduction

This Chapter gives a gentle introduction to data and their representation in the form of signals and images. A more rigorous treatment on data acquisition is given in the next Chapter.

Data is any useful value, which is usually measured as an outcome of a physical process. If the physical process continuously produces useful values, we will have a continuous string of data. These physical processes could be either natural or artificial. Example physical processes include the vibration amplitude of a string, heartbeat amplitudes, the intensity of emission of radio waves by astronomical bodies, amount of rainfall at a particular place, the market value of a certain stock and so on. As you would have observed, one feature that is common in these examples is that each of these values change as a function of *time*. The measured values can also change with respect to any other variable. So, there is one measured *value*, which changes with respect to another *variable*, which could be time. It is possible to plot the *variable* and the *value* in the form of an X-Y graph. In such a graph, the variable is plotted in the X-axis and the measured value is plotted in the Y-axis. The X-axis is called the independent variable and the Y-axis, the dependent variable. Such processes could be represented mathematically as follows:

$$y = f(x) \qquad \text{(Equation 1.1)}$$

where y could be the market value of a certain stock and x the time or day. This equation is read, as "y is a function of x."

Useful data can also be obtained from space, say from an area, which can be represented to be with X- and Y-coordinates. For example, if we observe an x-ray film, the intensity of the picture will vary from one point in the film to another. For every point in the film, we can assign an X- and Y-coordinate, and an associated intensity at that point. If 'I' is the intensity, we can then write

$$I = f(x,y) \qquad \text{(Equation 1.2)}$$

Other examples of such data include the temperature distribution on a surface, amount of rainfall in a state, thickness of a paper at various points, elevation above mean sea level of a geographical area, and so on.

We say that the equation 1.1 represents one-dimensional data or a *signal*, whereas equation 1.2 represents two-dimensional data or an *image*.

1.2 Signals and Images – A Few Examples

We understand that a signal is a representation of a measured value with respect to time, while an image is a representation of a measured value with respect to X- and Y-coordinates. Let us see a few examples of signals and images, before we try to understand their characteristics. Figure 1.1 shows a typical signal.

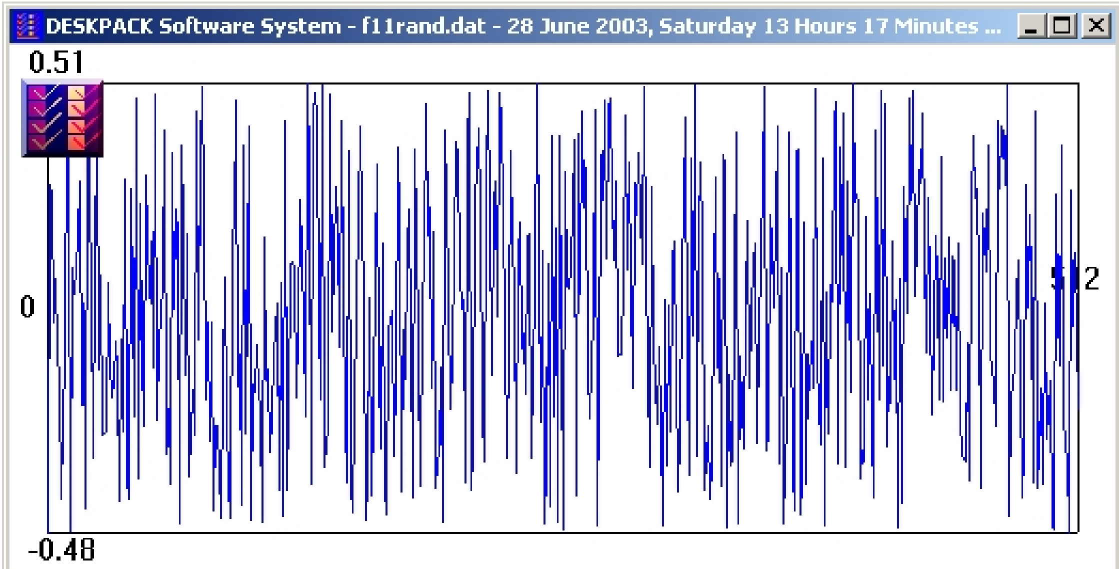

Figure 1.1 A Typical Signal

Before we delve into the characteristics of the signal shown in Figure 1.1, we notice that it has 512 measured values, and that the values vary from about -0.48 to +0.51. The measured values seem random in nature, as they do not show any specific pattern. One important characteristic of any physically measured value, as in the signal shown Figure 1.1, is that for any given 'x' (say, time instant), there is a unique and single 'y' value.

Now, let us see how a typical image looks like. Figure 1.2 shows a typical image.

