7)

VARIATION REDUCTION IN A CONTINUOUS WEB PROCESS

by

William C. Pack
B.S. Civil Engineering, University of California at Berkeley, 1992

Submitted to the Sloan School of Management and the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN MANAGEMENT
and
MASTER OF SCIENCE IN CIVIL AND ENVIRONMENTAL ENGINEERING

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 1998

Signature of Author____
Sloan School of Management
Department of Civil and Environmental Engineering
May 8, 1998

Certified by____
Professor David H. Staelin, Thesis Advisor
Department of Electrical Engineering and Computer Science

Certified by____
Professor Arnold Barnett, Thesis Advisor
Sloan School of Management

Certified by____
Daniele Veneziano, Reader
Department of Civil and Environmental Engineering

Accepted by____
Larry Abeln, Associate Dean
Sloan Master's and Bachelor's Programs

Accepted by____
Joseph Sussman, Chairman
Departmental Committee on Graduate Students

VARIATION REDUCTION IN A CONTINUOUS WEB PROCESS

by

William C. Pack

Submitted to the Sloan School of Management and the
Department of Civil and Environmental Engineering on
May 8, 1998 in partial fulfillment of the requirements for
the degrees of Masters of Science in Management and
Masters of Science in Civil and Environmental Engineering

Abstract

Variation has long been an enemy to manufacturing quality and productivity. As technology increases our ability to record and process data we explore what kinds of improvements can be made with the new tools now available. The Tenneco Packaging facilities in Tomahawk, Wisconsin offer a particularly rich source of data with which to test these tools. Three good data sets have been assembled and explored as a result of this internship with varied success. One of the results of this work is the development of a design of experiments for machine speed on the mill's largest paper machine. Junehee Lee and Geoffrey Lauprete will perform further analysis on the data in conjunction with their doctoral studies.

Exploration for other potential sources of important variation was also performed including variation in first stage processing, raw material variation, and high frequency variation in final processing. Several incremental improvements have been suggested as per these studies and will hopefully lead to improved production, raw material cost savings, and reduced wear on the machinery.

Thesis Advisors: David H. Staelin, Department of Electrical Engineering and Computer Science
Arnold Barnett, Sloan School of Management

Acknowledgments

I gratefully acknowledge the support and resources made available to me through the MIT Leaders for Manufacturing (LFM) program, a partnership between MIT and U.S. Manufacturing. LFM is an inspired program that I have benefited from immensely. I am grateful to those who had the vision to bring the program together initially and those who continue to develop it today.

Along with the formal educators, I must thank my classmates for the knowledge they have shared with me for the meager price of my participation. Lifetime friendships will trace their origins to our time spent at MIT and, hopefully, formal partnerships will result from the shared experiences and dreams.

I must thank the particularly important mentors in my professional development. My parents have always been wonderfully supportive of my efforts and ambitions. Paul Curren, former boss and good friend, has been there for me at my critical inflection points. Paul always believed in me more than I did myself and was often the one responsible for that extra push that I needed to make the biggest leaps. Alexis Strauss and Christian Euvrard continue to inspire me with my memories of their exemplary leadership. And my friends who, from near and far, have continued to spice up my life and keep my mind open.

Table of Contents

Table of Figures

1. Introduction

1.1 *Objective*

The primary objective of this internship was to reduce variation in the production process at the Tenneco mill in Tomahawk Wisconsin. This was to be accomplished through the analysis of large matrices from a data rich information system and through onsite process studies. An important secondary objective was to improve our understanding of the potential role of extensive data sets in process improvements.

The Leaders for Manufacturing (LFM) Research Group number 4 (RG4) had done similar work with other data sets but hoped to facilitate the process by placing an intern at the facility. In prior work done with other LFM partners, the data was gathered by a process engineer and forwarded to MIT. This was the way the work also began with Tenneco and MIT. With an intern at the facility and Junehee Lee, a research assistant, working on the data at MIT there would be someone on site to download and preprocess the data and someone at The Institute to process it. The on-site intern would also serve as a type of translator between those trying to understand the details of the manufacturing process at MIT and those at the facility trying to better understand the analysis.

2. Problem Description

2.1 Plant History

The Tomahawk Mill has a rich history of changes in its production, process, and ownership. The original facility began making pulp in April of 1924, with production of less than 100 tons per day. Today the plant produces nearly 1500 tons of paper per day with equipment that ranges in age from brand new to seven decades old. Over that period the mill has changed ownership more than half a dozen times, three times in the last decade.

The tasks covered locally range from the harvesting of the raw materials to shipping of finished product, and everything in between. The mill generates much of its own power through dams and by burning bark, gas, coal, and tire rubber. It produces all the steam necessary for drying the paper and runs a wastewater facility that treats the total suspended solids (TSS) municipal equivalent of 25 million gallons per day (a municipal facility for a community of 200,000 people treats a comparable amount of TSS).

2.2 Background

Because of the size and complexity of their system, owners of the mill have been interested in increasing their ability to centralize the control of the various pieces of the production and find out what effects, if any, one particular area of the plant may have on another. As late as the early 80's, the mill had decentralized control with little more than scattered pneumatic controls at the machines. In 1984 they began installing a Bailey system which gave the operators more information and control and allowed engineers and others to access the data.

The Bailey system helped with local control and optimization but little had been done to tie all the parts of the process together. In 1990 they began gathering information together in a centralized database. The number of data collection points has continued to increase to the point now where they have close to 4000 data points, or tags, which are automatically downloaded into the central system. A year's worth of data is kept active and is readily accessible to anyone connected to the computer system. After the data is over a year old, it is compressed and archived.

Tenneco joined the Leaders for Manufacturing Program as a limited partner in 1996. One of Tenneco's first efforts after joining the program was to begin working with Research Group #4 on ways to use their data to understand and improve their production. The Tomahawk facility has one of the more comprehensive databases and was consequently chosen for the study.

2.3 *Strategy*

The main objective was to identify the main sources of variation in the mill's processes through multivariate analysis tools. Since the downloading, preprocessing, and translation of the data would not require all of the onsite intern's time, other process-characterization and process-improvements experiments and studies would be conducted as time allowed. The results from these other studies will also be discussed.

2.4 *Previous Work*

This thesis work at Tenneco differs from most other internship projects in that it continues work begun by LFM RG4. As such it might be interesting for readers to gain some background in this earlier work. Specifically: Mark Rawizza, *Time-Series Analysis of Multivariate Manufacturing Data Sets;* Timothy Derksen, *The Treatment of Outliers and Missing Data in Multivariate Manufacturing Data;* and Ronald Cao, *Multivariate Analysis of Manufacturing Data* all master's theses from the MIT department of Electrical Engineering and Computer Science. Also, of particular interest will be the upcoming doctoral thesis of Junehee Lee who has been my counterpart at MIT on this project. His thesis will address many specific technical details of the analysis that was done on the data sets from the Tomahawk Mill.

3. Methods of Data Analysis

3.1 Data Organization

The data was organized in a square matrix of observations and variables. For example, if we were organizing data on what conditions were ideal for growing corn, we might have 20 locations each with one acre of crop. Each location would serve as an observation. The parameters that make the sites different would be the variables. In this example, variables might be nitrogen content found naturally in the soil, rainfall, amount of sun exposure, how much the farmer talked to his crops, number of bug species found in the area, pounds of crops produced, type of combine used by each farmer, etc. We label the rows with the observations and label the columns with variables. We call the number of observations "n" and the number of variables "m". The following nxm matrix is an example of how we would organize our data.

	Nitrogen	Rainfall	Sun	Lbs. of Corn	Bugs	• • •
Topeka, KA						
Tomahawk,						
Boston, MA						
Phoenix, AZ						
•						
•						
•						

We further sub-divide the variables into process (or explanatory) variables and quality variables. Quality variables are the target variables to be effected or optimized, and the process variables are the remainder, which presumably have some impact on the quality variables. In our example we might choose the pounds of corn produced as the quality variable that we want to maximize and the rest of the variables as process variables.

In the case of manufacturing data we can create very large data sets with thousands of observations and thousands of variables. The observations are usually time stamps or snap-shots of the variables magnitudes at given intervals, where some variables are control settings. The observations or snap-shots could be taken every minute, every 15 minutes, they could be averaged values over several hours, totals taken over a given time period, or any other breakdown that might be helpful in organizing or assessing the data. The variables can also be selected by where they are in the process or taken as an entire group. The quality variable that you choose to evaluate depends on the information you are trying to distill.

3.2 *Covariance and Correlation*

Once we have chosen our quality variable and our process variables and made several observations, we want to know if there is any information to be gained about our quality variable from our process variables. For example, if we were trying to find the relationship of rainfall to pounds of crops, we would expect there to be an increase in crop yield with an increase in rainfall (to a certain point). On the other hand we would expect there to be a decrease in crop output with an increase in insect species. The degree of influence of a particular process variable on the quality variable can be measured through their covariance.

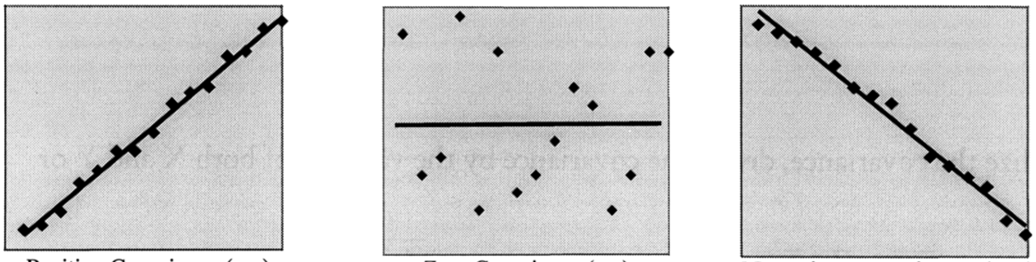
We define the covariance between two variables X and Y as the expected value of the product of each variable subtracted from its mean or:

$$Cov(X,Y) = E\left[\left(\overline{X} - X\right)*\left(\overline{Y} - Y\right)\right] \tag{3.1}$$

The covariance can also be written as the mean of the product of X and Y minus the mean of X multiplied by the mean of Y or:

$$Cov(X,Y) = \overline{XY} - \overline{X}*\overline{Y} \tag{3.2}$$

We can see from equation 3.2, when X and Y are both large, they will dominate over the product of their two means. Conversely, when X is small when Y is large or vice versa, the first term in equation 3.2 is dwarfed by the second term. If the two move independently, the sum of the various products should cancel each other out and the value should converge close to zero. This can be illustrated graphically the three simple graphs:



Positive Covariance (x,y)

Zero Covariance (x,y)

Negative covariance (x,y)

Figure 3.2.1 Covariance Graphs

The potential drawback to covariance is that it is dependent on the units of the variables. You could, for example, have a covariance of 100 between variables A and B and a covariance of 10 between A and C. Does that mean that A and B are more closely related than A and C? Not necessarily, C could be in miles and B could be in inches and if you converted C to inches it could have a covariance with A much larger than 100.

To eliminate this kind of error, the data is often normalized. Normalization of data means taking the difference between the observed value and the mean of all the observations of that variable and dividing it by its standard deviation or:

$$Z_{ij} = \frac{X_{ij} - \overline{X}_j}{\sigma_j} \tag{3.3}$$

where

$$\overline{X}_j = \frac{\sum_{i=1}^{m} X_{ij}}{m} \tag{3.4}$$

and

$$\sigma_j = \left(\frac{\sum_{i=1}^{m} \left(X_{ij} - \bar{X}_j \right)^2}{m-1} \right)^{\frac{1}{2}} \quad \text{or} \quad \sigma_j = \left(\frac{\sum_{i=1}^{m} Z_{ij}^{\,2}}{m-1} \right)^{\frac{1}{2}} \tag{3.5}$$

To normalize the covariance, divide the covariance by the variance of both X and Y or:

$$Cor(X,Y) = \rho = \frac{\overline{XY} - \bar{X} * \bar{Y}}{\sigma_x * \sigma_y} \quad \text{or} \quad Cor(Z_i, Z_j) = \rho = \frac{\overline{Z_i Z_j}}{\sigma_x * \sigma_y} \tag{3.6}$$

This gives us the correlation coefficient ρ. The correlation coefficient is unit-less, gives all data sets a mean of zero and a variance of one, and ranges in value between 1 and -1. A value of 1 means perfect positive correlation, -1 perfect negative correlation, and 0 suggests perfect independence between the two variables. Using normalized data is especially helpful with manufacturing data sets because it allows you to compare data with different units.

3.3 Regression

The simple linear regression model, which is univariate, is described by the formula

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i \tag{3.7}$$

where Xi is the ith explanatory variable, β_o is the intercept, β_1 is the slope, and ε_i is the error for the ith data point. To find the closest estimate of Y with the data provided by X we calculate the least-squares estimates. The least-squares estimate will give us the values for β_o and β_1 which, when used in the formula,

$$S(\beta_o, \beta_1) = \sum_{i=1}^{n} \left[Y_i - \left(\beta_o + \beta_1 X_i \right) \right]^2 \tag{3.8}$$

yields the smallest possible value. Setting the above equation equal to zero and differentiating gives us the following formulas for β_o and β_1:

$$\beta_o = \overline{Y} - \beta_1\overline{X} \tag{3.9}$$

$$\beta_1 = \frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2} \tag{3.10}$$

In manufacturing, as well as many other applications, we are interested in multivariate relationships (multiple variables jointly affect another variable) not just univariate relationships (one variable's effect on one other variable). The linearity of the multivariate relationship is also often measured through regression. Multiple linear regression models expand this simple model by incorporating several variables:

$Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} ... + \beta_k X_{ik} + \varepsilon_i$

Where *i* represents the number of variables used to model Y.

As would be expected, when trying to characterize a quality variable with numerous process variables, some process variables will have a larger influence on the chosen quality variable than others. In fact, many of the variables may have very little influence at all and can be ignored. One of the major challenges in trying to model with a large number of process variables is trimming down the data to those that matter most. The process of taking a large data set, calculating the coefficient values and then eliminating the insignificant factors is commonly called backward elimination.

Another challenge in finding a set of clear explanatory (or process) variables is weeding out those variables that give the same or similar information. This phenomenon is known as multicollinearity and exists when there is linear or near linear dependence between one or more explanatory variables. If we consider the previous example about the production of corn, we could expect that the two variables "rainfall" and "exposure to sunlight" could have an inverse relationship since the sun would seldom be shining when it was raining. It wouldn't be raining every time the sun was not shining, but we could expect there to be some dependence, as the one condition would exist almost exclusively without the other condition.

The reason we are concerned about multicollinearity is because the values for β_1 and β_2 will be effected if collinearity exists between x_{i1} and x_{i2}. Since β_1 and β_2 are estimates, their values are expressed by a distribution of probable values. The main concern as coefficients is whether or not the value could actually be zero, which means they would have no effect on the value of the quality variable. Two different tests are typically performed to check for this possibility, the t-statistic and the p-value. The t-statistic is roughly the multiple of standard deviations that the value zero is away from the mean value of βi in the distribution and the p-value is the probability that the coefficient's distribution "crosses" zero. So, it is important to know if two variables are correlated as their interdependence could effect their estimated importance.

To measure multicollinearity of two variables we return to the correlation coefficient. Ideally, the coefficients will have no correlation (a value near zero) and we can use them with confidence. When the correlation between two variables is zero, we call them orthogonal.

3.4 *Principal Component Analysis*

Data reduction and the identification of multicollinearity (to an extent) can be eloquently addressed with the use of one relatively simple tool, Principal Component Analysis. I will refer to Principal Component Analysis, as KLT after its other name, Karhunen-Loéve Transformation because the acronym PCA in Tenneco circles stands for Paper Company of America, the previous owners of the mill.

KLT can be best understood geometrically. Imagine a three dimensional data set with all the data points spread out randomly inside a space shaped like a flat football as in the figure on the following page.

