Fundamental Problems with On-line Data Streams

Peter R. Lord

Raltex Inc.

3116 Monticello Drive

Raleigh, NC 27612

Introduction

Two major problem in on-line data acquisition are (1) dealing with the prodigious flow of data and (2) being able to predict the end result. The problem applies to a large range of parameters which might be measured but, in the present case, only data obtained by measuring the linear density of yarn in an on - line mode will be discussed. The data were obtained by Cotton Incorporated and this paper merely discusses some of the problems, and some possible solutions for dealing with the problems.

Data Acquisition

Scientists and technologists frequently use on-line transducers to produce streams of data which are sampled at high frequencies to produce time series or data in other formats. The data is transmitted as a series of data medium. At some time after storage, the data is retrieved and analyzed. Each sample contains X bytes of memory and for a sampling rate of a samples/ sec the information flow is s x bytes/ sec. Consider the implications of this. Many transducers are capable of producing prodigious information flow; values of some 100 MB/sec per channel are not uncommon. In an industrial setting, the continuous use of such an on -line sampling would produce an information flow of $100\times60\times60\times24 = 8.6\times10^{12}$ bytes/day per channel. Not only does such information have to be transmitted and stored, even if only temporarily, it has to be analyzed.

There are a number of ways to manage such vast quantities of data. In an industrial setting, the data are often only sampled: furthermore, only data outside control limits are recorded. The object of the exercise, in this case, is to find trends that endanger production levels or product quality. To set the proper control limits and measure the right parameters is a different matter. Even if the al necessary parameters are tested, a large volume of mind that in a spinning plant, there are rare events that occur only a relatively small number of times per 100,000mhours/ spindle, the analyzed might amount to some 3.6×10^{10}MB [3.6×10^{16} bytes] per channel. To describe a technological process requires that a number of parameters be considered, also some parameters might require several channels to acquire the information needed: such considerations further inflates the requirements.

Even if the data is completely and successfully analyzed, there is a matter of presentation of the results. Of course, there is always the hope that the data can be modeled mathematically so that the end result is a simple equation or perhaps an array of equations. As will be discussed later, random processes play a large role in textiles and a deterministic model alone is usually insufficient for prediction purposes. Such prediction capabilities being the most desirable outcome of any modeling that is undertaken.

Compression

Data compression is a possibility. If the data is compressed, the need for storage space within the computer is reduced but, of course, the files have to be expanded again before the data can be analyzed. Compression engine was placed between the transducer and the input device to the computer: the memory low rate to be handled would be reduced but the compression engine must be able to handle the flow and there would be, of necessity, a processing lag in the in formation flow. Thus there are limits to the help that compression can play in the transport of information from transducer to computer memory. However the gain achieved by conventional in the area of data storage is significant.

Many compression schemes work by reducing the memory taken up by redundant elements in the data string [or strings]. The elements may be at the Boolean level or in higher information structures such as words or geometric patterns. When an element is repeated n times within a data set, an uncompressed version will contain n times the image memory needed for one element in addition to the co-ordinates will be called "pointer memory." Thus the total memory needed for the uncompressed data set is n times the [image + pointer] memory. If the redundancies are removed, the total memory can be reduced to [image +n(pointer)] memory. Two examples will be given at opposite ends of a spectrum of combinations of images and replications.

The first is a case with many images and replications and it relates to a 2,000 word passage from a book and, for illustration, only the redundancies with words will be considered, and even then only the most frequently used words will be displayed. For the purposes of this demonstration, let words be replaced by symbols of one character. It is well known that the word "the" is the most commonly used, but the substitution of a single character for it saves only 2 characters. Table 1 shows extracts from the frequencies of the 35 most frequent words Thus the 193 substitutions shown in Table 1 saved 386 characters, which was 3.1% of the total. If the 35 most frequent words were considered, substitutions saved about 20% in the total characters. If all the words were used, the percentage increases but the exact amount is of no consequence for this analysis. Clearly the amount of redundancy is a matter of the language, subject and author style, but the principle is clear, Increased compression can be achieved at the computer code level where boolean notation is used but the extent is always limited.