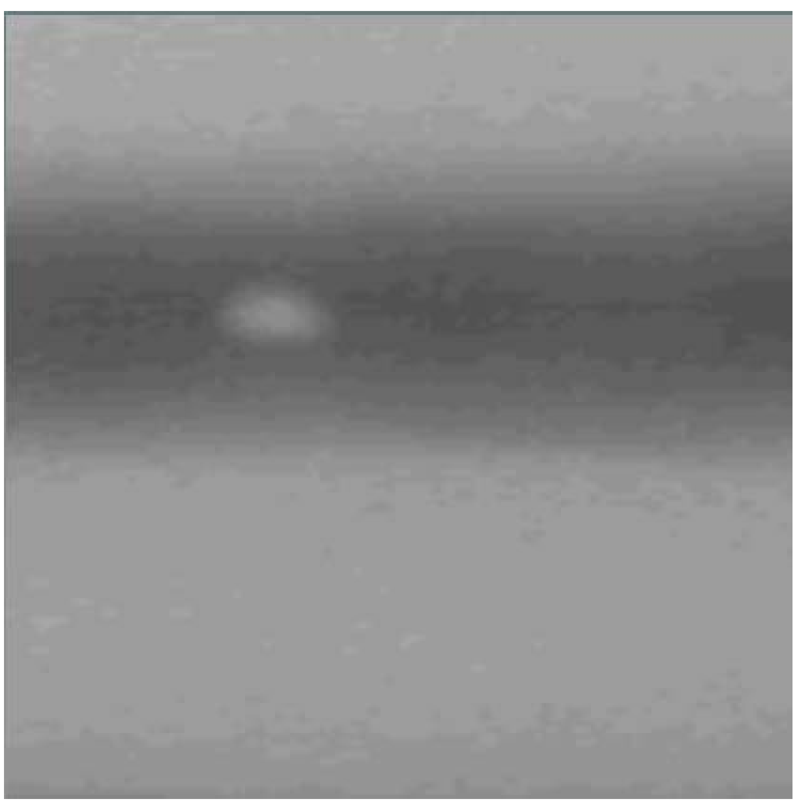

Figure 1.2 A Typical Image

Figure 1.2 shows a typical *gray*-scale x-ray image. In a gray-scale image, the intensity of the picture at a point (x,y) is represented in a scale that varies from fully white to fully black, depending upon the number of bits used. Occasionally, the various shades in a gray-scale image are also represented using pseudo-colours. Figure 1.3 shows such an image.

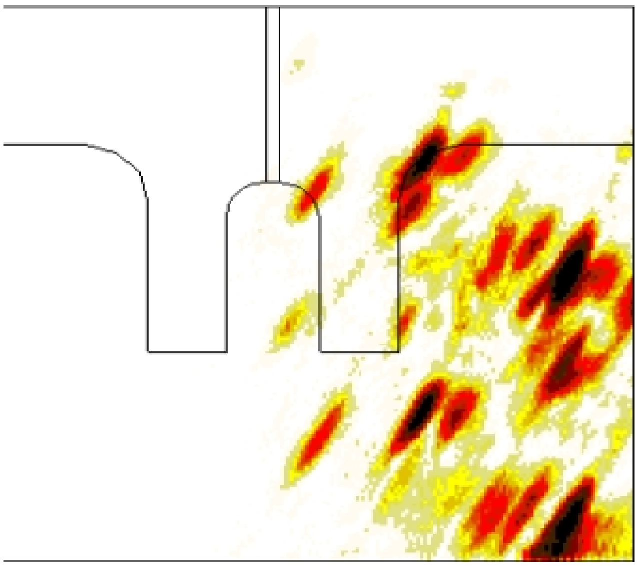

Figure 1.3 A Typical Pseudo-Colour Image

It is possible to remove these pseudo-colours and convert the image into a gray-scale image. Figure 1.4 is the gray-scale converted image of Figure 1.3.

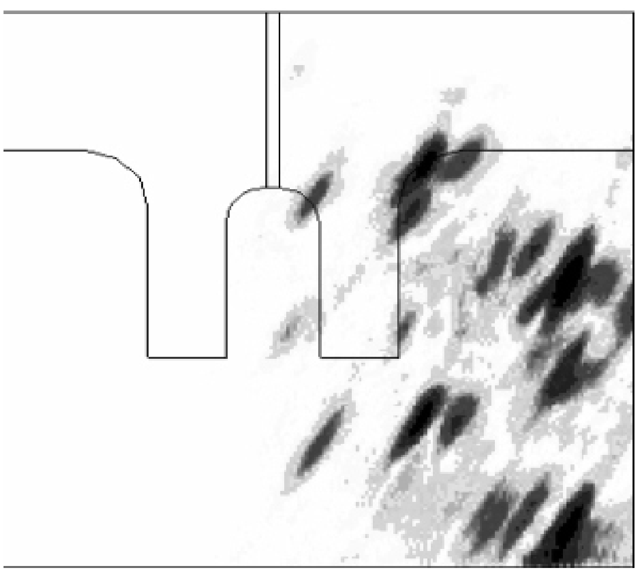

Figure 1.4 Gray-scale converted image of Figure 1.3

Because the same image (and hence, the same "source" from which the image has been obtained) can be represented either as a pseudo-colour image or a gray-scale image, extra care is needed while processing, analysing and interpreting these images. A good knowledge about the methods adopted while obtaining the image is crucial for the accurate interpretation of the image. This is also true in the case of signals, where the parameters of signal "acquisition" can alter the way the signal looks and interpreted. In order to understand the data obtained from these perspectives, we shall look into some of the characteristics of signal and image data in the following sections.