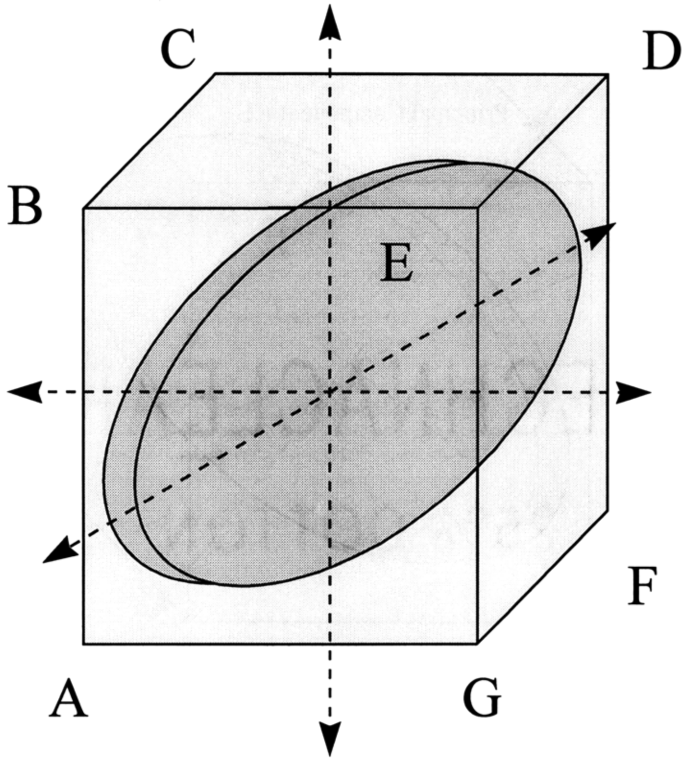


Figure 3.4.1 3-D Data Graph (normal axis)

One can see by the shape of the data set that the greatest variation exists in the direction from corner A to corner D, the next largest variation exists in the "width" of the data or roughly from corners E to G, and the least source of variation is found across corners C to E. KLT seeks out the largest source of variation using the least sum of squares like our regression discussion above only it calculates it in a dimensional space equal to the number of variables (3 in our example but usually much larger). This new vector, the largest source of variation, is known as the first principal component. KLT then seeks out the next principal component, or next largest source of variation with the stipulation that it must be orthogonal to the first principal component. The third principal component must be perpendicular to the first two, and so on until you have as many principal components as you have variables. The transformed axis might look like the figure on the following page.

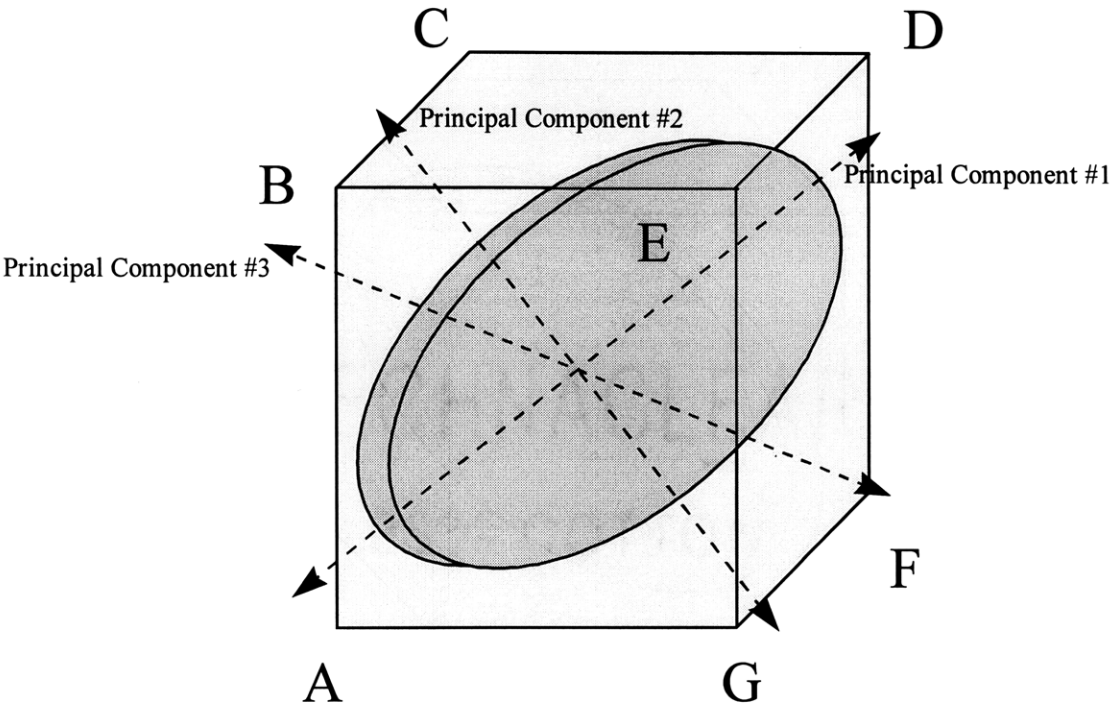


Figure 3.4.2 3-D Data Graph (adjusted axis)

The three dimensional example above may seem trivial but when you are working with a matrix with hundreds or thousands of variable you begin to understand how powerful a tool KLT can be.

Once we have calculated the principal components we have to translate them into information about the individual process variables. Principal component analysis gives us information in the form:

$$\begin{aligned} v_1 &= a_1x_1 + b_1x_2 + c_1x_3 \cdots \\ v_2 &= a_2x_1 + b_2x_2 + c_2x_3 \cdots \\ v_3 &= a_3x_1 + b_3x_2 + c_3x_3 \cdots \\ &\vdots \end{aligned}$$

Where v_1, v_2 and v_3 represent principal components; *a*, *b* and *c* represent coefficients; and x_1, x_2 and x_3 represent process variables. The eigenvalues of the individual principal components listed above give us the relative importance of the individual principal components vs. the other principal components in the group. The groups of coefficients (*a*, *b* and *c)* define the various principal components or eigenvectors. If the principal components clearly define subsets of all parameters, these groups are more likely to represent real process variations, which have the potential to be controlled.

What we need to do now is translate this information so that we know the weight or importance of the individual process variables decoupled from the principal components. This is done by choosing a quality variable (which is removed from the matrix before principal component analysis) and regressing it against all other variables. The resultant ranking shows us the importance of each individual process variable.

Unfortunately, this can introduce some multicollinearity back into the relationship between variables. To address the multicollinearity one needs to take the now short list of important variables and eliminate the obvious "clones" and test the suspicious ones by calculating the cross product or correlation coefficient.

3.5 Process Studies

One of the difficulties in using collected data to model or predict performance is limited information. If the amount of fertilizer is the most important variable in corn the quantities used are unknown, you are completely unable to take advantage of the predictive power of the variable. For this reason it is critical to not only take the data that is available and process it but also to analyze the system to make sure the most important information is getting into the system or is somehow recognized as a contributing source of variation.

It is also essential to understand how the quality variables are calculated. The weight of corn produced is a fairly straightforward metric but many times the quality of a good is measured by a rather complicated test. These tests vary in repeatability, how well they actually describe the attribute you are trying to measure, cost, and complexity. For these reasons, considerable time was spent on understanding how the Tomahawk system worked and how the quality variables were calculated. The particular methods used to study the processes will be discussed with the description of each project.

4. Process Description

Before going into the details of how quality parameters were chosen and how specific experiments were run, it would be helpful to describe how the process works at this mill and how all the pieces fit together.

4.1 Raw Materials

The main raw material inputs to the system are wood fiber, steam, chemicals, electrical energy and lots of water. The wood fiber arrives in various forms including chips, round wood, and recycled materials. The wood chips come from neighboring lumber mills. The chips come in fairly homogeneous loads of the various species found in the area (aspen, oak, and other mixed hardwoods). The chips are dumped out of the semi trucks they are delivered in and then stored in large piles by end loaders.

The round wood, or logs, arrive by train and by truck and are stored in large piles according to the species and time of delivery. The round wood is supplied by the company's internal lumber supplier and is also purchased from outside suppliers. Because of the methods used to process the lumber at the mill, all of the fiber that is processed through the entire system is from deciduous trees, no coniferous wood is used.

There are two sources of recycled material that is used as fiber sources at the mill. First is the "broke stock" this is material that has been trimmed out of the process at some point but is still a perfectly good fiber source. The other type of recycled material is Double Lined Craft board or "DLK" (also called "waste"). DLK basically the scrap from box plants or other sources of clean cardboard. DLK is important, as it is often price competitive with other sources of fiber and it provides a source of longer fibers, which can be particularly important with certain grades of paper.

4.2 Energy/Steam

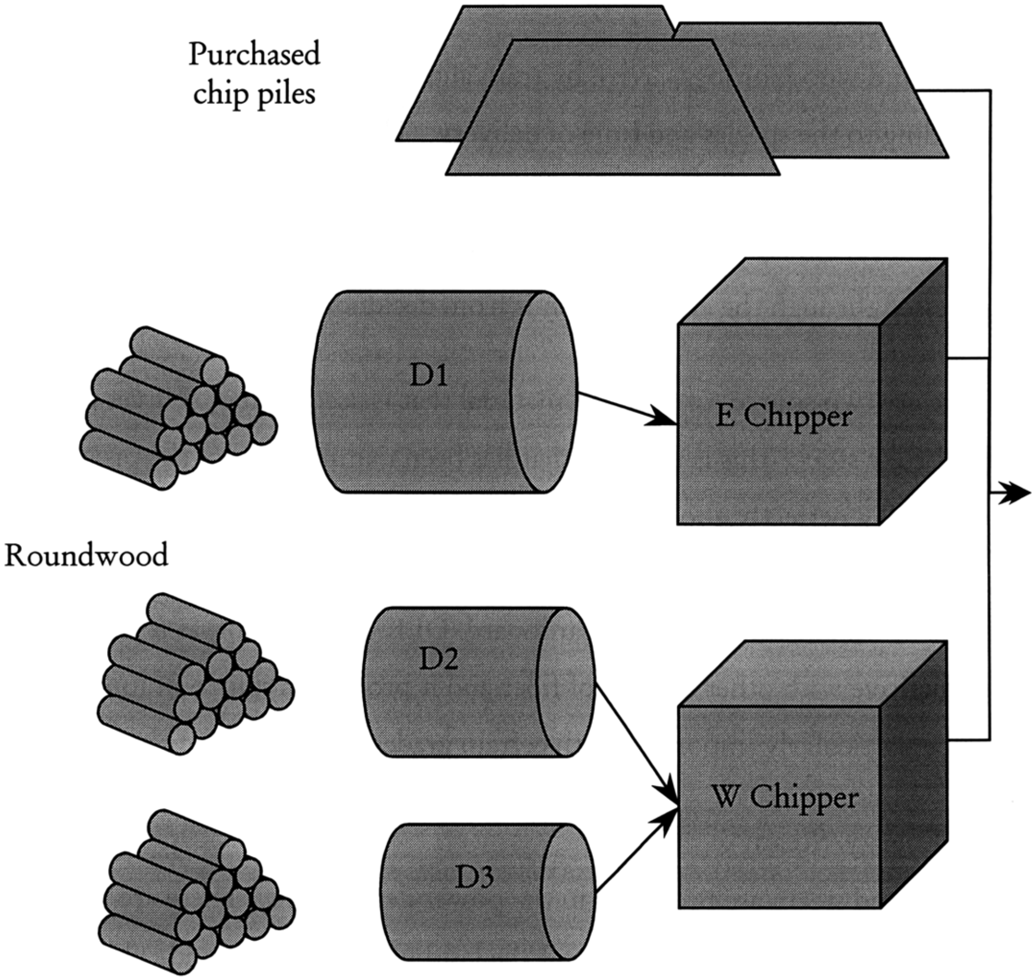
Energy and steam are produced in the power department. There are four boilers, two run on coal, one on natural gas, and the last burns both bark and rubber from tires.

Energy is supplemented by a hydroelectric source and by electricity purchased from the local Power Company.

4.3 The Woodroom

There are three wood lines in the woodroom to process the roundwood. Each line consists of a log deck and a debarker. The logs are placed on the log deck where they are incrementally fed into a debarker. The debarker is like a large cloths dryer which slowly tumbles the logs causing them to scrape against each other and peal off the bark. The logs work their way through the debarker until they fall out the other end on to a conveyor that carries them to a wood chipper. The first line feeds into the east chipper and the other two lines feed into the west chipper.

Figure 4.3.1 Woodyard/Woodroom

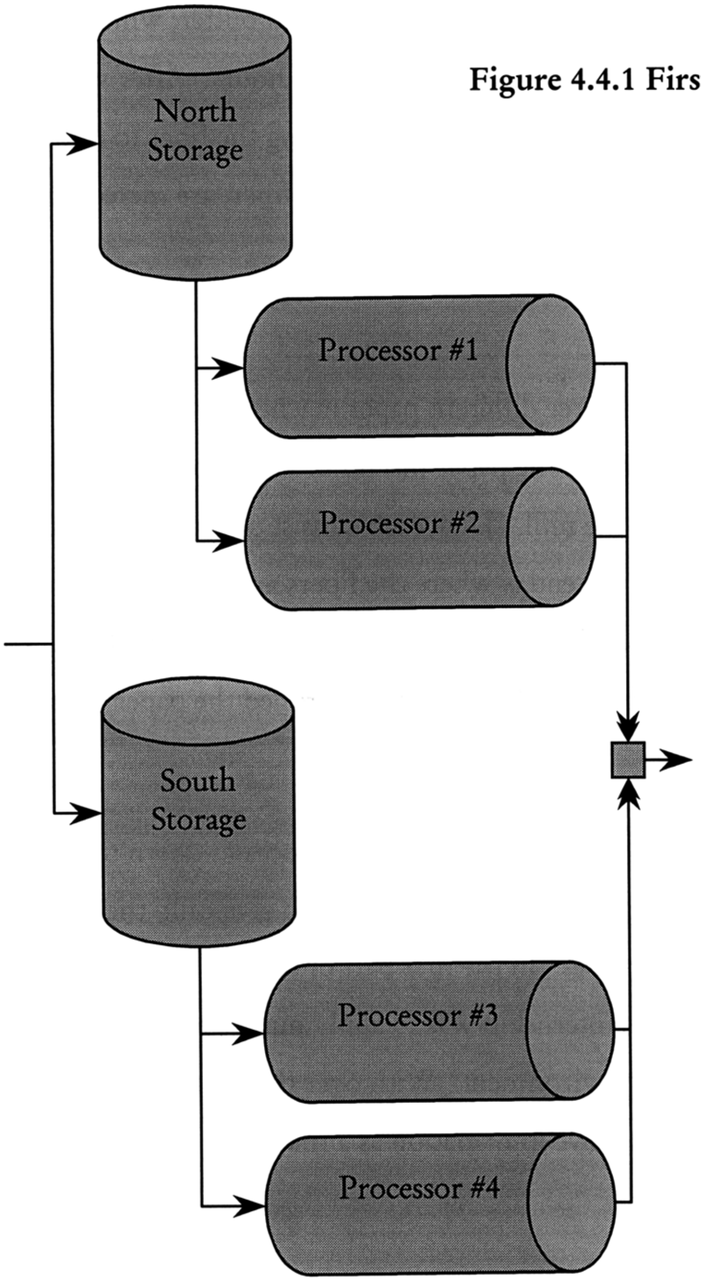


The purchased chips are fed in and mixed with the roundwood chips after the chippers. From the mixing point the chips are stored in one of two silos. The north and south chip silos are the same size and can hold several hours of chip inventory. The two silos are fed by the same system so only one silo can be filled at a time.