Word	Word Frequency	Character Saving/Word	Characters Saved	Percentage Saved
the	193	2	386	3.1
of	116	1	116	0.9
in	97	1	97	0.8
and	84	2	168	1.3
to	61	1	61	0.5
was	43	2	86	0.7
it	36	1	36	0.3
[Data omitted]	etc.	etc.	etc.	etc.
but	10	2	20	0.2
history	10	7	70	0.6
market	10	5	50	0.4
this	10	3	30	0.2
also	9	3	27	0.2
	TOTAL*		2508	20.1

* for the 35 most frequent words

Table 1.

The second case is where a single geometric pattern is replicated many times. The pattern may be printed on paper with either irregular of irregular co-ordinates like wall paper. Taking the irregular placing first. This is at the opposite end of the spectrum mentioned earlier: there being only one image and many replications. If the memory used to describe the image is very large and the pointer memory very small, then there is a great potential for compression. Of images and replications which are coupled with varying potentials for compression. Also there is the potential for compression due to redundancies within the images themselves.

If the pattern replications are periodic, then further "compressions" are possible because the co-ordinates become capable of being described in mathematical terms and the pointer information can be expressed according to some mathematical formula. This thought brings us to the use of the scientific method where the aim is often the reduction of variables to formulae.

Formulation and Prediction

To be able to predict events it is necessary to fit data to formulae as nearly possible. Often an independent group of compared to another to find what relationships exist between them, it becomes possible to predict what outcome will result from the occurrence of a set of independent variables. The relationship may be progressive or periodic. Polynomials are periodic relationships. For instance, some data streams might contain elements of both progressive and periodic variations. If there is no random component, one element could be a trend on which, say, a sinusoidal variation is superimposed. For long-term record, one is more concerned with the periodic variation and the data groupings are usually adjusted to minimize any progressive variation. In practice, there is always some random component and in textile technology that component can be very large.

Consequently, data for long records, such as might be found with on-line testing, is often classified as random or periodic. With periodic data, it is only necessary to describe the details of one period and the multiple positions in space or time: alternatively, a mathematical equation can be used. With random data, probability density curves have to be constructed from past data, from which the probabilities of a result in future data can be calculated [but there is no certainty to the result]. The probability density curves merely give an estimate of the probable outcome. If the probability distribution is similar from one data segment to another, there can be a good predictive potential. If the probability distribution changes along the data stream, the prediction potential is degraded.

The periodic and random component normally have to be separated because different techniques are required to analyze them. Harmonic analysis [such as the use of FFTs] is used to determine the wavelengths or frequencies involved in the periodic motion. Random variations tend to get averaged out by these mathematical processes. Conventional statistical analysis disregards the order of the data in a time series and it is normal to create a histogram [which is a probability distribution curve]. Information is lost in regard to the periodic nature of the data. There is no assurance that just one sort of analysis will fully describe the subject phenomena.

Periodic Variation

The most likely sources of periodic variation is data obtained from textile-related measurements are caused by mechanical errors or periodic motion of machine components. Processing is carried out in a series of machine with progressive drafting along the textile flow path. This means that the final material produced [say, yarn] has components from each of the processes and most of these individual processes elongate some of the errors produced by the mechanical arrangements within the machine concerned and elongate all the errors produced by previous machinery in the line up. Furthermore, errors created by the early machines in the process line have their errors elongated several times before the product emerges from the whole process line. The consequence is that there is an extraordinarily wide spectrum of error wavelengths involved at the yarn stage. Usually, the range might extend from wavelengths measured in millimeters to those measured in megameters. The long errors are rarely measured just because they are too long to handle under normal circumstances. One cause of such long errors is from variations in the bale.

Harmonic analysis techniques involved the interaction between short contiguous segments within the time series and a periodic variable [the periodic variable sometimes being divisions of the entire time series]. To avid aliassing, the sampling must be at least twice per complete cycle of the highest frequency component [i.e., the sampling rate must be higher than the Nyquist value]. The shortest error wavelength might be 10mm, in which case a sample should be taken at least once per 5mm. The minimum sampling frequency would be 1000/5 = 200 samples/meter. Examples of harmonic analysis include modulation and FFTS. Modulation includes wavelets and sinusoidal modulation.