1.3 Characteristics of Signals

Let us now take a careful look at the Figure 1.1, which represents a random signal. Here are some observations:

- The signal has 512 points – we say that the record length of the signal is 512. In other words, if the first point in the X-axis is represented by x(0), the last point is represented by x(511), taking the total number of points to 512
- The signal varies from about –0.48 to +0.51 – we then say that the range of the record is +0.51 – (-)0.48) or about 1.0
- For every x(n), there is an unique y(n), where y(n) is the measured value
- As we move from x(0) to x(511), we observe that the values y(0) thru y(511) do not follow any particular order or sequence – they seem to be varying at random.

In contrast, take a look at Figure 1.5, which represents another type of signal – a sinusoidal signal.

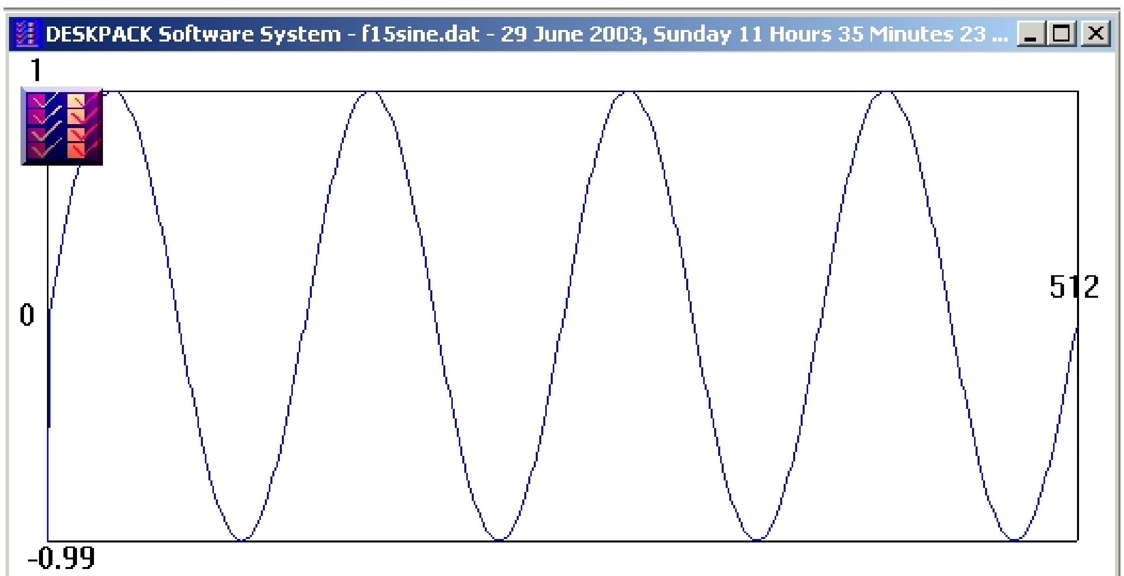

Figure 1.5 A typical Sinusoidal Signal

Figure 1.5 looks somewhat familiar. Though this signal too has 512 points, its lower bound is –0.99 and its upper bound is 1.0. More importantly, the y(n) values in Figure 1.5 show a pattern – the values are "periodic". One can observe that this single record of the signal (comprising of 512 points) has four crests and four troughs – each pair of crest and trough forming one period of the signal. Such signals are called periodic signals.

Signals in the real world could be anywhere from a pure sinusoidal wave as in Figure 1.5 to a "pure" random wave, as in Figure 1.1.

It is quite possible to have more crest-trough pairs within the 512-point signal record. Also, the signal record length can vary too – they could be 64, 128, 256, 512, 1024, 2048, 4096 points and so on.

1.4 Characteristics of Images

One characteristic that will distinguish an image from a signal is that the measured value 'y' is a function of both X- and Y-coordinates, as shown in Equation 1.2. Each measured value 'y' is digitally represented by a number of bits (each bit is either a '0' or a '1'). The number if bits used is also called the depth of an image. In normal digital images, each 'y' value could have a depth of 8, 12, 16 or 32 bits. If 12 bits are used to represent a 'y' value, 2^{12} gray-levels are possible.