4.4 First Stage Processing

From the silos, the chips are fed into first stage processors where they are combined with steam and chemicals. The north silo feeds processors #1 and #2 while the south silo feeds processors #3 and #4.

Figure 4.4.1 First Stage Processing



This is the only time that the fibers will be in an environment that is super heated and so rich with chemicals. This is where the majority of the breakdown of the large fiber groups occurs.

The fiber is then expanded in a vacuum environment and then rinsed to remove the chemicals that were used to break the fibers down. This is intended to stop the chemical treatment but the physical manipulation of the fiber continues until final processing.

4.5 Buffers/Refining

After washing, the pulp is stored in the major system buffer, which has sufficient capacity to supply all the paper machines for many hours. After storage, the fiber is refined in two steps. This refining consists of cutting the fiber to optimal length. At this point the two other fiber sources, broke and waste fiber, are mixed in with the virgin fiber and readied for final processing.

4.6 Final Processing

The pulp is sent to one of three different paper machines where it receives final touch up and then is made into paper. The paper machines are major investments and are the production constraint of the mill. They each consist of two major sections, a wet end and a set of dryers. The wet end is where the fibers are oriented and the majority of the water removal takes place. The dryers are the large, steam filled drums where last of the unwanted moisture is removed. Once dried and rolled the paper is trimmed and prepared for shipping.

We have just reduced one of the great industrial processes down to a few pages, to a few simple steps. The description is obviously meant to give only the simplest outline to the process for reference purposes. In the first part of this paper it was made clear that one of the main purposes of this internship was to apply numerous variation reduction techniques in a data rich environment. With over three thousand variables recorded at the mill, we will use the above information as a road map of what is taking place.

5. Data Set #1

As a first cut at understanding the system and the data, our objective with the first data set was to quickly gather an all-inclusive matrix of data so we could see what we were up against. After discussions with RG4 and plant personnel it was decided to gather a two-week data set with data collected at 6-minute intervals.

5.1 Data Gathering

In an effort to minimize the time needed to download the data and clean it, the first step was to take the existing list of available data tags and find which of them were repetitive or no longer in use. We started with over 4000 variables, we were able to reduce the list by 640 tags that were no longer relevant (final count 3636 tags). Since the data system at the mill downloaded into MS Excel, which has a limited matrix size, the larger matrix had to be broken down into 37 (634 x 100) matrices.

With the matrix size and content established, it was just a matter of pulling the information from the central server into the individual files and compiling them for shipment back to MIT.

5.2 Data Cleaning

Once the data had been successfully transferred, it had to be translated from the Excel format to a format that could be understood by MATLAB (MATLAB is a particularly powerful software/language that was used with this project).

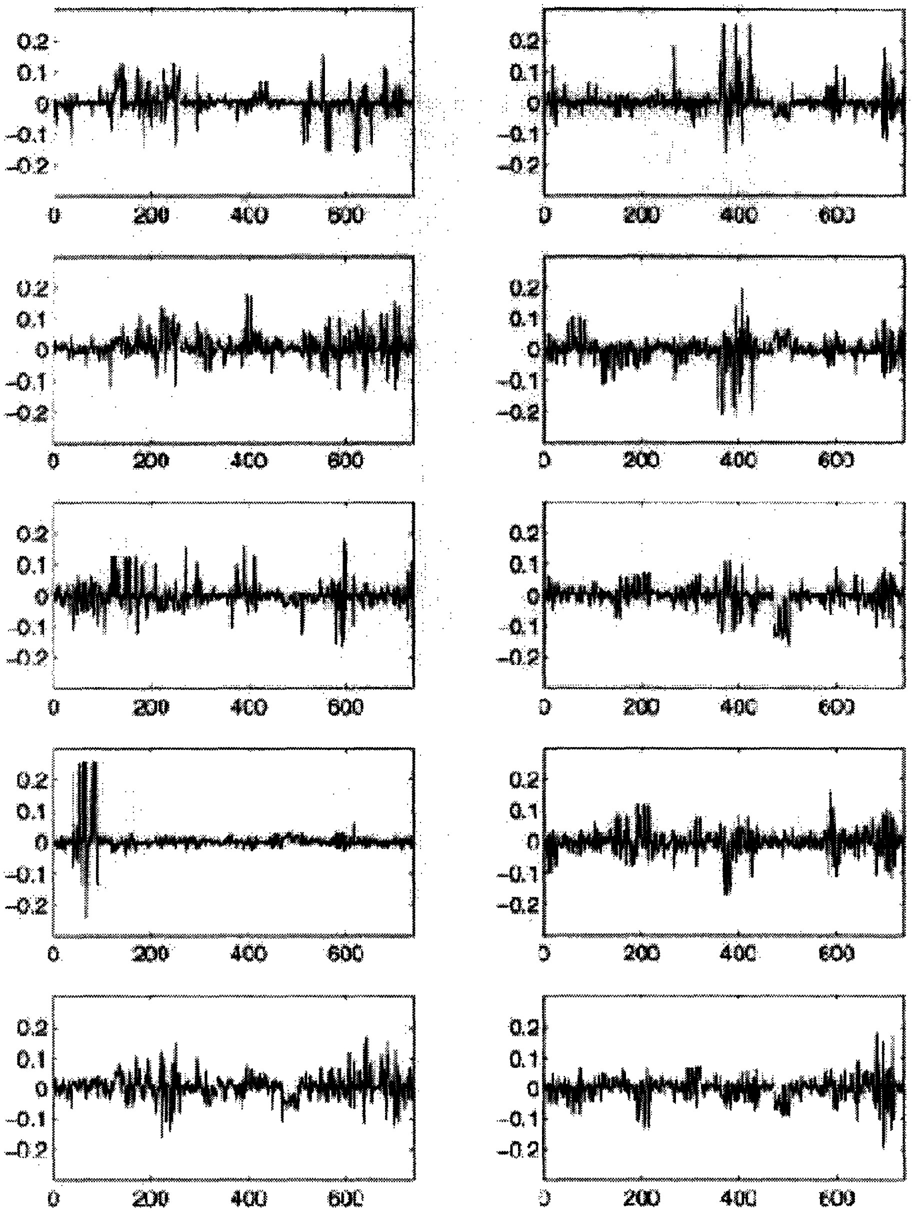
One of the early hurdles in processing a data set this large is finding all the outlying and missing data and deciding what should be done with the missing data. The two main types of bad data we were trying to eliminate were data that was inappropriately extreme and output that had returned non numeric data. Considering the size of the data set, it was decided to simply identify and remove those columns of data which contained this "bad" data and work with the remaining data. Of the original data, 2337 variables "survived" or contained enough information to be useful using our chosen cleaning method.

5.3 Data processing

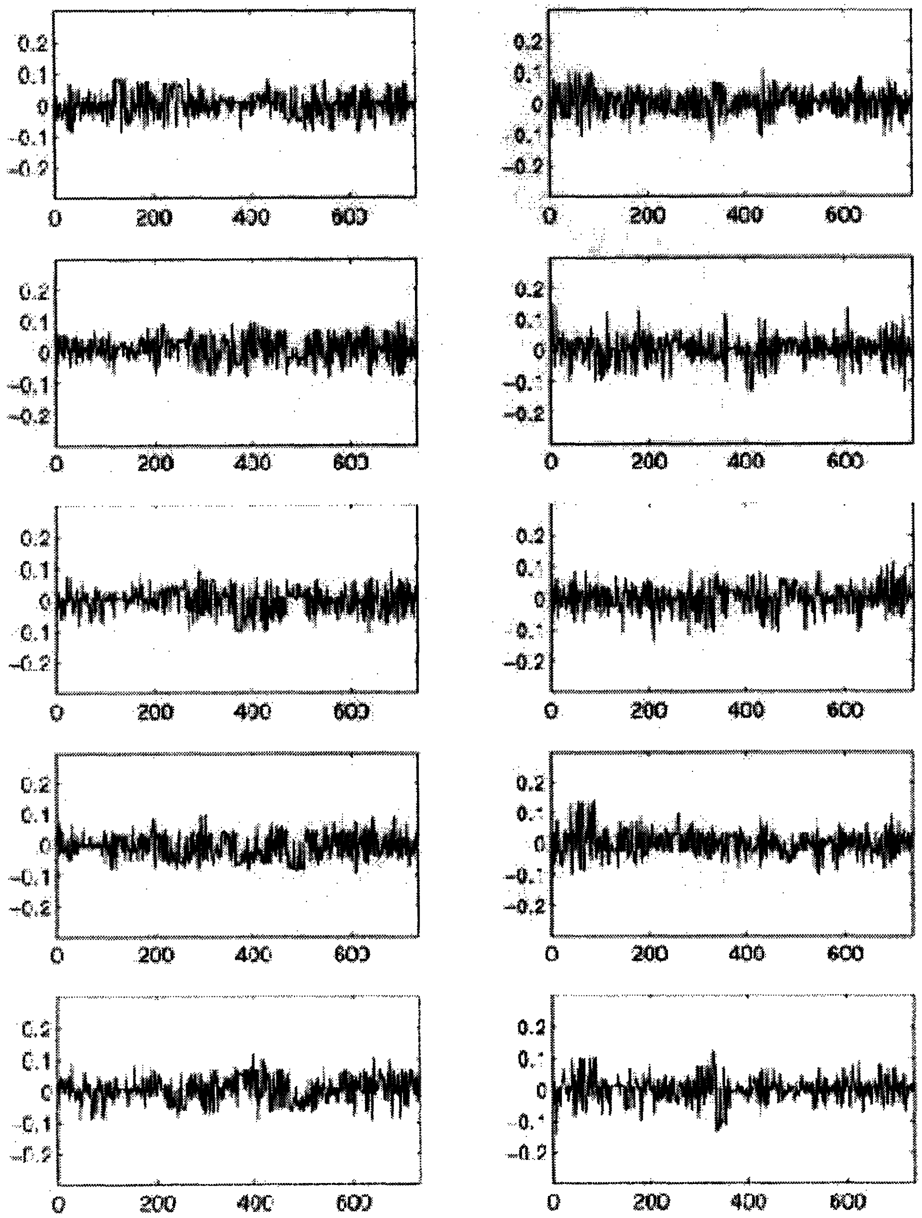
The first step in processing the data is to decide what is the variable or variables for which you want information. In a manufacturing process the obvious place to look is at the specification metrics for the product. For the mill the most important metrics are the tests they perform which give them information on the strength of the paper. These tests include CMT and the Ring Crush tests, which are run on the rolls of paper and have to meet minimum thresholds to be acceptable for shipment. Other interesting quality variables are those that reflect how well the machines are running. These would include machine speed and the number of breaks that occurred during runs (a break is when a tear occurs across the entire sheet. Breaks result in halted production and require the re-threading of the machine and hence cause considerable interruptions to production.).

5.4 Data Results

The following are graphs of the more important eigenvectors from the first 10 principal components of the first data set (they are ordered top to bottom, left to right so plot #1 is top left and #10 is bottom right). The x-axis represents the variable location across the data set and the y-axis represents the normalized value of that particular variable.

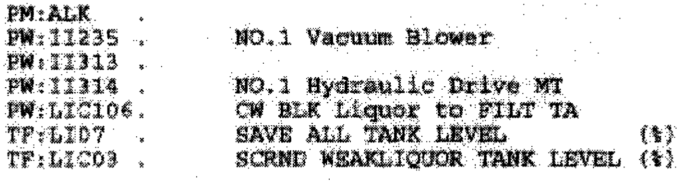

0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600
0.2
0.1
0
−0.1
−0.2
0
200
400
600

There is little obvious information to be gathered from the plots in their current state. They look fairly random in their distribution and tell us little about individual, or groups, of variables. When these same eigenvectors are filtered with a high pass filter, we can see there is more information. In particular, we see that eigenvectors 1,2,3,4,7 and 9 show groupings of variables that differ significantly (two octave) from the rest of the set.



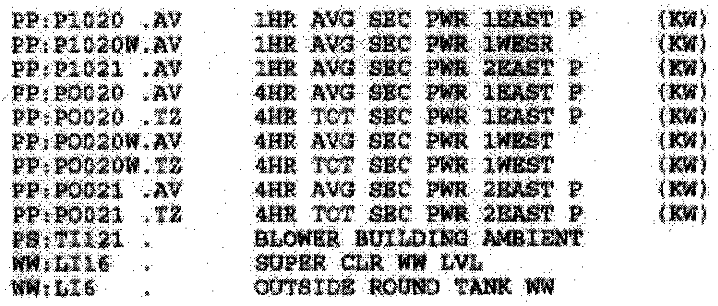
These groups of principal components can be divided up into the following important groups of variables.

From EV #1,



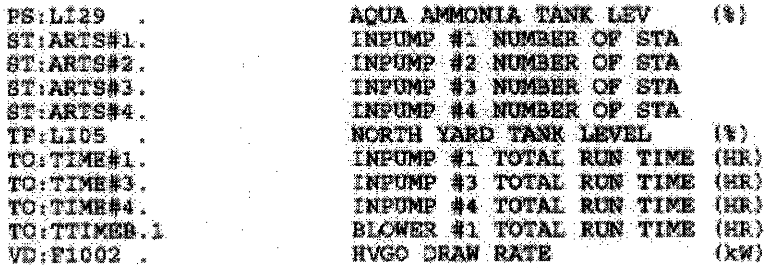
```
PM:ALK    .
PW:II235  .      NO.1 Vacuum Blower
PW:II313  .
PW:II314  .      NO.1 Hydraulic Drive MT
PW:LIC106.       CW BLK Liquor to FILT TA
TF:LI07   .      SAVE ALL TANK LEVEL         (%)
TF:LIC03  .      SCRND WEAKLIQUOR TANK LEVEL (%)
```

From EV #2,



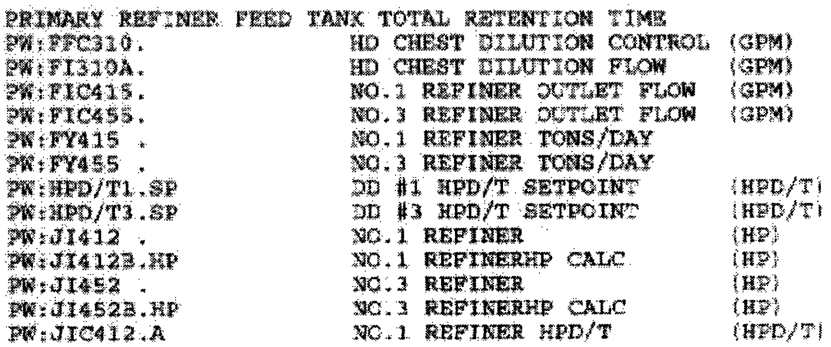
```
PP:P1020 .AV      1HR AVG SEC PWR 1EAST P      (KW)
PP:P1020W.AV      1HR AVG SEC PWR 1WESR        (KW)
PP:P1021 .AV      1HR AVG SEC PWR 2EAST P      (KW)
PP:PO020 .AV      4HR AVG SEC PWR 1EAST P      (KW)
PP:PO020 .TZ      4HR TCT SEC PWR 1EAST P      (KW)
PP:PO020W.AV      4HR AVG SEC PWR 1WEST        (KW)
PP:PO020W.TZ      4HR TCT SEC PWR 1WEST        (KW)
PP:PO021 .AV      4HR AVG SEC PWR 2EAST P      (KW)
PP:PO021 .TZ      4HR TCT SEC PWR 2EAST P      (KW)
PS:TI121 .        BLOWER BUILDING AMBIENT
WW:LI16  .        SUPER CLR WW LVL
WW:LI6   .        OUTSIDE ROUND TANK WW
```