To avoid errors from long-term random variations, sufficient number of replications of the periodic variable must be used. The number of replications of the periodic variable must be used. The number to obtain an acceptable confidence interval depends on the variability of the data, but if the level requires m replications for a given confidence interval, then the longest wavelength that can be safely reported is the length of the time series divided by m. Thus, if we measure 10,000 meters of yarns, the longest error wavelength that will still give the confidence level we require is [10,000/m]. If m were to be 10, the longest reliable error wavelength from the analysis would be 1km. If the sampling rate is 2 samples/ mm of yarn produced, at least 2×10^6 samples would be needed to give acceptable results. This illustrates the difficulties whereby long lengths of yarn have to be measured to get reliable results from errors that are relatively short as compared to a single days production.

Random Variations.

Random, of quasi-random variations cannot be predicted exactly otherwise they would not be random. However, if the population is properly sampled over a sufficient number of readings [to ensure that and appropriate number of events of included in the analysis], histograms can be constructed from which probability density curves can be inferred. Conventional statistical analysis gives a global estimate of the variance within the population tested but it is often difficult to assign the variance to all the various sources.

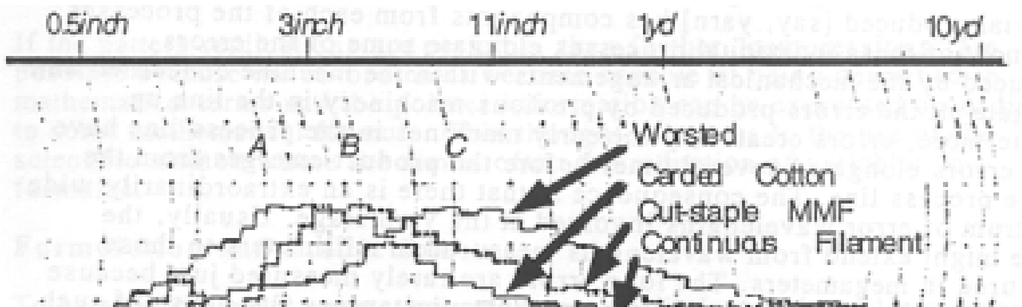

Figure 1. Spectrograms of Yarns with Different Fibers

Within limited lengths of yarn, there are quasi-random errors that display a certain pattern in a spectrogram. There is a "hill" at an error wavelength that, according to textile lore, is related to fiber length. Examples of these are given in Fig. 1, The spectrograms are displayed with amplitudes within nature of the curves.

A similar result should be obtained with a time series of yarn linear density to which an FFE has been applied. This supposes that sufficient yarn has been tested to give an acceptably accurate result. Fig.2 shows the results of such a test. The original data was generated by Aminuddin and Li and will be described at every 20th point in the time series which made the sampling interval 10mm. The resulting periodogram is not so accurate as theirs and there is a risk of some aliases distorting the result. Nevertheless, the similarity between the periodogram and the spectrograms given earlier is recognizable except for the large rise in amplitude at error wavelengths greater than 1 deter [shown in gray]. Hence it is desirable to consider what controls the accuracy in this zone.

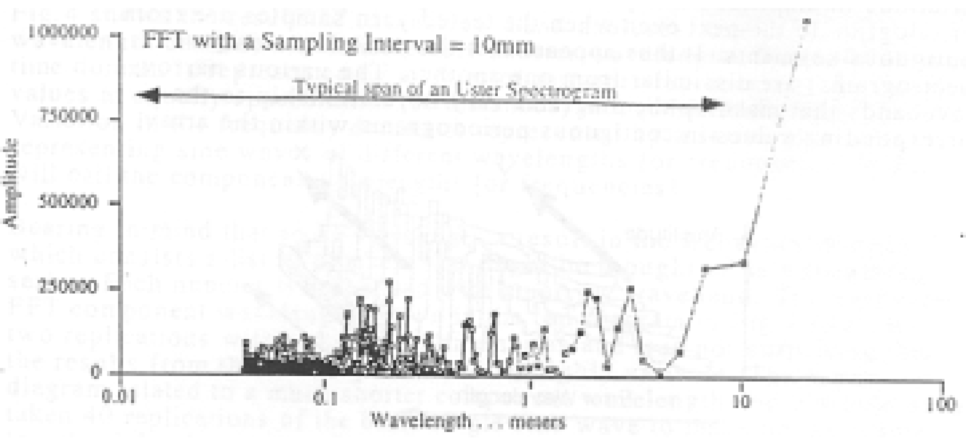

Figure 2. Periodogram Generated from an FFT of a relatively short Time Series of Linear Density Readings