Occasionally, for the sake of computational efficiency, a gray-scale image could be converted into a "binary" image (or just a black and white image). Figure 1.6 shows a typical gray-scale image, while Figure 1.7 shows its binary counterpart.

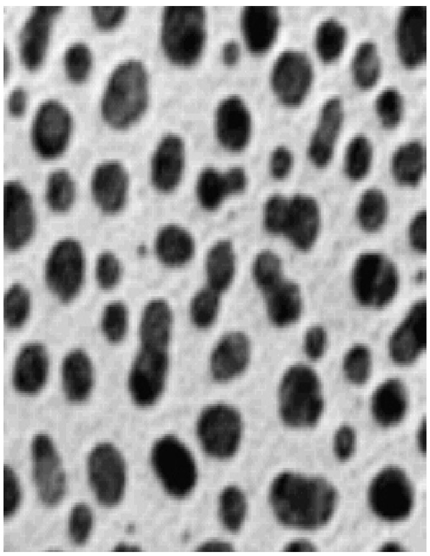

Figure 1.6 A typical gray-scale image of "blobs"

Figure 1.7 The binary image of the gray-scale image shown in Figure 1.6

One good use of converting a gray-scale image into a binary image is to "measure" the properties of the objects present in the image. In Figure 1.7, for example, it could be worthwhile to measure the average diameter (or the lengths of major and minor axes) of the blobs present.

It must be remembered that both signals and images are represented in a computer as files, which contain the measured values. For example, the random signal shown in Figure 1.1 is stored as ASCII (American Standard Code for Information Interchange) or "text" values in a file that has a form as shown in Figure 1.8. Note that the values are stored in a column format in a data file. This is not the only way to store information, but is one of the more predominant ways of storage.

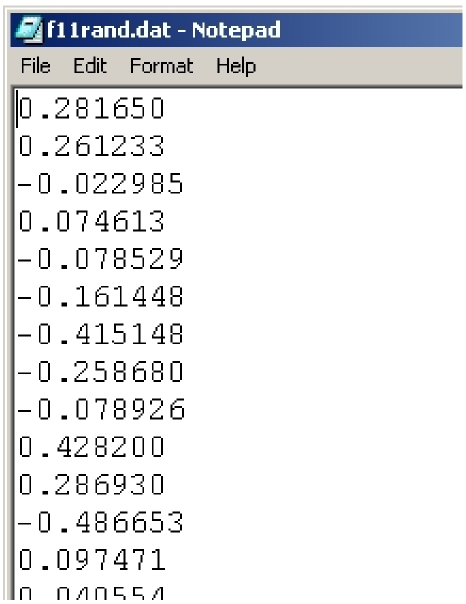

Figure 1.8 Typical format of data (in this case, a signal) storage in a computer file

One can adopt the same method to store the data of an image, in a computer file. However, in most cases the measured values in the case of an image are stored in an entirely different manner in a computer file. This is due to the fact that image data are large and hence occupy more space. Hence, compressing the data for image storage makes sense. Compression also helps in easier and quicker transmission of data from one computer to another. Depending upon the type of compression used, the image data file has an extension *.bmp (bitmap file), *.jpg (a JPEG file) or *.tif (a TIFF file). The JPEG is becoming an industry standard for storing image files. Compressed data files such as a JPEG file is called a "binary" file (as opposed to a text file shown in Figure 1.8), which can be properly read and rendered only by a machine.

Knowledge of how data is stored in a computer is thus an essential skill, for processing, analysing and correctly interpreting the data.

In this Chapter, we have walked through the first few steps in describing and recognising data in the form of either signals or images. Working on these fundamental ideas, we shall find out more about data acquisition and representation in the next Chapter.

1.5 Questions

- Is there a data "depth" (number of bits used to represent a measured value) for signals too?

- Why should the record length be 32, 64, 128, 256, 512, 1024 etc., i.e., 2^N where N is an integer?
- What is the advantage of representing image data in pseudo-colours?
- Is there a way to represent three-dimensional data?
- Can you give one example of a three-dimensional data?
- Like a record length for a signal, is there a preferred width and length for an image?
- What could be the process that converts a gray-scale image in to a binary image?
- Is it possible to convert a binary-image back to a gray-scale image?
- What physical processes could lead to a random signal? A periodic signal?
- In Figure 1.5, how to measure the period of the signal shown using a classroom-ruler? What information is missing that would give the value of the period of the signal shown?
- Is there any advantage in viewing Figure 1.1 (a random signal) as an image?
- In Figure 1.8, the storage format of the signal shown in Figure 1.1 is given. Figure 1.1 has both Y-values and X-values. Whereas Figure 1.8 has only one column. Is this Y or X values? Where's the other set of values?
- Try opening a *.jpg file in Notepad and examine the contents. Caution: Do **NOT** save the contents.