From EV #3,



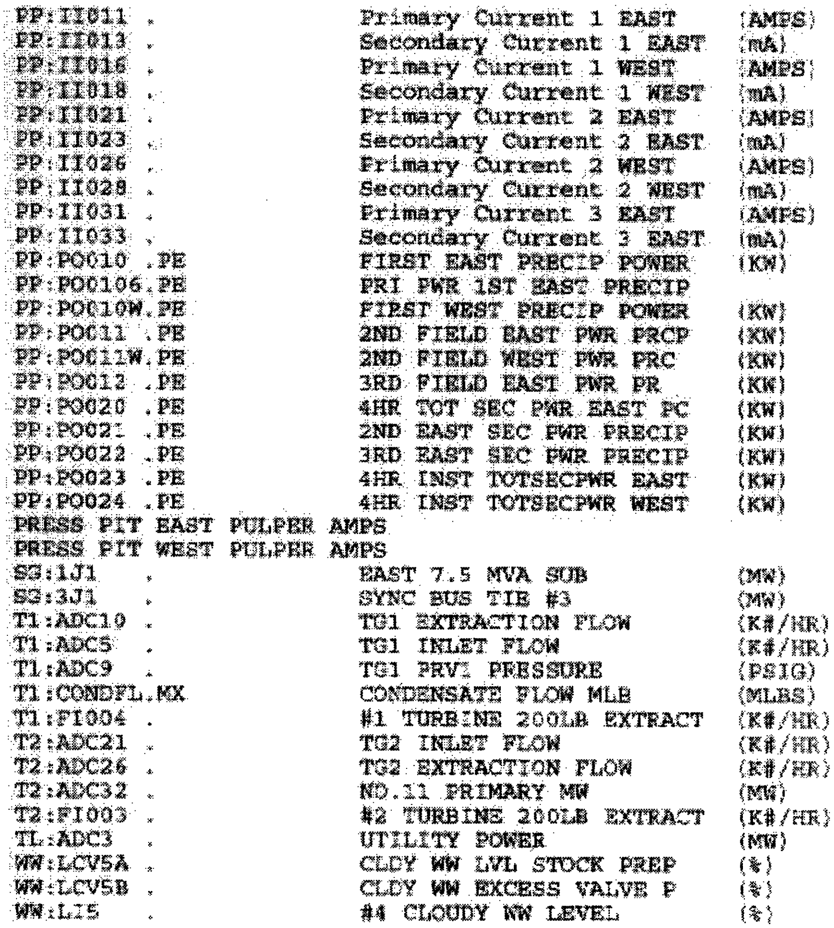
```
PS:LI29   .               AQUA AMMONIA TANK LEV    (%)
ST:ARTS#1.                INPUMP #1 NUMBER OF STA
ST:ARTS#2.                INPUMP #2 NUMBER OF STA
ST:ARTS#3.                INPUMP #3 NUMBER OF STA
ST:ARTS#4.                INPUMP #4 NUMBER OF STA
TF:LI05   .               NORTH YARD TANK LEVEL    (%)
TO:TIME#1.                INPUMP #1 TOTAL RUN TIME (HR)
TO:TIME#3.                INPUMP #3 TOTAL RUN TIME (HR)
TO:TIME#4.                INPUMP #4 TOTAL RUN TIME (HR)
TO:TTIMEB.1               BLOWER #1 TOTAL RUN TIME (HR)
VD:F1002  .               HVGO DRAW RATE           (kW)
```

From EV #4,



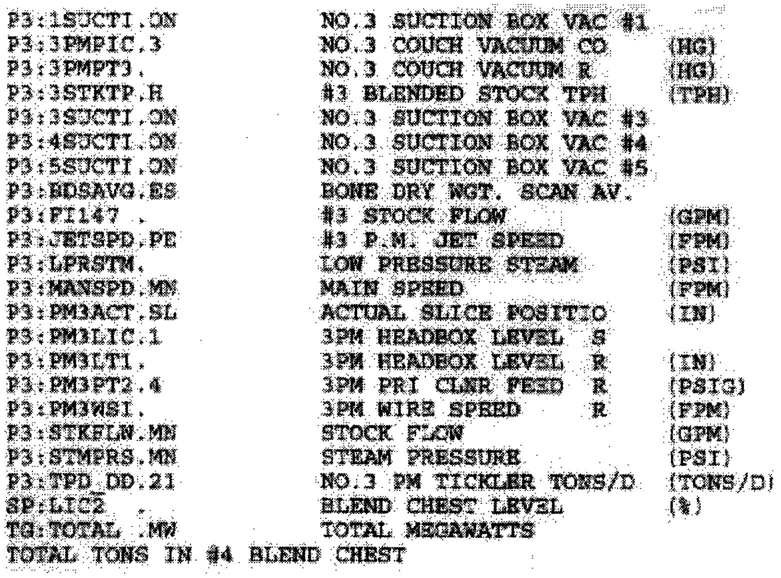
```
PRIMARY REFINER FEED TANK TOTAL RETENTION TIME
PW:FFC310.              HD CHEST DILUTION CONTROL (GPM)
PW:FI310A.              HD CHEST DILUTION FLOW    (GPM)
PW:FIC415.              NO.1 REFINER OUTLET FLOW  (GPM)
PW:FIC455.              NO.3 REFINER OUTLET FLOW  (GPM)
PW:FY415 .              NO.1 REFINER TONS/DAY
PW:FY455 .              NO.3 REFINER TONS/DAY
PW:HPD/T1.SP            DD #1 HPD/T SETPOINT      (HPD/T)
PW:HPD/T3.SP            DD #3 HPD/T SETPOINT      (HPD/T)
PW:JI412 .              NO.1 REFINER              (HP)
PW:JI412B.HP            NO.1 REFINERHP CALC       (HP)
PW:JI452 .              NO.3 REFINER              (HP)
PW:JI452B.HP            NO.3 REFINERHP CALC       (HP)
PW:JIC412.A             NO.1 REFINER HPD/T        (HPD/T)
```

From EV #7,



```
PP:II011 .              Primary Current 1 EAST    (AMPS)
PP:II013 .              Secondary Current 1 EAST  (mA)
PP:II016 .              Primary Current 1 WEST    (AMPS)
PP:II018 .              Secondary Current 1 WEST  (mA)
PP:II021 .              Primary Current 2 EAST    (AMPS)
PP:II023 .              Secondary Current 2 EAST  (mA)
PP:II026 .              Primary Current 2 WEST    (AMPS)
PP:II028 .              Secondary Current 2 WEST  (mA)
PP:II031 .              Primary Current 3 EAST    (AMPS)
PP:II033 .              Secondary Current 3 EAST  (mA)
PP:PO010 .PE            FIRST EAST PRECIP POWER   (KW)
PP:PO0106.PE            PRI PWR 1ST EAST PRECIP
PP:PO010W.PE            FIRST WEST PRECIP POWER   (KW)
PP:PO011 .PE            2ND FIELD EAST PWR PRCP   (KW)
PP:PO011W.PE            2ND FIELD WEST PWR PRC    (KW)
PP:PO012 .PE            3RD FIELD EAST PWR PR     (KW)
PP:PO020 .PE            4HR TOT SEC PWR EAST PC   (KW)
PP:PO021 .PE            2ND EAST SEC PWR PRECIP   (KW)
PP:PO022 .PE            3RD EAST SEC PWR PRECIP   (KW)
PP:PO023 .PE            4HR INST TOTSECPWR EAST   (KW)
PP:PO024 .PE            4HR INST TOTSECPWR WEST   (KW)
PRESS PIT EAST PULPER AMPS
PRESS PIT WEST PULPER AMPS
S3:1J1   .              EAST 7.5 MVA SUB          (MW)
S3:3J1   .              SYNC BUS TIE #3           (MW)
T1:ADC10 .              TG1 EXTRACTION FLOW       (K#/HR)
T1:ADC5  .              TG1 INLET FLOW            (K#/HR)
T1:ADC9  .              TG1 PRV1 PRESSURE         (PSIG)
T1:CONDFL.MX            CONDENSATE FLOW MLB       (MLBS)
T1:FI004 .              #1 TURBINE 200LB EXTRACT  (K#/HR)
T2:ADC21 .              TG2 INLET FLOW            (K#/HR)
T2:ADC26 .              TG2 EXTRACTION FLOW       (K#/HR)
T2:ADC32 .              NO.11 PRIMARY MW          (MW)
T2:FI003 .              #2 TURBINE 200LB EXTRACT  (K#/HR)
TL:ADC3  .              UTILITY POWER             (MW)
WW:LCV5A .              CLDY WW LVL STOCK PREP    (%)
WW:LCV5B .              CLDY WW EXCESS VALVE P    (%)
WW:LI5   .              #4 CLOUDY WW LEVEL        (%)
```

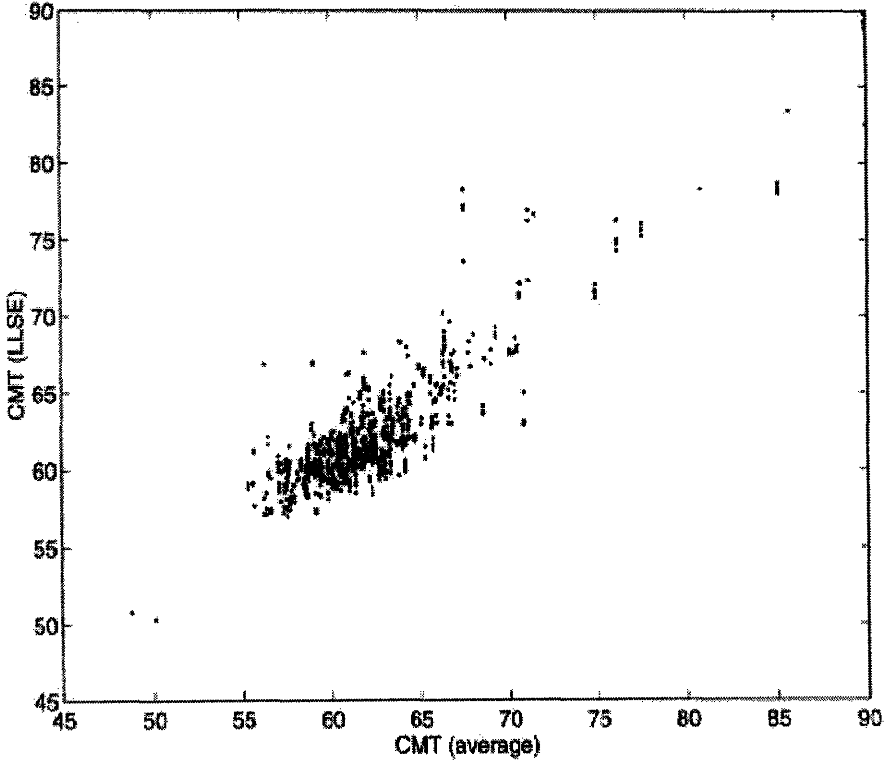
From EV #9,



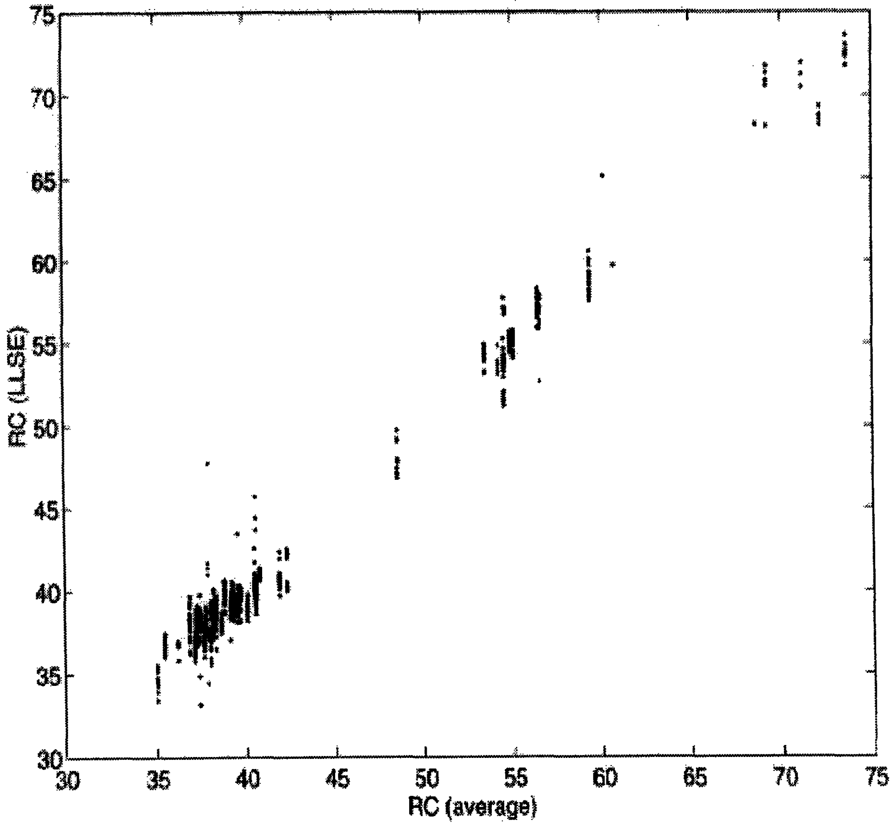
```
P3:1SUCTI.ON          NO.3 SUCTION BOX VAC #1
P3:3PMPIC.3           NO.3 COUCH VACUUM CO      (HG)
P3:3PMPT3.            NO.3 COUCH VACUUM R       (HG)
P3:3STKTP.H           #3 BLENDED STOCK TPH      (TPH)
P3:3SUCTI.ON          NO.3 SUCTION BOX VAC #3
P3:4SUCTI.ON          NO.3 SUCTION BOX VAC #4
P3:5SUCTI.ON          NO.3 SUCTION BOX VAC #5
P3:BDSAVG.ES          BONE DRY WGT. SCAN AV.
P3:FI147 .            #3 STOCK FLOW             (GPM)
P3:JETSPD.PE          #3 P.M. JET SPEED         (FPM)
P3:LPRSTM.            LOW PRESSURE STEAM        (PSI)
P3:MANSPD.MN          MAIN SPEED                (FPM)
P3:PM3ACT.SL          ACTUAL SLICE POSITIO      (IN)
P3:PM3LIC.1           3PM HEADBOX LEVEL  S
P3:PM3LT1.            3PM HEADBOX LEVEL  R      (IN)
P3:PM3PT2.4           3PM PRI CLNR FEED  R      (PSIG)
P3:PM3WSI.            3PM WIRE SPEED     R      (FPM)
P3:STKFLW.MN          STOCK FLOW                (GPM)
P3:STMPRS.MN          STEAM PRESSURE            (PSI)
P3:TPD_DD.21          NO.3 PM TICKLER TONS/D    (TONS/D)
SP:LIC2  .            BLEND CHEST LEVEL         (%)
TG:TOTAL .MW          TOTAL MEGAWATTS
TOTAL TONS IN #4 BLEND CHEST
```

What these graphs tell us is that these groups of variables had an above average impact on the first ten principal components of the data set. These first ten principal components represent the largest sources of variation in the process so we should look to these lists of variables to see what effect they are having on production as a whole.

As a separate exercise, we correlate the data set against two quality parameters, the CMT test and the Ring crush test. The graphical results are shown on the following page. The x-axis represents the true value and the y-axis represents the predicted value.



Validating Set of CMT test (with regular and strong paper)



Validating Set of RC test (with regular and strong paper)

While at first blush these graphs seem to show promise, there was so much noise from the data from different machines that the information was not as useful as hoped. This is because, as will be seen later with the second data set, when we further process the data by removing all of the variables that are listed as important and obviously correlated with quality variable, our predictive powers are reduced.

6. Data Set #2

Because of the difficulty we had with noise in the first data set, we decided to be more focused in the second data set. Instead of taking data from all available tags, we went through the list of available data and chose tags from one machine that were deemed most important using engineering judgement.