It can also be mentioned in passing, that there is a risk from rare events. The appropriate data series should include a significant number of these rare events so that all important risks are included in the analysis. The rare event, such as yarn defects, end break ages, etc,., are measured in thousands of meters, sometimes hundreds of thousand of meters: consequently the amount of yarn to be tested is very large if the risks are to be assessed properly. This sort of risk is not assessed using spectrograms but by the incidence of the faults themselves. However, a comprehensive on-line system should be capable of assessing such defects.

Loss of Information in Analysis.

Because of the replication inherent in harmonic analysis, the random component is reduced and the mathematical process loses some valuable information. On the other hand, conventional statistical analysis is based on probability distributions which are built up by re-ordering the data to give statistical frequencies in the probability density curves. This destroys the information relating to the order of the data. Neither from of analysis is sufficient of its own.

Segmental FFT Analysis of Variations in Linear Density of Yarn

It has been pointed out that spectrograms of variations in yarn linear density have characteristic patterns. There is a temptation to think that one spectrogram represents the whole yarn lot but clearly this is not so. As shown in Fig.3, an array of spectrograms taken at precisely spaced internals shows not only differences in variance but also differences in variance distribution [i. e., within any given waveband]. The particular diagram shows some very large mechanical errors, which are periodic, and which replicate reasonable well from one spectrogram in the array to another. The quasi-random errors replicate less well. Indeed, when most of the periodic errors arising from mechanical imperfections are removed, the variations in amplitude in a particular waveband can vary widely from and periodogram to the next even when the tested yarn samples are from contiguous segments. It thus appears that the periodograms [or spectrograms] are dissimilar from one another. The various narrow wavebands that make up the diagrams vary in relationship to the corresponding values in contiguous periodograms within the array.

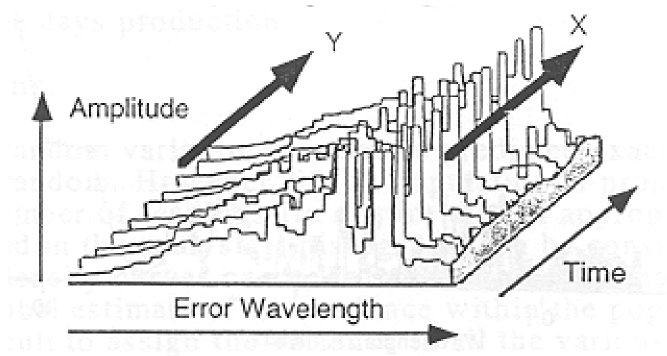

Figure 3. Mechanical Errors in Spectrograms

Consequently it is desirable to look at the variations in amplitude in the time direction such as X indicated in Fig. 3. Consider many vertical slices parallel to X and Y. If the slices contain information in each bands [or channel] used in the periodogram [or spectrogram], then information abort the variability within the bands can be gathered. If sufficient number of periodograms are used in the array, it is possible to calculate the probability distribution for each band.

Fig.4. shows the results of two narrow-band analyses, one at a long error wavelength and one at a considerably shorter wavelength. Variation in the time domain is represented by a series of numbers representing the sample values at equally spaced times [or distances] along the time series of numbers will call the component

wavelength [or frequencies].