6.1 Data Gathering

We had a target of around 600 variables and kept the set to 597 variables. Because the newer paper machine accounts for almost two thirds of the production of the mill, we focused on the variables that we thought had a direct effect on that machine. Because we were once again focusing on the CMT test as one of the quality variables, we decided to download the data at the same time that each paper reel finished. In other words, each paper roll finishes at a certain time and that time is recorded as the time that the CMT test is sampled. So, we found the time that each sample was taken and then found out what was the value of the other variables we were interested in that time. Since the CMT tests are taken at roughly an hourly time frame it looks like that was the method used to sample the data. However, the data is structured according to the reel sampling.

Another adjustment we made with the second data set was to take data only when the paper machine of interest was making one particular grade of paper. At the mill, one grade of paper was routinely run during the week and other grades (or thicknesses) of paper were run on the weekends. As one can imagine, when the paper is thicker many of the properties of the paper making system will change. With our pre filtering, we hoped to eliminate this noise.

6.2 Data Cleaning

Because so much data was lost in the cleaning of the first data set and because the second data set was much smaller, it was decided to clean the data set manually. This meant going through the entire matrix and finding all the bad data. Once found, the data would have to be replaced by data that was an interpolation of the appropriate value.

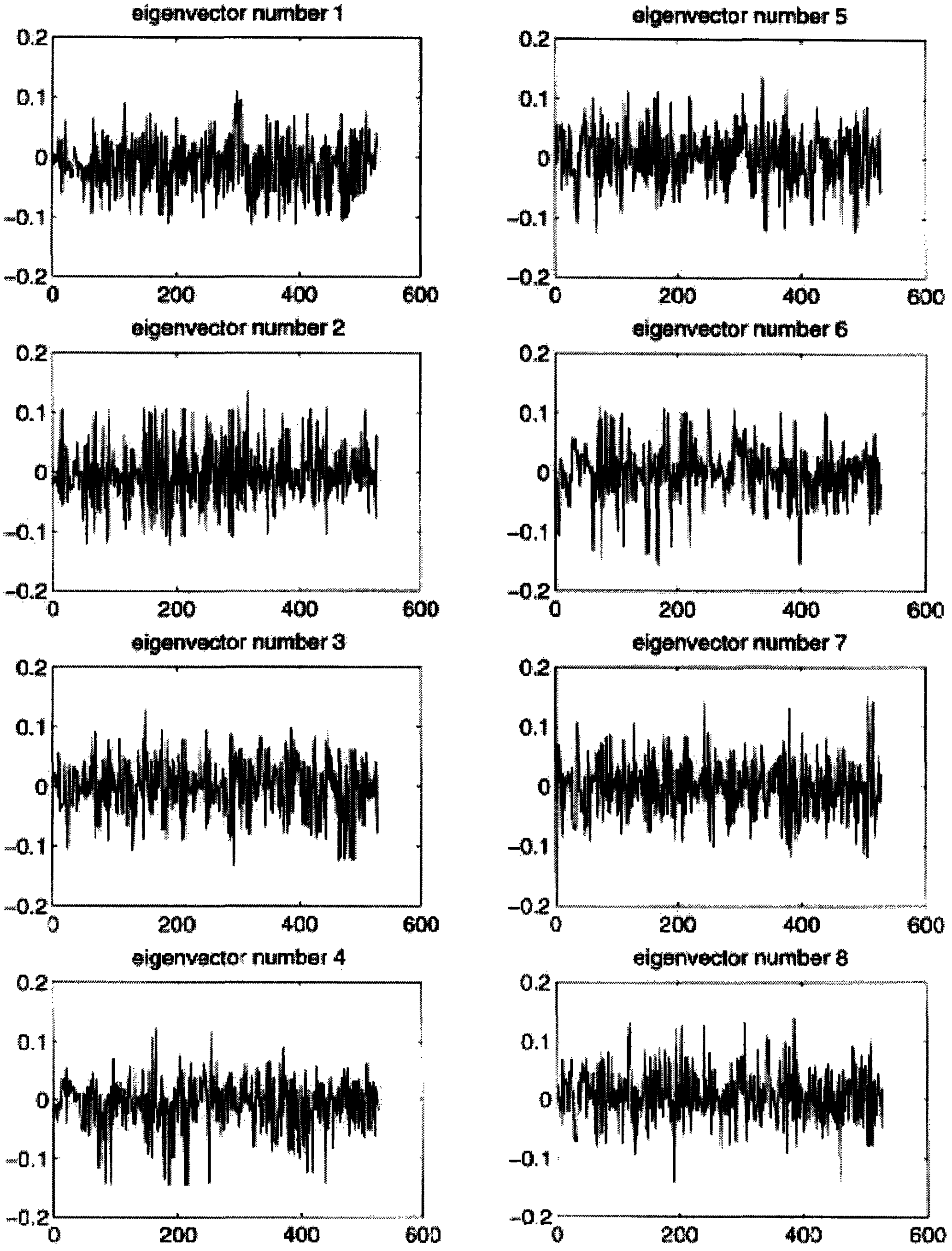
This was a long and painstaking process but allowed us to keep all the information with maximum accuracy.

6.3 Data processing

As with the first data set, we used quality metrics like CMT and machine speed to analyze the data

6.4 Data results

The following are graphs of the eigenvectors for the second data set. As with the first data set, the x-axis represents the variable location across the data set and the y-axis represents the normalized value of that particular variable. It is obvious from the plots that there is little information to be gathered from the eigenvectors in their current state. They look almost perfectly random in their distribution and, as before, tell us little about groups of variables.


eigenvector number 1

eigenvector number 5

eigenvector number 2

eigenvector number 6

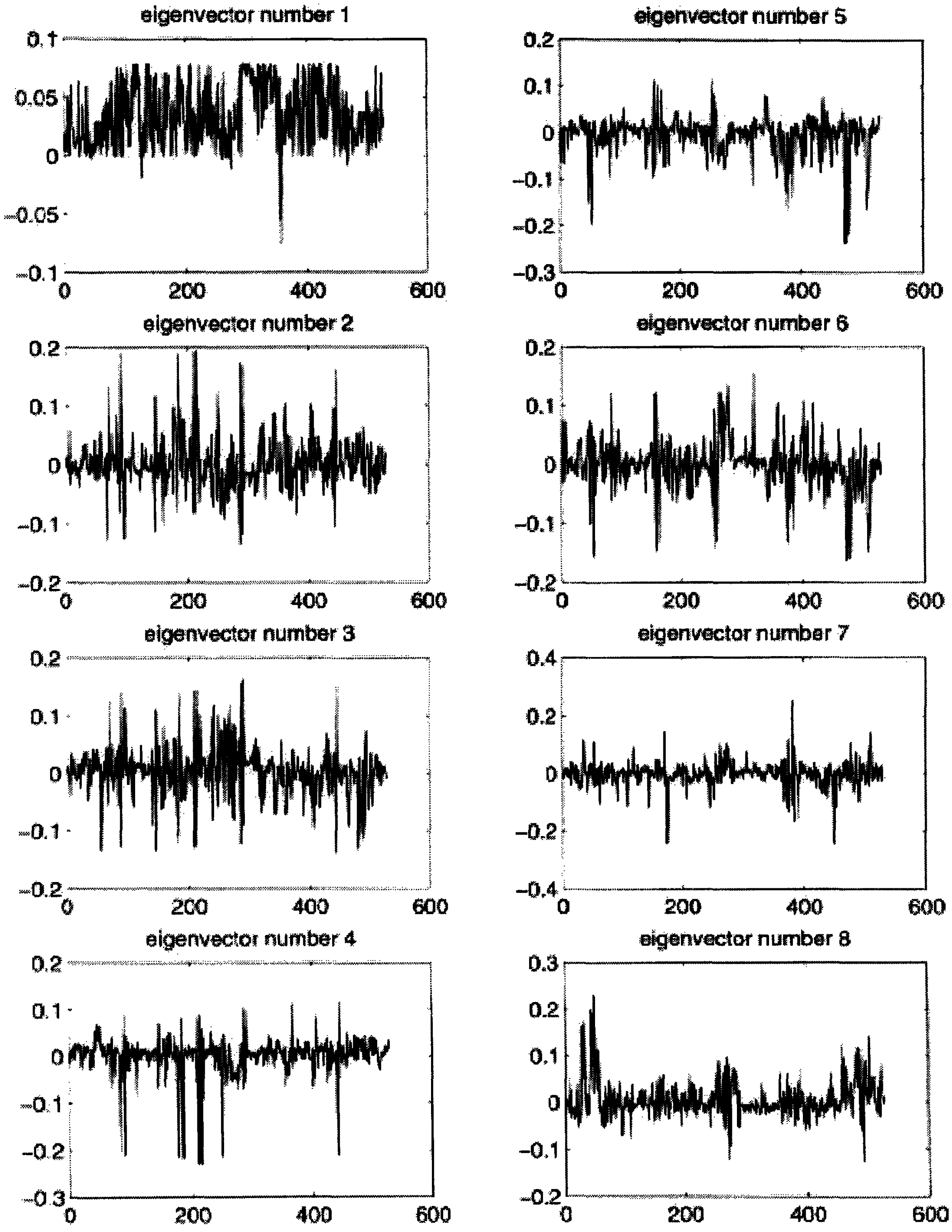
eigenvector number 3

eigenvector number 7

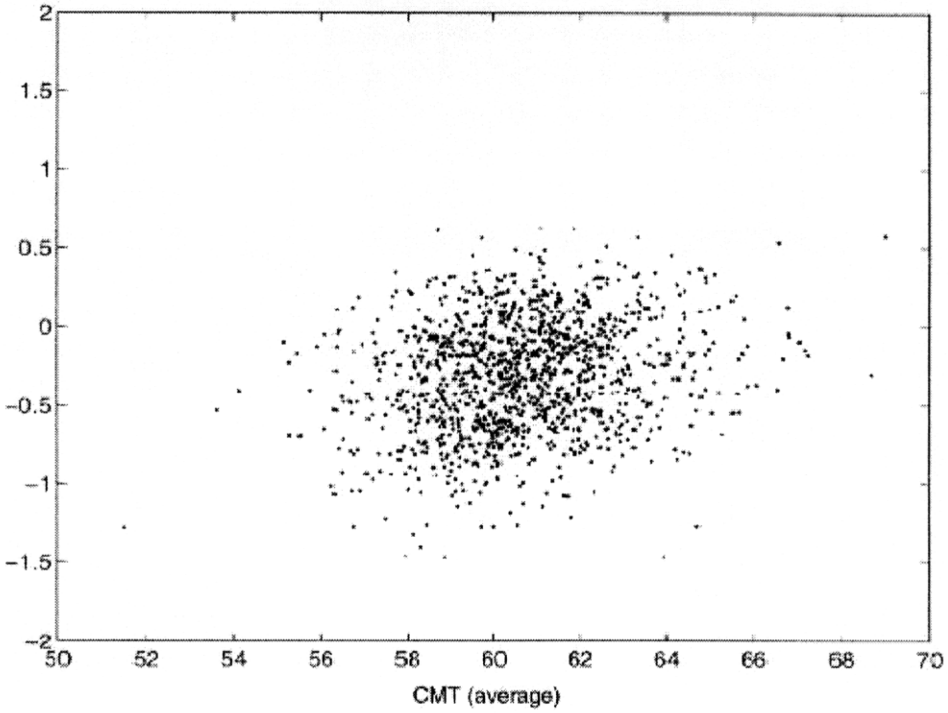
eigenvector number 4

eigenvector number 8

There is little obvious information to be gathered from the plots in their current state. They look fairly random in their distribution and tell us little about individual, or groups, of variables. When these same principal components are filtered with a high pass filter, we can see there is not much more information than the unfiltered eigenvectors.



When the second data set is regressed with the quality variable of the CMT test, we get the following plot. The x-axis represents the true value and the y-axis represents the difference from its expected value.

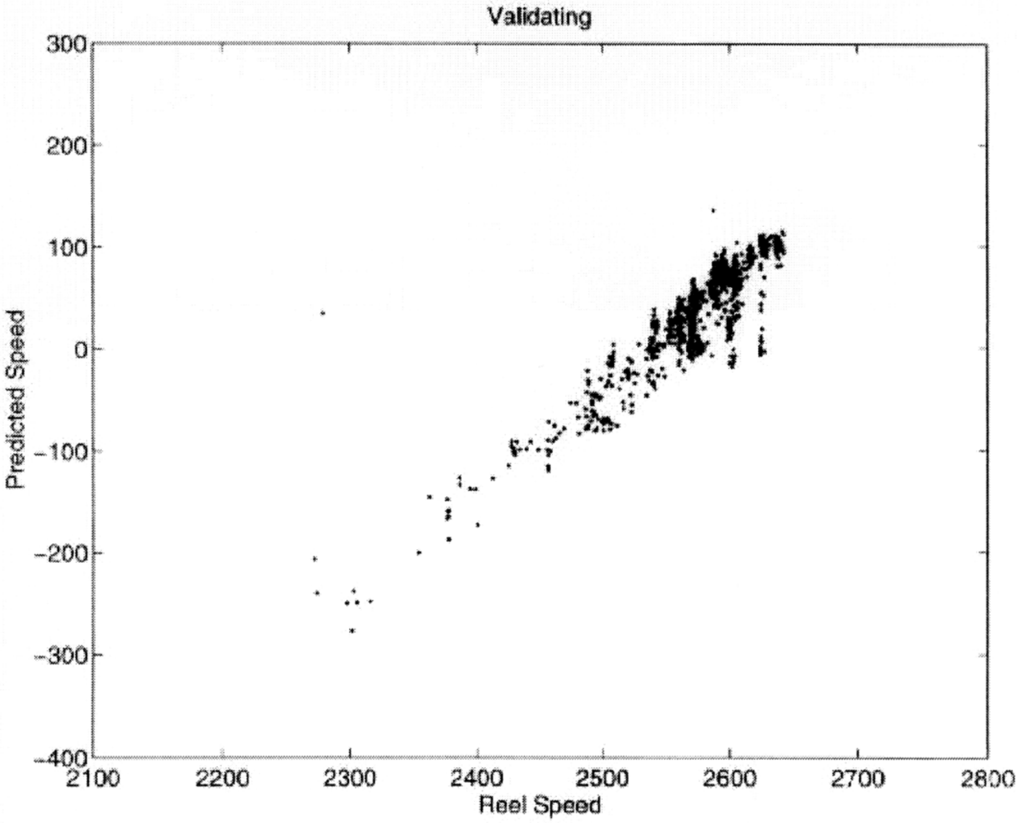


As can be seen by the plot, the distribution is extremely random and offers little strength as a predictor of CMT given our chosen process variables.

When the data set is regressed with the quality variables of machine speed we get the following lists of important variables and the accompanying plot.

Tag Name	Descriptor	Relative value
P4:PIC202	Headbox Pressure	15.62
P4:PIC202.SP	Headbox Pressure Set Point	15.16
P4:THSTPT.PE	Delta Stpt vs Actual	13.9
P4:PI209	Headbox Level Drive Side	8.07
P4:HBXPRS:MN	Total Head Inches	5.87
P4:JETWIR.MN	Jet/Wire Ratio Set Point	5.56

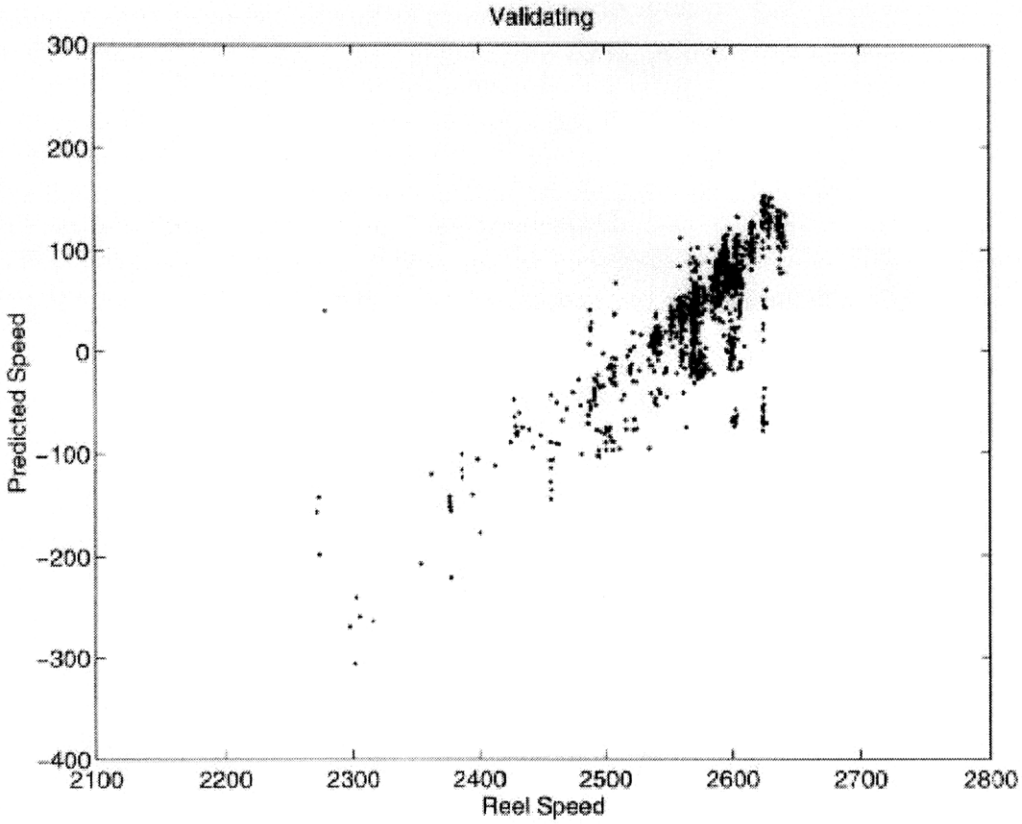
P4:PIC202.CO	Headbox Pressure C.O.	4.48
P4:II198	Fan Pump % Load	4.33
P4:SI199	Fan Pump Speed	4.21
B8: ST%	#8 Blower Steam Flow %	4.14



One can see from the speed plot that this regression is powerful and allows us to accurately predict what speed we should have given the information from this short list of coefficients. Unfortunately, this is where we run into problems of multicollinearity or clone variables. When we take a closer look at the list of most influential variables, we see that many of the 10 most important variables are variables that are obviously closely correlated with machine speed (highlighted variables).

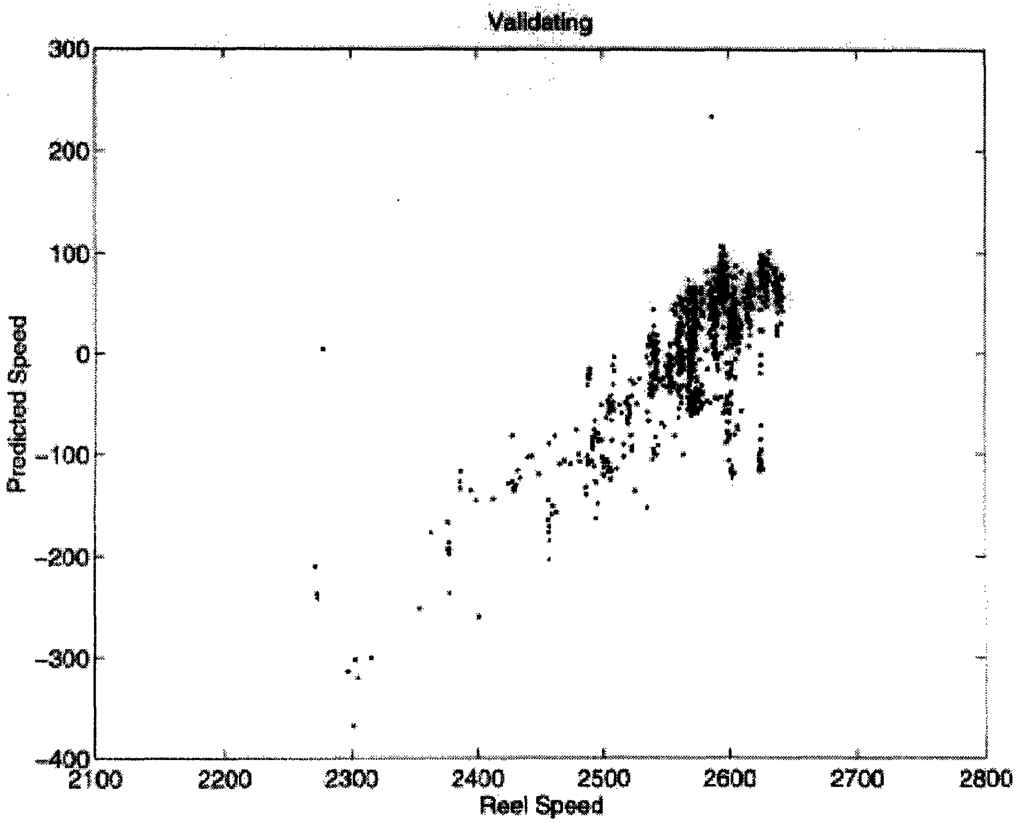
To address this issue we take the clone observational variables out of the matrix and re-run the analysis. When we do, we find PKLT returns the following list of most influential variables and the plot of predicted values:

Tag Name	Descriptor	Relative value
P4:RD/WP%.PE	% rush drag/wire speed	7.95
P4:TPACTL.TU	Actual throughput	7.88
P4:BASWT	Basis weight	7.40
P4:STMPAP	#4#steam/#paper	6.29
P4:BDSAVE.ES	Bone dry scan average	5.75
P4:FI20	PM #4 stock flow (gpm)	5.42
P4:LI190	Silo level	5.16
P4:STKFLW.MN	Stock flow	4.92
P4:STKTP.H	#4 stock tons/hour	4.88
P4:STKSPT.MN	Stock flow set point	4.80



Once again, we find some variables that are obvious observational clones of machine speed so we repeat the process of variable elimination and recalculation. This second iteration produces the following list of variables and graph:

Tag Name	Descriptor	Relative value
P4:BDSAVE.ES	Bone dry scan average	10.26
P4:STMPAP.	#steam/#paper	8.09
P4:REFSPT.4	Refiner #4 power	7.33
P4:BASWT	Basis weight	6.43
P4:REF4KW.MN	Refiner #4 power	6.10
P4:LI190	Solo level	5.94
P4:STKSPT.MN	Stock flow set point	5.68
P4:DRAW6	1st to 2nd dyer draw	5.42
P4:FI302	200 # steam flow	4.74
P4:TPDREF.4	No. 4 DD refiner	4.74



This time we find only one variable that has an obvious correlation with machine speed yet we also see that our predictive powers have been reduced. Fortunately, however, it looks like that there is still predictive power in the regression. From this list of variables, a designed experiment was created and can be tested (see attachment #1) to demonstrate any ability of these controllable variables to permit increases in machine speed while maintaining product quality.

7. Data Set #3

As time for the internship was drawing to a close it was decided to gather one last data set that could be used for further research. This data set, like the first data set, was drawn from the complete list of active tags. The time interval for this last data set was one-hour data. An entire year's worth of data was collected. While this took considerable time, it will hopefully provide sufficient information to draw some solid conclusions.

7.1 Data Cleaning

Because of the size of this data set and desire to maximize the amount of information retained from the data set, it was decided to try to clean the data with a series of smaller programs. The first program was made to remove all variables that had less than 90% good data.

7.2 Data processing

The processing and results from this data set will not be included in this thesis. As mentioned previously, look to the doctoral thesis of Junehee Lee and Geoffrey Lauprete for more information.

8. Variation in First Stage Processing

As mentioned earlier in the thesis, while the on-site intern was not downloading, cleaning, or interpreting the data sets, there was time to do some independent studies. These studies focused on finding important sources of variation in the process and, hopefully, possible solutions for these sources of variation.

One area of interest was the first stage processing. This area has four independent parallel systems (see figure # 4.4.1). The fact that the fibers receive this particular type of treatment in this one area, and nowhere else in the process, and the fact that the four lines are totally independent make them interesting subjects for the potential introduction of variation.

8.1 Variation Measurement

One of the issues that makes it difficult to track variation in this area is the lack of a quick, accurate quality variable that can be monitored frequently and inexpensively. Most of the tests that can be done at this stage that yield any kind of useful information have to be performed in a laboratory and take considerable time to perform. As could be expected, typically, the more laborious tests yielded the most helpful and accurate information.

Three variables were tracked over time to attempt to observe the behavior of the different processors. The following measurements are static measures from samples removed from the system on the date shown. If data are missing it is because that particular test was not running on that particular day or the processor was not operating.

The 1st and 2nd processors receive raw materials from the same source and different from that of the 3rd and 4th processors. Because of this difference it will be interesting to see if the two groups of processors (1 and 2 vs 3 and 4) show different behavior. If they did produce different test results, it would lead us to think that the raw material feeds are significantly different.

The data below, from the first variable, suggests that the 1st and 4th processors are behaving similarly yet the 2nd and 3rd processors are behaving more independently. The summary at the bottom of the table is the results of a difference of means test. It suggests that regardless of the apparent similarities between two of the digesters, over all, the digesters seem to be giving independent results.

Processor (variable #1, %)
Target 20-25%

		1	2	3	4
Date	6-Jun-97	26.41		15.48	30.94
	9-Jun-97	24.62		18.08	30.62
	12-Jun-97				
	16-Jun-97			21.03	32.35
	18-Jun-97		25.66	17.8	28.84
	19-Jun-97	30.69	20.01	15.07	
	20-Jun-97	27.77	21.64	17.73	
	23-Jun-97				28.52
	24-Jun-97		22.13	18.57	31.02
	25-Jun-97		25.27	19.42	22.18
	26-Jun-97	26.4	20.83	19.12	27.1
	27-Jun-97	22.17	22.21	20	26.46
	30-Jun-97	26.33	22.48	21.71	34.4
	1-Jul-97	28.28	20.69	23.01	29.15
	8-Jul-97	26.27	24.26	25.67	32.88
	10-Jul-97	28.95	22.59		33.54
	15-Jul-97	28.11	23.88		
	16-Jul-97	30.28	26	26.63	27.42
	18-Jul-97	24.4	19.84	20.7	
	21-Jul-97	26.37		17.65	26.64
	23-Jul-97	29.73		21.88	22.23
	24-Jul-97	28.25	27.09		
	28-Jul-97	28.56	24.37	18.21	26.74

Mean	27.27	23.06	19.88	28.88
Variance	4.98	4.99	9.60	12.63
Var/n	0.31	0.33	0.53	0.79
Means could be the same as the grand mean:	FALSE	FALSE	FALSE	FALSE

Grand	Average	24.73
	Variance	20.77
	Var/n	0.32

The next table of data, on the second variable, shows larger swings in the larger means yet processors #3 and #4 seem to be within the range of the grand mean. Regardless, there doesn't seem to be any really helpful information to be gleaned from the data other than the fact it seems to have a large variance.

Processor (variable #2, %)
Target 25-30%

		1	2	3	4
Date	6-Jun-97	8.92		9.04	8.77
	9-Jun-97	8.7		8.83	8.64
	12-Jun-97				
	16-Jun-97			21.68	35.36
	18-Jun-97		22.34	23.22	34.63
	19-Jun-97	30.59	14.33	33.09	
	20-Jun-97	36.59	10.84	26.1	
	23-Jun-97				51.57
	24-Jun-97		9.6	23.55	25.75
	25-Jun-97		6.65	9.33	36.55
	26-Jun-97	36.97	8.64	20.97	28.24
	27-Jun-97	33.26	12.31	20.33	29.93
	30-Jun-97	27.44	6.64	27.08	22.33
	1-Jul-97	30.92	11.79	33.95	21.26
	8-Jul-97	38.47	13.14	49.5	23.07
	10-Jul-97	29.29	10.92		35.9
	15-Jul-97	34.48	16.81		
	16-Jul-97	30.64	11.54	14.52	39.92
	18-Jul-97	43.61	13.24	23	
	21-Jul-97	42.83		11.26	24.57
	23-Jul-97	38.52		22.74	35.57
	24-Jul-97	32.76	18.06		
	28-Jul-97	42.53	12.69	18.67	23.49

Mean	32.15	12.47	22.05	28.56
Variance	100.61	16.62	103.43	116.92

Var/n	6.29	1.11	5.75	7.31
Means could be the same as the grand mean:	FALSE	FALSE	TRUE	TRUE

Grand	Average	23.95
	Variance	136.80
	var/n	2.10

Variable #3 is considered to be the most accurate and precise of the readily available tests and shows an interesting pattern. The 1st and 2nd processors seem have the same mean, and maybe more significantly, many of the values are very close to each other on days when the test was taken for both processors. Similarly, except for the early data, the 3rd and 4th processors also seem to move together. This supports the earlier idea that the different raw material sources for the two sets of processors could be adding to the variability of the product.

Another interesting observation from this data is the parallel movement of the results from the 3rd and 4th digesters (see graph below). The data also shows that the 3rd processor is running at a level consistently lower than that of the 4th processor. This would suggest that the two processors are seeing the same raw materials changes yet are working at different settings. More specifically, it appears that the 3rd processor needs to have it's settings changed to move it's mean to the same value as the other processors.