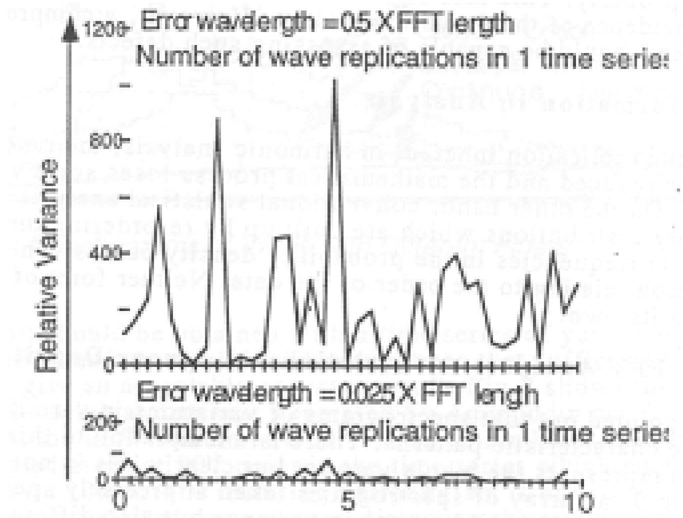

Figure 4. Variations from one Yarn Segment to another within

the same Narrow Waveband

Bearing in mind that an FFT produces a result in the frequency domain which consists a list of numbers that might be thought of as a frequency series. Each number is associated with a narrow waveband. The long-wave FFT component wavelength shown in the top diagram in Fig. 4 had only two replications within the whole time series and it is not surprising that the results from that waveband should be highly variable. The lower diagram related to a much shorter component wavelength and it would have taken 40 replications of the basic single sine wave to make up the whole length of the time series. In other words there were 40 replications and much of the error in that waveband was smoothed out by the mathematical process involved in calculating an FFT. The indicated value of amplitude was much more likely to be an accurate estimate of the enduring value in the second case than in the first.

The data in Fig. 4 had a time basis, whereas a statistician would reorder the data to give a probability distribution. If the number of occurrence [i. e., statistical frequency]within each successive waveband is plotted in the probability density curves are shown in Fig. 5, where the recorded probability is whether the amplitude falls within a specified range[i.e., in a specified bin].

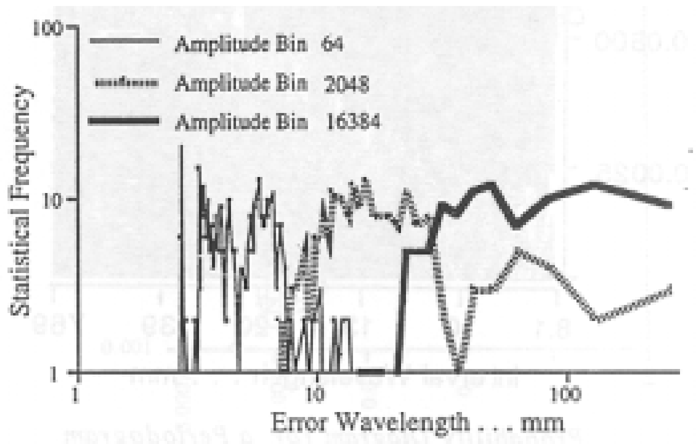

Figure 5. Probability Density Curves

The data is the same as used in Fig. 4 and was only a part of the total data available. Had the larger set of data been used, one would expect the probability distributions to become smoother. In addition, the choice of bin size affects the statistical frequencies and the degree of smoothness. An arbitrary choice was made in this respect purely for the purpose of demonstration. The value quoted for each curve in the diagram is an average for the particular bin, also the numbers should be considered as being to an arbitrary scale. The meaning of amplitude will be discussed later. Also the replication effects discussed come into play and Fig.5 contains a gray rectangle where the results shown should be regarded with some suspicion.

The probability density curve for a low level amplitude bin peaked at a relatively low error wavelength whereas that for a range of amplitude at a higher level moved to the right [i.e., to a longer wavelength]. Thus short-term errors seemed to produce small amplitudes in a manner which could be estimated with some degree of confidence. As the error wave length increased, the amplitudes also grew and were similarly predictable within a certain confidence interval. At longer wavelengths, the confidence intervals deteriorate. However, the replication errors already referred to are abated if more points are used in the time series. The gray zone moves to the right as the time series length increases. It appears as though the probability of obtaining a certain amplitude of error within a narrow waveband can be determined from a time series of sufficient length.