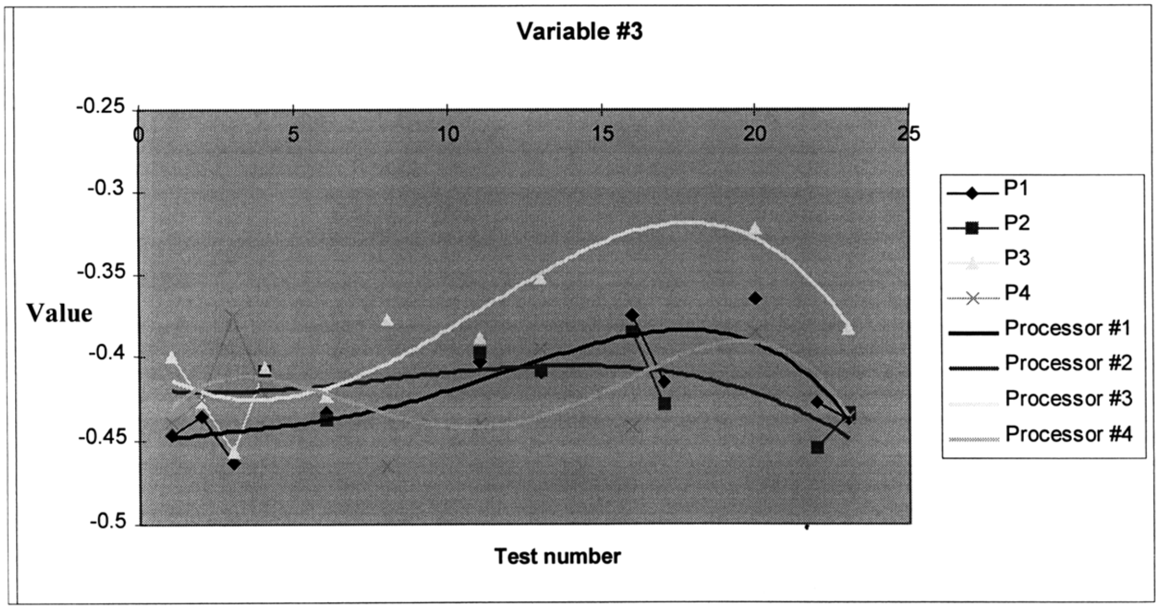
Processor (variable #3)

		1	2	3	4
Date	6-Jun-97	-0.4468		-0.3993	-0.4388
	9-Jun-97	-0.4353		-0.4268	-0.4249
	12-Jun-97	-0.4629		-0.4567	-0.3751
	16-Jun-97		-0.4083	-0.4065	-0.4214
	18-Jun-97				
	19-Jun-97	-0.4329	-0.4371	-0.4228	
	20-Jun-97				
	23-Jun-97			-0.3769	-0.4652
	24-Jun-97				
	25-Jun-97				

26-Jun-97	-0.4024	-0.3972	-0.3883	-0.4393
27-Jun-97				
30-Jun-97	-0.4094	-0.4088	-0.3524	-0.3951
1-Jul-97				
8-Jul-97				
10-Jul-97	-0.3753	-0.3846		-0.4416
15-Jul-97	-0.4153	-0.4283		
16-Jul-97				
18-Jul-97				
21-Jul-97	-0.3649		-0.3224	-0.3854
23-Jul-97				
24-Jul-97	-0.4269	-0.4539		
28-Jul-97	-0.4373	-0.4341	-0.3832	-0.4361

Mean	-0.41904	-0.41904	-0.39353	-0.42229
Variance	0.00088	0.00053	0.00148	0.00081
Var/n	0.00005	0.00003	0.00009	0.00005
Means could be the same as the grand mean:	FALSE	FALSE	FALSE	FALSE

Grand	Average	-0.4133
	Variance	0.0010
	var/n	0.0000



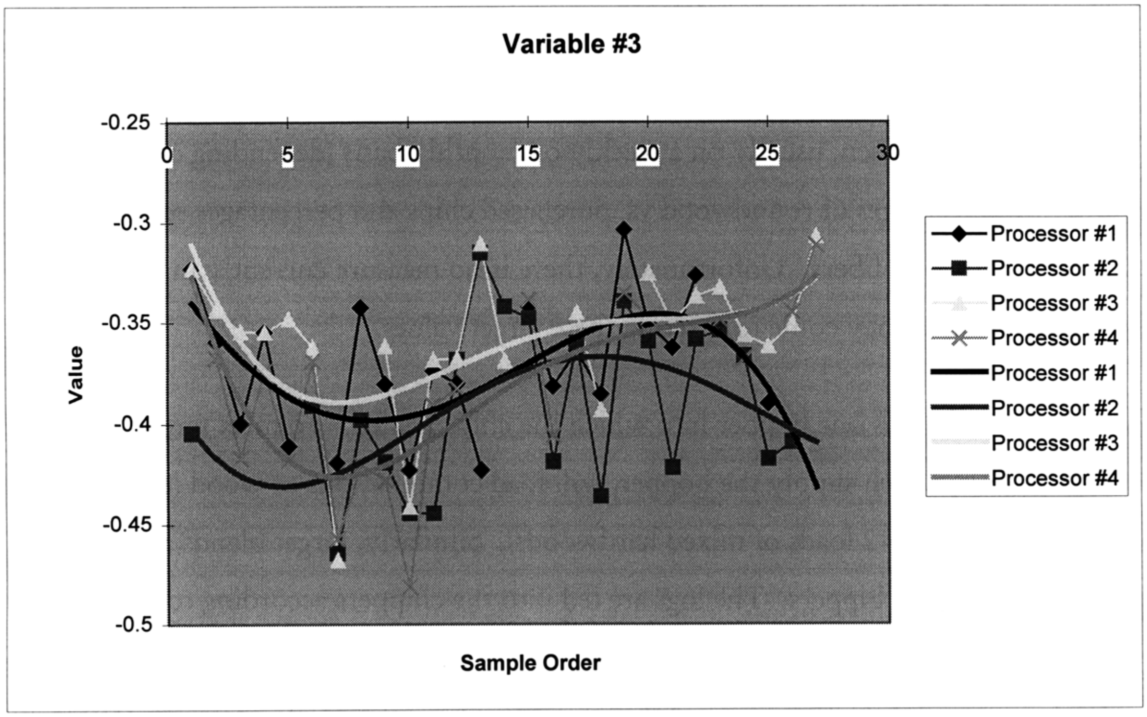
At this point in the mill several hardware changes were made in this area and when we continued to track the data we lost some of the pattern we saw earlier. The following data shows variable number 3 data taken over a period of significant adjustments and changes.

Variable #3

Date	Processor #1	Processor #2	Processor #3	Processor #4
10/7/97	-0.3229	-0.4047	-0.3229	
10/8/97	-0.3593		-0.3436	-0.3673
10/14/97	-0.3997		-0.3543	-0.4168
10/15/97	-0.3546		-0.3546	-0.3738
10/16/97	-0.411		-0.348	-0.418
10/17/97		-0.391	-0.362	-0.369
10/18/97	-0.4192	-0.4644	-0.4676	-0.457
10/18/97	-0.3424	-0.3981		-0.4202
10/19/97	-0.3803	-0.4191	-0.3616	-0.4281
10/19/97	-0.4229	-0.4445	-0.4411	-0.4811
10/21/97	-0.3738	-0.4445	-0.3683	-0.3909
10/22/97	-0.3803	-0.3683	-0.3676	-0.3826
10/23/97	-0.4229	-0.3152	-0.3105	
10/24/97		-0.342	-0.3689	
10/28/97	-0.3419	-0.3477		-0.3385
10/30/97	-0.3813	-0.4189		-0.4058
11/3/97	-0.364	-0.36	-0.345	
11/6/97	-0.3854	-0.4361	-0.3931	
11/7/97	-0.3036	-0.3393		-0.3352
11/8/97	-0.3506	-0.3591	-0.3252	-0.3476
11/10/97	-0.3624	-0.422	-0.3503	
11/11/97	-0.3265	-0.3583	-0.3368	
11/13/97		-0.3538	-0.3324	-0.3485
11/14/97		-0.3665	-0.3562	-0.3679
11/17/97	-0.3899	-0.4175	-0.3617	
11/18/97		-0.4092	-0.3506	-0.3448
11/19/97			-0.3069	-0.3112

average	-0.37119	-0.39001	-0.36010	-0.38851
variance	0.00112	0.00167	0.00127	0.00176
St D	0.000001	0.000003	0.000002	0.000003

The plot from this data also shows a pattern between several of the digesters but the movement is not as close between the pairs 1-2 and 3-4 as the previous plot.



When we run correlations between the different processors, we get the following results:

Correlations	Processor #1	Processor #2	Processor #3	Processor #4
Processor #1	1	0.416502422	0.533375874	0.834823293
Processor #2	0.416502422	1	0.710113037	0.815643311
Processor #3	0.533375874	0.710113037	1	0.83878821
Processor #4	0.834823293	0.815643311	0.83878821	1

Which suggests parallel movement by some of the processors but not the ones we might have anticipated.

8.2 Sources of Variation

The inputs to the processors are fiber, chemicals and heat. Other sources of variation include processor maintenance and component age. Of these inputs, the most important variable is the fiber source. Both the heat source and the chemicals are tracked constantly as is the *rate* of fibrous material fed into the processors. Aside from unforeseen breaks, the processor maintenance is scheduled. What is difficult to track is the particular make-up of the chips going into the processors.

8.3 Variation Control

Target values are chosen, usually on a weekly or monthly basis (depending on material supply) for percentages of roundwood vs. purchased chips and percentages of aspen fiber vs. mixed hardwood fibers. Unfortunately, there is no measure currently in place to measure how close they come to meeting these target values.

Mechanically, there is one hopper into which the chips are fed. To meet target mixtures, the end loaders, which supply the hopper, will load combinations of wood (e.g. 2 loads of aspen followed by 2 loads of mixed hardwoods). Similarly, target blends are chosen for the roundwood chippers. The logs are fed into the chippers according to whichever type of log comes out of the debarker. The debarkers are feed from the log decks. The log haulers unload the inbound wood shipments and keep the log decks supplied with roundwood.

9. Raw Material Variation

The raw materials are supplied to the mill from dozens of different sources. Aside from species variation, the material can vary in size, biological age, age since harvesting and overall biological quality. Because of the type of paper produced at the mill the process can accommodate large differences in the quality of raw material purchased by the mill. This flexibility offers the mill a considerable raw material price advantage over mills which have more specific needs, but it also facilitates larger swings in raw material quality.

9.1 Variation Measurement

Of the observable sources of variation in raw material quality, the most obvious and readily controllable is the wood species. At the beginning of the internship there was little being done to measure the variation in the incoming fiber other than the target percentages mentioned above. During the time of the internship several tests were performed to evaluate chip quality and species.

9.2 Variation Control

Because of space and manpower constraints, the unloading and storage of the purchased chips consisted of piling the material in large piles near the chip hopper. The chips were never sorted into different species which made it subsequently more difficult to load the chip hopper according to target percentages. By the end of the internship, the mill management, with the cooperation of the wood yard personnel came up with a solution to the difficulty in sorting the chips and now there is better control over the different species.

10. A Possible Solution to Early Variation

With the existing configuration of the mill there are several possible methods to address the existing level of variation in raw materials to the processors. But before these options are delineated, I would first like to discuss why it is of such importance.

10.1 Theory of Constraints Before the Bottleneck

The bottleneck processes at the mill are the paper machines. Everything else in the system has some degree of excess capacity, as the paper machines must keep running. The theory of constraints suggests that everything before the bottleneck should be inspected before it gets to the bottleneck since lost production at the bottleneck is lost forever and cannot be reclaimed. If a part is fed through the bottleneck and it is defective from a process that preceded the bottleneck, the whole system has lost the time required to make that product.

While a web process is continuous, meaning you cannot single out units for inspection, the concept still applies. In as much as is possible, the system should be modified and optimized before the bottleneck. In the case of a paper machine as your bottleneck, this means you want to run the highest quality processed fiber possible into your machines to maximize production.

10.2 Why Variation Reduction is Particularly Important before Final Processing in a Web Process.

In a web process, the issue of quality is compounded by the equally important issue of material consistency. The artistry in papermaking is being able to take a number of imperfect measures and be able to predict the best possible speed to run your paper machine. The cost of being too optimistic is breaks in the web which cost considerable time in lost production and cause increased wear on your machines. The price of being too conservative about runability is that one runs the machines more slowly than they could run and a certain amount of production is lost forever. If the runability is constantly changing and you can't anticipate its movement perfectly, you are destined to lose considerable production.

Imagine that the potential runability of the incoming material acts like a sine wave around a target value. In a web process you would run the machine more slowly than its capacity and slowly speed up production until you exceeded the capacity of the material. At this point you would slow down your process until you became more confident and slowly began to increase the speed again. The difference between the theoretical optimum and the actual machine speed is a consistent loss of production.

There are two issues to address in the search for increased runability: the reduction of this variance which drives your machine speeds down because of dips in fiber quality and the increase in the overall mean of fiber quality. One wants to stabilize quality and, as much as possible, raise it. This challenge is considerable but the gains should be as well.

10.3 Robust solutions through process control

To reduce the variation of the fiber quality to the machines, one must reduce the variation in the quality of the fiber leaving the processors. To make the processors run more closely with one another and run more consistently, they must receive similar raw material mixes and be calibrated so they all run the same. To make sure the processors receive the same raw materials the mill has to gain control in the wood yard.

An important step was taken to reduce the variation in the chip supply when the mill began separating the chips in the wood yard. Now, when the mill targets a certain ratio of aspen chips to other mixed hardwoods, it can more accurately control the loads of each group. In addition to this, the mill could designate certain species to certain chipper lines thereby increasing the amount of control over that area of chip supply.

These improvements would help control the mix but a more foolproof method is at the mill's disposal because of their unique configuration. Since the mill has two storage units before the processors, it has the option of dedicating the two different lines to the two primary groups of chip that are used at the mill. If the north storage unit and the 1st and

2nd processors were dedicated to mixed hardwoods and the south storage unit and the 3rd and 4th processors used exclusively for aspen fiber, one could gain perfect control over the species ratios by controlling each stream's output. The added advantage of this configuration is the reduction of the variation in the fiber ratios and the two different lines can be adjusted to maximize yield of the two different groups of raw materials.

While it is recognized that these kinds of control improvements are not often achieved quickly or easily, the recognition of the potential gives important direction. While the logistics of such changes will take some time to implement, there are some more incremental, low cost improvements that could be made fairly immediately. Two of the more important quick changes would include: a) performing a design of experiments to make sure the processors are running similarly (especially the 1-2, 3-4 pairs) and b) changing the north-south silo diverter from a binary, north/south diverter to a system that would allow both storage units to be filled at the same time.

By modifying the diverter and working to fill the chip silos at the same rate, the silos will be filled with the same type of material and the size of the material swings to the silos will be reduced. For example, if the system were subject to a surge of all aspen chips most of those chips would be sent to a single silo. If the load were to be split, all four processors would receive the same type of chips and therefore reduce the time the system would be subject to the anomalous surge. Having all the processors receive the same chips would also facilitate synchronizing them in the design of experiments.

11. High Frequency Variation in Final Processing

Besides the reduction of variation before final processing, another obvious area of concern would be the identification and reduction of variation in final processing. Due to the fiber cost and web strength issues, considerable effort, expense, and machinery are deployed to address the issue of uniformity in the web sheet. Ideally, there would be perfect consistency in fiber content and moisture in the sheet in both the cross direction (the non-continuous direction) and in the machine direction. In reality, there is so much vibration in the large machinery and other imperfections that this remains a considerable obstacle.

11.1 Variation Identification

There are two main patterns that can be identified in a web. One can think of it like a road with both ruts running parallel to the road and washboard bumps that run transversely across the road surface. The ruts are typically attributable to unevenness in the slice where the fiber is fed onto the wet end of the machine or some other imperfection that concentrates the moisture or fiber into rows.

To identify the severity of the "ruts" in the web, the paper machines have, incorporated into the machine, a scanning system that identifies and feeds back the quantity of fiber and moisture in the web. This information is averaged and feed into an automatic adjustment system that controls the gap where the fiber feeds into the machine.

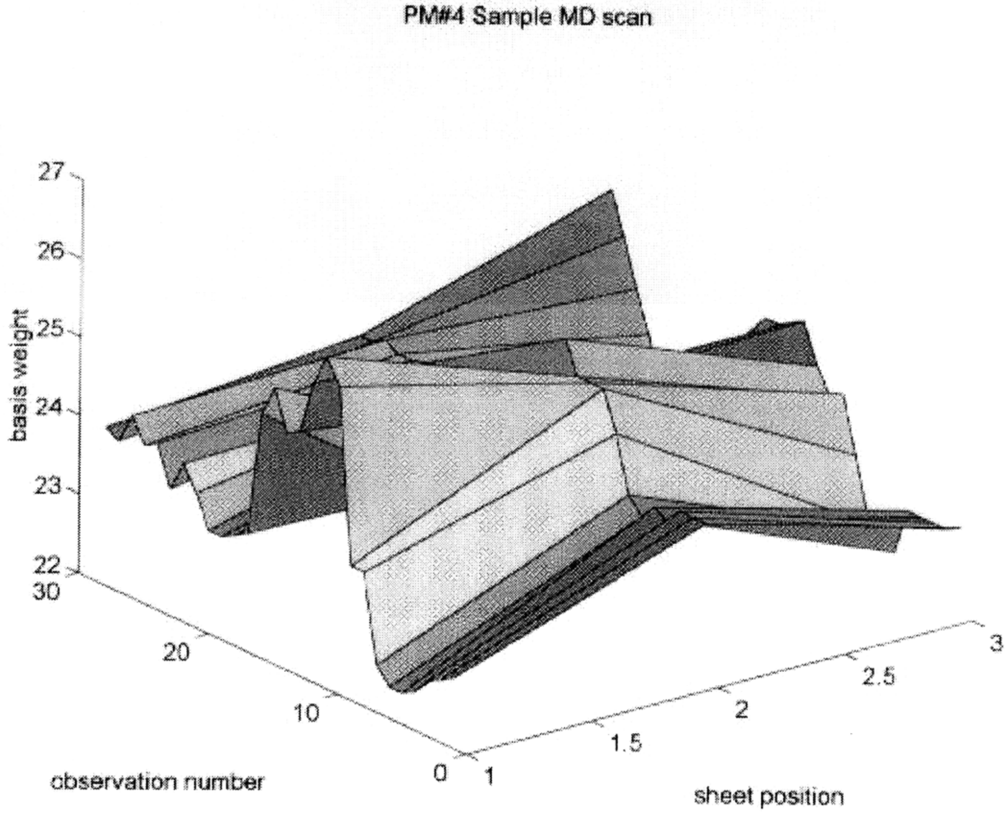
Imperfections in the other direction are more difficult in as much as they are recorded and fed into an automated system. When the washboard pattern appears in the paper it is typically due to the vibration of some particular piece of machinery in the process. This phenomenon can be particularly allusive as it can come and go as the machine speed or paper thickness changes. Another obstacle exists in that the web sheet is usually moving at speeds that make it impossible to see high frequency imperfections with the naked eye.

11.2 Variation Measurement

The scanner that works back and forth across the web to identify cross direction inconsistencies measures at fixed points as it travels. If you were to connect the scan points along the sheet, you would create a long zigzag pattern that would travel from one side of the sheet to the other. As mentioned above, this works well for cross direction control since these imperfections do not tend to change quickly over time. It is, however, much too slow for the washboard.

To observe the possible existence of washboard imperfections it is necessary to record at high speeds or slow the paper down. If you choose to slow the paper down it must be done off-line and at considerable cost. One way around this problem is to watch the paper as it is transferred and fed into the winder. At the beginning of the transfer, the paper is slowly accelerated at which point you can stand under the sheet and watch as the overhead light passes through the sheet. A step beyond this is to observe the web until the sheet shows strong washboard and then take a scrap of the sheet and hold it up to a light source such as a window or light table.

A more advanced technique (although still fairly primitive) is to take a sample that is thought to have a washboard problem and manually feed it under the cross directional scanner while the paper machine is down and the scanner is set on single point scan. This method allows you quantify the difference in the "peaks" and "valleys" of the washboard. The following graph is an example of one sample that was measured in this way:



These are all ways to get an idea of the severity of the problem. The best method is to put the cross directional scanner on a single point read and record your information at a high frequency. This requires some high tech equipment but can be quite beneficial.

Because of reoccurrence of this phenomenon on one particular machine, one of the engineers from corporate headquarters was sent to the mill to perform studies with high frequency reading equipment. The output from the experiment showed high variation in the same region (30 inches) of the wave in the above graph (20 inches) but different enough to lead us to believe that we had not quite found the source of trouble.

11.3 Elimination

The process of eliminating the washboard problem is well beyond the expertise of the author and is best left in the hands of the specialist at the facility and the corporate engineers. The hope is that the issue, now raised, can now receive the proper attention and result in faster machine speeds with less breaks.

12. Conclusions

The costs of variation in manufacturing processes are well documented. Variation can effect finished goods shipment, raw materials content (and hence cost), product performance, equipment performance, equipment life, etc. The purpose of this internship has been to work with the host company to explore the existing operations for important sources of variation.

The internship facility was particularly interesting to the Leaders for Manufacturing Research Group #4 because of the potential to use multivariate analysis as an exploration tool. To date, working with the existing compiled data sets has resulted in the following information:

- The CMT test is a noisy quality test with considerable variation inherent in the test method. This is understood by industry as the test itself states "...the precision, repeatability, (within a laboratory) is 4.5% with 10 specimens/average..."[1]. The result of this information is, if you want accurate information about the CMT test, you have to take many samples. The other option would be to explore for other test by which to qualify the product. Because of industry standards, this would be difficult to change but the parent company could explore the possibility of experimenting with other standards with their internal customers.
- We have been able to iterate, with KLT, down to a reasonable list of influential variables that could be used in a design of experiments. The next step for this list would be to either run the experiment or analyze it again (with engineering judgement) and run KLT again. The amount of adjustment needed to run the experiment should not have a detrimental effect on production and could yield some interesting information.

Other experiments have shown the potential for variation reduction in raw material or first stage processing variation by (listed in order of assumed cost and ease of implementation):

[1] TAPPI Flat crush of corrugating medium test methods T 809 om-87

- Modifying the diverter on the chip silos so that both silos fill with the same mix of chips. This will make it easier to compare the four different processors and align their output. It should also reduce the "slugs" of material that could arrive to either of the silos.
- Continued experimentation to control the processors so they are all cooking the fiber the same. In particular, processor #3 seems to be out of sync with the other processors. Once the processors are running together, find optimum cooks for different mixes of fiber.
- Implement further separation controls in the woodyard to help the operators better control the mix of chips and roundwood. This could include large outdoor signs that remind the operators what the mix is supposed to be on during any particular period, or other systems (better if suggested by the operators themselves) which serves as a reminder to watch the mix.
- Designation of chipping lines and silos for specific species mixes combined with the optimization of the processors for the given mixes. This would be a step improvement in control for the early stages of processing at the mill. Maintenance would have to be tracked more carefully and buffers would need to run at higher capacity until confidence in the system was established.

13. Bibliography

Cao, Ronald "Multivariate Analysis of Manufacturing Data" Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1997.

Derksen, Timothy J. "The Treatment of Outliers and Mission Data in Multivariate Manufacturing Data." Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1996.

Devore, Jay L. *Probability and Statistics for Engineering and the Sciences* Brooks/Cole, 1991.

Goldratt, Eliyahu M. *The Goal.* North River Press, 1992.

Hogg, Robert V. and Johannes Ledolter *Applied Statistics for Engineers and Physical Scientists* Macmillan Publishing, 1992.

Posnack, Alan J. "Profiting From Constraint Management" Productivity Partners, 1996.

Rawizza, Mark A. "Time-Series Analysis of Multivariate Manufacturing Data Sets." Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1996.

Sjostrom, Eero, *Wood Chemistry Fundamentals and Applications.* Academic Press, 1981.

Smook, G. A. *Handbook for Pulp and Paper Technologists.* Joint Executive Committee of the Vocational Education Committees of the Pulp and Paper Industry, 1896.

Attachment #1

Design of experiment on increasing the process speed of the paper machine 4 of Tenneco Packaging Co. plant at Tomahawk, WI

by Junehee Lee, David Staelin

November 5, 1997

14. 1. Problem Statement

LFM RG4 has been working on the Tenneco Packaging paper production plant data set using multi-variable filtering such as principal component analysis and multi-variable linear regression. This report is focused on prediction of, and possible methods to increase the process speed represented as P4:RELSPD.MN, which is the reel speed measured in feet/minute. Before the linear regression, we applied principal component filtering to reduce the noise embedded in the various parameters so that the dimensionality of the process parameter space is reduced to a reasonably small number. After that, we applied the multi-variable linear regression technique to the smaller dimensional parameter space to find the relation between the speed and various parameters. Note that the candidates for the potentially important process parameters were carefully chosen by Tenneco personnel who understand the physics of the process very well. In this memorandum, we address the predictability of the process speed. Using the predictability and the relations between parameters found in the linear regression, we design an experiment that may verify our findings. If successful, this experiment could potentially result in a better operating point where the process speed is higher without further risking breaks in paper caused by extreme process conditions.

2. Predictability of Speed (P4:RELSPD.MN)

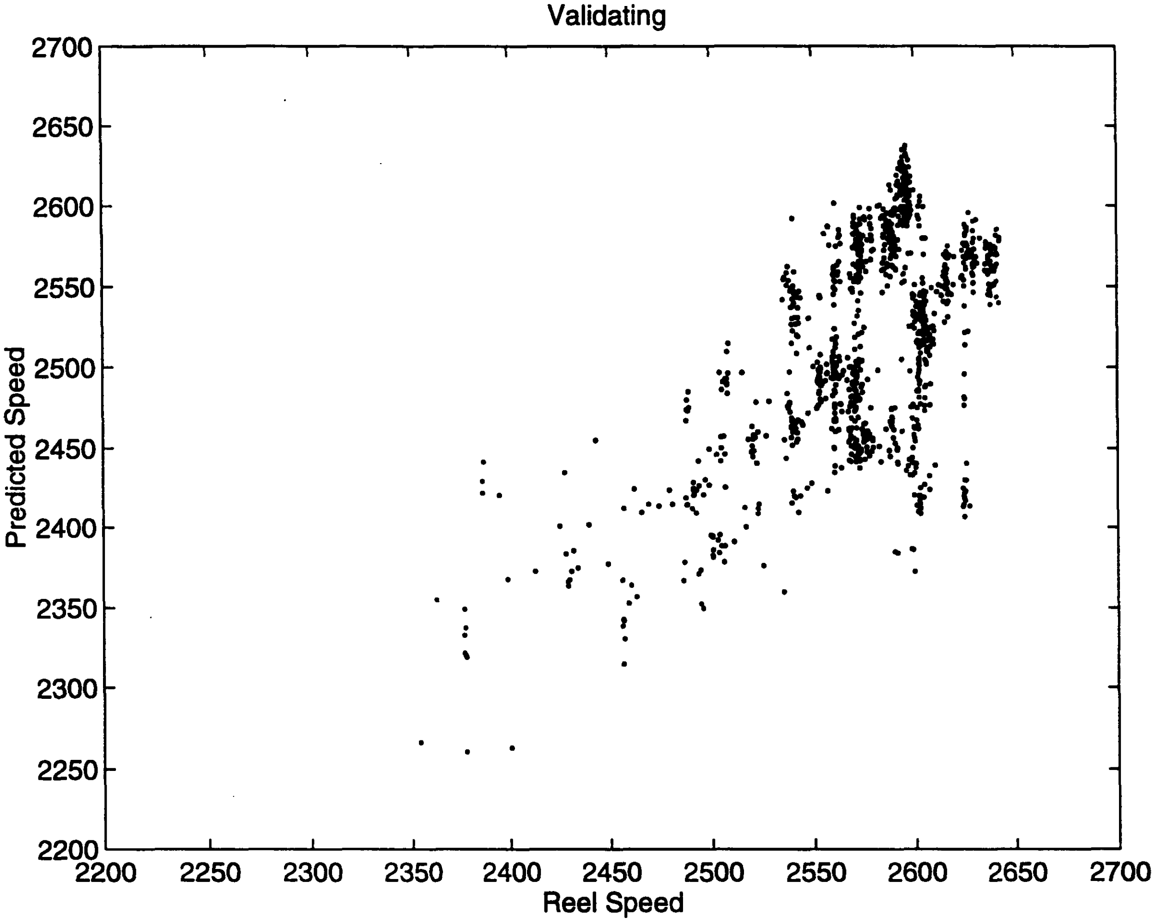
The analyzed data set contains 508 parameters measured once every hour for about 5 month. This data set focuses on the paper machine 4. First, we applied principal component filtering to the 508 parameters to reduce the noise. Having done that, we used standard linear regression techniques to find the relation between the parameters and the reel speed. For a reason which will be made clear later, we normalize each parameter by its standard deviation.

P4:BDSAVG.E Bone Dry Scan Avg.	-8.933
P4:STMPAP. Steam/Paper	-7.888
P4:STKSPT.MN Stock Flow Set Point	6.619
P4:REFSPT.4 Refiner Power	6.387
P4:BASWT . Basis Weight	-5.997
P4:LI190 . Silo Level	-5.744
P4:TPDREF.4 No 4 DD Ton/D	4.827
P4:REF4KW.MN Refiner No 4 Power	4.671
P4:FI113 . 4 Refiner Flow	4.650
P4:DRAW6 . 1st Dryer to 2nd Dryer	4.416

The table on the previous page is the result of linear regression for the 10 most important parameters. The second column contains the regression coefficients for parameters. The numbers in the second column represent the increases in speed when you increase the corresponding parameter by one standard deviation. For example, the first row says that if we increase P4:BDSAVG.ES by one standard deviation, the speed should decrease by 8.933 feet/minute.

When we apply linear regression, we divide the whole data set into two; the early data and the later data. The first set is used to train the relation between the parameters and the speed. The second set is used to validate the relations that are found in the regression of the first data set. We do this to avoid the over-training of data. More often than not, the trainirg data set gives a good linear relation. We can say that there is a time-lasting linear relation between parameters and speed only when the validating set also shows a linear relation.

The following graph shows the predictability of speed for the validating data set. The horizontal axis is the measured P4:RELSPD.MN and the vertical axis is the predicted speed using the relation found in regression of the training set.



Although it is not a straight line, we can see that there is a "good" linear predictability in speed. In the next section we propose a designed experiment that could potentially lead to an operating point with a higher process speed.

3. Designed Experiment

The following table lists the 10 most important parameters. The second column is the same as for the previous table. The third column is the mean of each parameter, and the fourth column is the standard deviation. We want to design an experiment based on the list. We choose the first 10 parameters to design the experiment. We propose two settings of process conditions in which one setting is designed for a faster process speed and the other setting is designed for a slower process speed.

For the faster process speed setting, we set the first 10 parameters in the list to their mean value shifted by its standard deviation. When a coefficient is a positive number, the parameter corresponding to the coefficient is set to be the mean value plus the standard deviation, while a parameter is set to be the mean value minus the standard deviation when the corresponding coefficient is a negative number. For the slower process speed, we reverse the operating point, that is, when a coefficient is a positive number, the parameter corresponding to the coefficient is set to be the mean value minus the standard deviation, and vice versa. The fifth and the sixth column show the operating point for the faster process and the operating point for the slower process, respectively.

Parameter	Coefficient	Mean	Standard Deviation	Fast Operating Point	Slow Operating Point
P4:BDSAVG.ES	-8.933	0.023644716	2.54951E-04	0.023389765	0.023899667
P4:STMPAP.	-7.888	0.001573469	5.71990E-05	0.00151627	0.001630668
P4:STKSPT.MN	6.619	3.484988766	0.091278819	3.576267585	3.393709947
P4:REFSPT.4	6.387	0.157090068	0.008683626	0.165773694	0.148406442
P4:BASWT .	-5.997	0.025858132	3.05672E-04	0.02555246	0.026163803
P4:LI190 .	-5.744	0.077147755	0.001098157	0.076049598	0.078245912
P4:TPDREF.4	4.827	0.409095672	0.029040914	0.438136586	0.380054758
P4:REF4KW.M	4.671	0.15716338	0.008919106	0.166082486	0.148244274
P4:FI113 .	4.650	1.455127016	0.100604802	1.555731818	1.354522214
P4:DRAW6 .	4.416	0.015547849	0.001676501	0.01722435	0.013871348

4. Expectation

We expect that the difference in process speed could be as much as 120 feet/minute by setting the above parameters at the fast operating point rather than the slow operating point. We do not expect a significant degradation in other quality variables such as paper break probability because we only change the process operating point from its average by one standard deviation, which is well within the natural variation of parameters.

The actual result can be affected by the co-linearity of parameters. For example, we may not be able to control each parameter independently. Therefore, we may not be able to reach the operating point that is suggested above. In addition, it is possible that the observed changes in the above parameters were partly the result of the change in speed, rather than the cause of it. In those cases, it is probable that the actual improvement in speed will be less than 120 feet/minute.

5. Future Analysis

A close eye should be kept on other quality variables such as CMT and ring crush while the experiment is being conducted. A real improvement in speed should come without any degradation in other quality variables. If degradation in CMT and/or ring crush happens, we should conduct another set of experiments in which we find the operating point that is optimized in terms of speed, CMT and ring crush. We define the optimum operating point as the operating point that results in the maximum possible speed without violating the quality constraint in CMT and ring crush.

3118-9