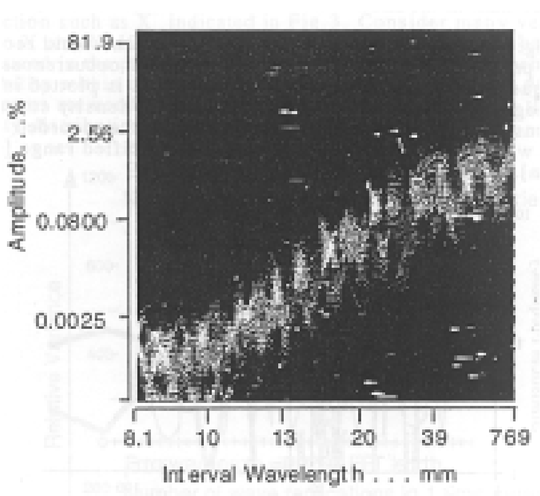

Figure 6. Predicting Reliability in Various Intervals of an FFT

Predicting Reliability in Various intervals of an FFT

Histograms were made of a large number of slices [of narrow wavebands] and probability density data were marshaled as a three-dimensional matrix. The matrix can be simplified by using greyscale shading to indicate the statistical frequency. With such a scheme the base of the three-dimensional graph has abscissae of error wavelength and ordinates of error amplitude. Areas where the statistical frequency was zero or very low appear black in the diagram and the positions of high statistical frequency is shown as light gray or even white. Thus end views of the slices are erected on abscissa of Fig. 6 to produce the most likely amplitude/ wavelength relationship for a length of yarn that is somewhat longer than the one tested. As pointed out longer interval wavelengths suffer from increasing error. As yet the calculations have not been applied to longer time series.

The data clearly shows tat lack of replications in the FFT process at long error wavelength gets causes error in the estimated amplitude. As the error wavelength gets longer in proportion to the length of the time series, does the amplitude itself. Fig. 7 indicates that the standard deviation of the amplitude is proportional to the mean value of the amplitude.

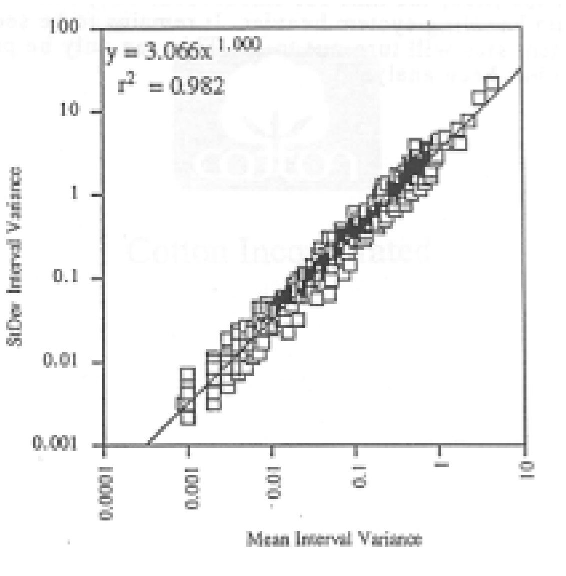

Figure 7. Variation in Amplitude of Error in an FFT Array

The goodness of the fit suggests that not only mean value of amplitude can be estimated but so can the standard deviation of it. This is more important than might be thought because the amplitude can be interpreted in terms of the variance of the yarn. Despite these prospects, it is necessary to conduct more experiments on long data runs before reliable claims can be made.

If the data had have been displayed with a frequency axis rather than wavelength [frequency α 1/wavelength], the total area under the curve would have represented variance of the yarn. The area under the curve within one wavelength channel is [frequency band-width × amplitude] and this represents the variance component within that channel. The frequency band width is calculated on the basis of {Velocity × [(1/wavelength$_1$)-(1/wavelength$_2$)]}. In other words, the amplitude referred to earlier is really a measure of the variance component in the yarn within the narrow error waveband. It is clear that the major portion of the yarn variance for each yarn lies within a limited span of error wavelength which varies from fiber to fiber. The error is not white noise because it has a component of complex order arising from changes on fiber length caused by processing.

Consequently, bearing in mind that the amplitude of a waveband is really the interval variance component of the yarn within that waveband, we can reliable relationship over the length tested.

Reconciliation between Data-gathering and Analysis

The first part of this paper discussed the difficulties of transporting and storing data and the second part discussed ways of predicting the behavior of quasi-random results. The size of the computer files needed to carry out will give a satisfactory analysis. A large number of segments gives a good estimate of the probability density distributions but if the time series segments are short, the acceptable range of error wavelengths is small. If the segments are larger, the range of acceptable error wavelengths improves but the size of the files, the time for calculation, etc., increase, making the load on the data handling system heavier. It remains to be seen what the optimum segment size will turn out to be. This can only be predicted when sufficient data has been analyzed.